A FACTOR ANALYTIC STUDY OF THE INTERNAL STRUCTURE OF THE BRIGANCE COMPREHENSIVE INVENTORY OF BASIC SKILLS-II

By

DANIEL HYDE BREIDENBACH

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

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
College of Education

MAY 2009

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of DANIEL HYDE BREIDENBACH find it satisfactory and recommend that it be accepted.

_____

Brian F. French, Ph.D., Chair

_____

Michael S. Trevisan, Ph.D.

_____

David F. Feldon, Ph. D.

_____

Holmes Finch, Ph.D.

ACKNOWLEDGMENTS

I am not overstating the truth when I say that this dissertation could not have been completed without the support and assistance of my advisor, Dr. Brian French. His unfailing encouragement gave me the faith that I could complete the work, and his comments, suggestions, and insight brought out my best efforts. My heartfelt thanks go out to him. Thanks, also, are owed to his family for sharing him with his students.

I also thank the rest of my committee: Dr. W. Holmes Finch, Dr. David Feldon, and Dr. Michael Trevisan. Their input helped me refine my work and allowed me to complete this project in a timely manner.

A special note of thanks is due to Dr. Steven Nettles, Dr. Lawrence Fabrey, and the management team at Applied Measurement Professionals for their understanding and support as I finished my work. In addition, I thank Curriculum Associates, Inc., who provided funding for this project.

Finally, I owe more than I express to my wife, Lisa, and children, Ian and Margaret. Lisa would not let me give up. She pushed me when I needed to be pushed, and supported me when I needed support. Thank you for all your extra efforts to allow me to continue. Ian and Margaret: thank you for putting up with me. My dissertation is done; now let’s go outside and play!

A FACTOR ANALYTIC STUDY OF THE INTERNAL STRUCTURE OF THE BRIGANCE COMPREHENSIVE INVENTORY OF BASIC SKILLS-II

Abstract

by Daniel Hyde Breidenbach, Ph.D.
Washington State University
May 2009

Chair: Brian F. French

The Brigance *Comprehensive Inventory of Basic Skills-II* is the newest version of a long-standing instrument that is presented as useful for identifying student achievement, identifying and monitoring strengths and weaknesses, obtaining data to support referrals for further diagnostic testing, and reporting progress for individual educational plans. Since the CIBS-II is intended to comply with requirements of the NCLB act, as well as the IDEA, validity studies are required. This study provides evidence to support the use of CIBS-II scores as indicators of students' progress in various academic domains. This study is part of the overall standardization and validation project for the instrument.

Nine subtests are included in the standardization and validation study of the CIBS-II. This study seeks confirming or disconfirming evidence as to the proposed composite score structure. The scores from the standardization sample are used to find evidence of essential unidimensionality of subtests through the use of DIMTEST and to investigate the composite score structure through the use of confirmatory factor analysis.

DIMTEST results indicate that five subtests cannot be considered unidimensional. Several CFA models were fit to the standardization data, including the proposed

composite score model and multiple plausible rival models. The sample was randomly split in half to allow one subgroup to be used to test models while holding the second subgroup in reserve to cross-validate the best-fitting model. The best-fitting model was in accordance with the proposed composite score structure. This model was cross-validated with the second random subgroup to ensure that the final model was not replicating specific features of the sample and to support the conclusion that the selected model fits the entire sample.

Results provide support for the proposed composite score structure, but the support is tempered by evidence of multidimensionality in five subtests and by high interfactor correlations and structure coefficients, which are consistent with evidence of multidimensionality. Suggestions are made regarding additional studies to resolve these concerns.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Education reform can be viewed as a never-ending process. Metaphors such as "cycle," "pendulum swings," and "bandwagon" are common in descriptions of this process. As a result of growing concerns over educational progress in the United States (e.g., National Center for Education Statistics (NCES), 2000; NCES, 2008; National Commission of Excellence in Education (NCEE), 1983) and of international comparison studies (e.g., TIMSS; Mullis, Martin, Gonzalez, & Chrostowski, 2004), education has entered what has been referred to as "the accountability era" (Dwyer, 2005).

Public and political dissatisfaction with teaching and learning (Hart & Teeter, 2002) eventually led to the No Child Left Behind Act (NCLB, 2002). More than half of parents of school age children (52%) think that the U.S. education system needs "major changes or a complete overhaul" (Hart & Teeter, 2002, p. 2). The primary goal of education reform in the latter half of the 20th Century and the early years of the 21st Century has been to improve the level of achievement for all students in general and to reduce the achievement gap, that is, reduce racial, gender, and socioeconomic inequality in educational attainment (e.g., Croom, 1997; Hiebert, et al., 1997; Porter, 2005).

The sense of "accountability" in this era is applied at multiple levels: states, school districts, specific schools, and teachers, individually and severally, are increasingly held accountable (i.e., responsible) for students' educational progress. Under NCLB, states and school districts can lose funding and local control if their students fail to make adequate educational progress. Districts hold schools accountable through

pressure on administrators, pressure that is passed on to teachers, again through the threat of losing funding or control. NCLB requires that if a school fails to meet state-mandated performance for five years in a row, the school must be "reconstituted," which can mean replacement of teachers and administrators or reorganization of the school as a charter school (Howell, West, & Peterson, 2007).

Measuring educational attainment is a complex undertaking. NCLB has created the criterion of adequate yearly progress (AYP), which is intended to track whether teachers are helping all students improve (NCLB, 2002). This criterion has been criticized because it places unequal demands on high-achieving schools versus traditionally underserved schools (Peterson, 2007). Students who have very low initial achievement may fail to make AYP even though they show substantial achievement gain (Linn, 2005). Under NCLB, states created their own educational standards and their own criteria for achieving the "Proficient" level. Consequently, different states have different definitions for AYP (Lewis, 2005), and even within a state, schools cannot be meaningfully compared based on AYP (Linn). The variable that is used to hold districts, schools, and teachers accountable is test scores—specifically, standardized achievement test scores (Berry & Howell, 2008). However, overreliance on achievement tests in accountability systems can "produce perverse incentives and seriously inflated estimates of gains in student performance" (Koretz, 2002, p. 753).

Accountability and Validity

Achievement test scores are used for multiple purposes, including purposes declaimed as inappropriate: for example assessing teachers' effectiveness (Joshua, Joshua, & Kritsonis, 2006) or influencing the sale of homes (Kohn, 2000). Such uses fail

to take into account modern notions of test score validity, which stress that test scores should only be considered valid for making inferences about the originally intended use of the scores. Validation studies for achievement tests typically address the suitability of the tests for making inferences about particular strengths and weaknesses of individual students or the relative standing of students compared to others. The studies do *not* typically evaluate the tests' suitability for making comparisons between students or entire schools (American Psychological Association, 2001).

In light of the increased push for accountability and the associated increase in public scrutiny of test scores, it is crucial that achievement tests meet the highest standards in all aspects of the testing process (e.g., *Standards for Educational and Psychological Testing*, AERA, APA, & NCME, 1999). In particular, well conceived and properly reported validity studies not only provide evidence that test scores are meaningful, but they also inform test users of the intended uses of the scores, which is a necessary, though not a sufficient condition, for proper use of the scores.

The present-day accountability movement can trace its roots to the 1983 report *A Nation at Risk* (NCEE, 1983), which questioned the quality of the U.S. educational system and served as a call for reform, including steps to track the results of the reform. The NCLB Act of 2001 has its roots in various federal efforts to encourage states to implement measures to improve their educational outcomes. For example, in 1994 the Improving America's School Act (IASA) was signed into law. The act required states to create and implement educational standards and an assessment system to monitor progress toward those standards (Walberg, 2003). The act thus set into motion a process that could ultimately lead to a different achievement testing system in each state.

However, by the targeted date of 2000, few states had an assessment system specified, much less implemented (Cohen, 2002).

The NCLB act (2002) spelled out some of the same goals as the 1994 IASA; however, NCLB's accountability provisions were much more clearly specified, with consequences spelled out in the law for states that failed to achieve the provisions of the law by specific target dates. Although an elaborate review and approval process was implemented to vet the states' accountability plans, states were given wide leeway in devising their assessment systems (Erpenbach, Forte-Fast, & Potts, 2003). For example, states could report norm-referenced or criterion-referenced scores, although in the case of norm-referenced scores, states were required to set a state-level definition of proficiency. For example, Iowa selected the Iowa Test of Basic Skills as its assessment instrument. Proficiency on this norm-referenced instrument was defined as scoring at the 41st percentile or higher, (2002 National norms—spring standardization study) (Erpenbach et al., 2003).

In addition to the many state achievement assessments that have been developed, many achievement tests are published commercially for diagnostic use, low-stakes monitoring of student performance, screening of students with learning difficulties, etc. (Koretz & Hamilton, 2006). In the 2005 and 2007 Buros Mental Measurement Yearbooks, 65 commercially available achievement tests were reviewed (Spies & Plake, 2005; Geisinger, Spies, Carlson, & Plake, 2007). The combined classification index, which classifies all tests reviewed since the ninth edition of the yearbook (i.e., since 1985), lists 109 achievement tests (Buros Institute of Mental Measurements, n.d.). Students and teachers are surrounded by achievement tests, and it is incumbent on test

publishers to provide evidence that test scores are useful for the publishers' intended interpretations. In addition, it is incumbent on test users to judge whether a given test will produce scores that are useful for the user's intended interpretation (Messick, 1989). It is noteworthy that the test user can only fulfill the user's responsibility if the producer has fulfilled the producer's responsibility.

In this environment of accountability, important decisions are made on the basis of achievement test scores. Test score validity refers to the degree to which these decisions, and the inferences on which the decisions are based, are justified by supporting evidence (Linn, 2005). Various forms of supporting evidence can exist, including evidence based on: (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing (AERA et al., 1999). In the past, sources of evidence were referred to as different types of validity, including content validity: the extent to which the instrument's items represent the domain of interest; predictive validity: the extent to which the instrument predicts performance on measurements (e.g., achievement) in the future; concurrent validity: the relationship between the instrument's scores and scores on other measurements given at the same time; and construct validity: the extent to which the instrument's scores allow meaningful inferences about some psychological construct (Crocker & Algina, 1986). Modern notions of validity favor a more unified view as opposed to multiple kinds of validity and treat validation as an ongoing process rather than a one-time study (see chapter 2); nevertheless, it is difficult to overstate the importance of carefully evaluating the validity of test scores, and new assessment instruments are, and should be, scrutinized for evidence to support the uses of scores for their intended purpose.

CIBS-II

The Comprehensive Inventory of Basic Skills-II (CIBS-II, Brigance 2009) is an example of a new achievement test, and its scores are intended to be used for multiple purposes. The CIBS-II is presented as an instrument useful for identifying students' level of performance, identifying and monitoring strengths and weaknesses, obtaining data to support referrals for further diagnostic testing, and monitoring and reporting student progress for individual educational plans (IEPs). Such uses of achievement test scores should be supported by evidence of the suitability of scores for those purposes. Since the CIBS-II is intended to comply with requirements of the NCLB act, as well as the Individuals with Disabilities Education Act (IDEA, 2004), validity studies are required. The present study is intended to provide some of the evidence needed to support the use of CIBS-II scores as indicators of students' progress in various academic domains. This study is part of the overall standardization and validation project for the instrument.

The CIBS-II is designed to be easily administered by school-psychologists, diagnosticians, or classroom teachers. Portions of the CIBS-II yield scores that can be interpreted as either criterion- or norm-referenced scores. As a criterion-referenced measure, the CIBS-II is designed to: (a) measure mastery of developmental and academic skills; (b) identify areas of strength and weaknesses; (c) serve as an indicator of student progress; and (d) assist in identifying goals and objectives for individual plans. As a norm-referenced measure, the CIBS-II is designed to: (a) meet state and federal assessment requirements for the identification of exceptional students for placement within special education services; (b) assess five areas of academic achievement (see below); (c) assess information processing skills in reading, math and written language

designated under the IDEA for the assessment of learning disabilities (IDEA, 2004); and (d) rapidly and briefly screen students to determine whether additional testing is needed.

Among the many subtests designed for 1st through 6th grade students, nine are included in the national standardization and validation study of the CIBS-II. (A separate instrument, the "Readiness Form," exists for kindergarten age students.) These subtests are designed to cluster into the following composites: Basic Reading skills (e.g., sight word vocabulary, phonetic analysis and phonemic awareness, survival sight words); Reading Comprehension, (e.g., reading vocabulary and passage comprehension); Mathematics (e.g., computation and math reasoning skills); Written Language (e.g., spelling and sentence writing); and Listening Comprehension (vocabulary and word-understanding separate from reading).

By presenting a composite score structure for the nine subtests in the standardized portion of the CIBS-II, the instrument's author has posited a latent structure for the instrument. Although the author and publisher have not presented any substantive or statistical explanation for the composite score structure, the nature of the score structure suggests that rather than measuring nine individual constructs or one general achievement construct, the nine subtests measure five broad constructs. In keeping with Standards 1.11 and 1.12 of the *Standards for Educational and Psychological Testing,* evidence supporting the composite score structure of the CIBS-II should be provided (AERA et al., 1999). This study is intended to seek confirming or disconfirming evidence as to the proposed composite score structure. The scores from the national standardization sample will be used to find evidence of essential unidimensionality (Stout, 2006) of subtests and

to investigate the composite score structure through the use of confirmatory factor analyses.

The next chapter will explore the history of achievement testing in the United States and the development of the modern unitary view of validity. Sources of evidence related to the internal structure of tests will be presented and briefly explained. The history of the CIBS-II will be presented, including a review of validity studies for previous versions of the CIBS. This background investigation will further establish the need for, and importance of, the present study.

Research Question

The overarching question in the present study is: To what extent do scores from the standardization sample of the CIBS-II support the composite score structure suggested by the publisher? Thus the study focuses on the internal structure of CIBS-II scores. The study will use the confirmatory factor analysis framework, and it will include an investigation of the dimensionality of subtest scores, which also addresses the internal structure of the scores.

CHAPTER TWO

REVIEW OF LITERATURE

The purposes of this chapter are to review the development of achievement testing in the United States, to summarize the history of validation studies in general, to provide an overview of the particular statistical tools to be used in this study, and to argue for the importance of the present study.

Achievement Testing

Early Years

The first group-administered achievement testing in the United States was implemented in the 1840s as an effort to monitor schools' effectiveness in Boston, Massachusetts. These test scores were intended to allow for comparison among schools and classrooms. In subsequent decades, such tests usually took the form of high school entrance examinations (Resnick, 1982), and the intended use of their scores changed from comparison of schools to identifying the most able students for placement in high school. Even though these tests were not administered to representative groups of students, they were used to compare schools on the basis of student achievement, which shows not only a long history of achievement testing but also a long history of questionable uses of test scores (Koretz & Hamilton, 2006).

World War I caused the next major wave of changes in standardized testing in the United States. The entry of the United States into the war created a massive increase in the size of the armed forces. The forces required an efficient way to classify recruits as being officer candidates versus infantrymen. Group tests were employed to measure the

intellectual abilities of recruits. This need to test ability ushered in the expansion of intelligence testing as schools began to use such tests to place students in homogenous ability groups (Koretz & Hamilton, 2006; Resnick, 1982). The first large scale tests designed as achievement measures appeared around the same time. The Stanford Achievement Tests were published in 1923 and the Iowa Tests of Basic Skills were developed in the 1930s. Both of these were expressly intended to measure student learning (i.e., achievement) over a broad range of content areas (Koretz & Hamilton, 2006). These early achievement measures were intended to help diagnose student academic needs so that teachers could adapt their instruction for their students (Resnick, 1982).

Origins of the Accountability Movement

Throughout these early years, and continuing into the 1950s, large-scale achievement testing was used mainly for student diagnostic and placement purposes and to monitor the academic performance of students in local jurisdictions, with little attention from the state or federal government (Koretz & Hamilton, 2006). However, the Soviet Union's launching of the Sputnik satellite in 1957 touched off a strong feeling of discontent with the United States school system (Popham, 1978). Among efforts to improve education, particularly in science and mathematics, Title I of the Elementary and Secondary Education Act of 1965 required measures to evaluate the law's effects. This provision led to the development of the National Assessment of Educational Progress (NAEP; NCES, 2000) and marked the first use of standardized assessments to monitor students' academic progress nation-wide (Koretz & Hamilton, 2006).

NAEP was initiated in 1965 as a program to assess the achievement of students in fourth, eighth, and twelfth grades. Initially, only selected item scores were reported, as opposed to test scores, and individual student scores have never been reported. By the late 1990s, NAEP had become more influential, with test scores reported at the state level and with federal education funding tied to state-level performance (Brennan, 2006).

Criterion Referenced Scoring

Minimum-competency testing was developed in the 1970s as the first large-scale example of holding students and teachers accountable for student performance (Popham, 1978). Minimum-competency testing, as implied by the name, was designed to measure whether students had reached a predefined level of competence (i.e., achievement). This shift in measurement led to the development of criterion-referenced measurement. Robert Glaser is credited with first contrasting norm-referenced versus criterion-referenced measurement in 1963 (Popham, 1978). In norm-referenced measurement, scores are used to determine examinees' standing relative to a standardization group, which is intended to be representative of the population of examinees. With criterion-referenced measurement, scores are used to evaluate an examinee's absolute level of attainment of criterion objectives.

Criterion-referenced measurement scores are most often used in one of two ways: (1) to determine what students know so that instruction can be tailored to their individual strengths and weaknesses and (2) to determine which students have attained mastery (e.g., for advancement or graduation) by comparing the students' scores to a predetermined cut-score. The first purpose led to the concept of measurement-driven instruction by using the test scores to shape instructional decisions (Popham, 1987). The

second purpose survives to this day in, for example, the standards-based reporting of NCLB (Koretz & Hamilton, 2006).

Accountability Revisited

The use of test scores to shape instruction, along with the continued prevalence of NAEP, led to increased attention to test scores, which contributed to widespread public dissatisfaction with the academic performance of United States students. The education-reform movement that swept the United States in the mid-80s eventually led to several states' enactment of standards-based test score reporting systems. Financial incentives (and sanctions) were put before schools and districts based on their scoring on state-mandated achievement tests. At the same time, several states began explicitly linking promotion between grades to exceeding a cut-score on the state achievement test. Associated with these developments was a shift away from minimum competency toward high expectations (Koretz & Hamilton, 2006).

Present Situation

In addition to a sharp increase in the amount of achievement testing that occurred in the 1990s and early 2000s, the characteristics and usages of the tests have changed as well. NCLB has played a large role in driving many of these changes. For example, under NCLB, fewer students are exempt from yearly achievement testing, students' scores are reported relative to targeted scoring levels (e.g., reaching or surpassing the "Proficient" standard), and a complex measure called Adequate Yearly Progress has been introduced to track performance of schools (Koretz & Hamilton, 2006).

NCLB instituted many requirements for state achievement testing. Within the defined regulations room exists for wide diversity. States were at liberty to define their

own academic standards and to design their own testing system. However, all NCLB testing revolved around comparing students' scores to standard performance levels of Basic, Proficient, and Advanced. Although every state is required to use, at a minimum, these three performance levels, each state defines its own criteria for the standards, which results in wide differences in the meaning of performance levels across states. Most states use a criterion-referenced type approach in which a standard-setting study (Cizek & Bunch, 2007) is used to set cut-scores for each labeled proficiency level. Other states set their standards based on norm-referenced scores: the cut score for a level is based on achieving a certain percentile score relative to a specified standardization of the test.

Outside the realm of mandated state-level achievement testing, several other types of commercially produced achievement tests continue to enjoy wide use, including content area surveys, academic area achievement and diagnostic assessments, and special education diagnostic assessments (Ferrara & DeMauro, 2006). The most widely used content area surveys generally are used to describe a student's performance across a wide range of content areas, such as mathematics, reading skills, reading comprehension, writing skills, social studies, etc. Score reporting is typically norm referenced and based on nationally representative samples. Some such assessments also include performance level information in tandem with percentile scores.

Academic area achievement and diagnostic assessments are less closely aligned with specific grade-level academic content and instead assess students' achievement in rather broad academic areas (e.g., computation, written expression) and are specifically intended to report individual students' strengths and weaknesses. Such assessments are

usually selected by individual school districts or schools and are not intended for group reporting of scores.

Similarly, special education diagnostic assessments are intended to identify special education students and track the progress of these students. These diagnostic assessments are used to determine the existence of disabilities in students, to plan educational services and prepare instruction, and to provide ongoing evaluation of their progress in schools. Special education students' individualized education plans (IEPs) are sometimes specified in relation to progress on these diagnostic assessments.

Validity and Validation

The term *validity* as applied to tests and test scores has varied widely since its introduction into educational testing in the early part of the 20th Century, and its meaning is still studied, argued, and often misunderstood to the present day (Cizek, Rosenberg, & Koons, 2008; Hogan & Agnello, 2004). The most general sense of the validity of test scores is to ask: "What is the meaning of these test scores?" Often, especially early in the development of validity theory, this question was posed as: "Does this test measure what it is purported to measure?" For adherents to modern validity theory, the question typically becomes: "Is the intended interpretation of these test scores defensible?" or "Does empirical evidence and theoretical rationale support the intended inferences that are to be drawn based on these test scores?"

Conceptions of validity and validation have evolved tremendously over the years from the 1920s to the present. Messick (1989) presents a thorough account of the many transitions that validity theorists passed through in the transition from the focus on distinct types of validity to the current unitary validity concept. The following summary

is intended to highlight the types of validity evidence that have been the main focus of theorists through the years.

Criterion Validity and Content Validity

The earliest attention to the validity of test scores was in the form of criterion validity studies of achievement tests developed in the 1920s. Criterion validity became the predominant manner in which validity was defined through the 1930s and 1940s (Kane, 2006). The criterion model of validity has two versions: predictive validity and concurrent validity. Predictive validity referred to the extent to which the instrument predicted performance on measurements (e.g., achievement) in the future, while concurrent validity indicated the relationship between the instrument's scores and scores on other measurements given at the same time (Crocker & Algina, 1986). For early validity theorists, the goal of measurement was to estimate as accurately as possible the value of some criterion variable, so validity specifically referred to the relationship between test scores and criterion scores (Kane, 2006). Criterion-related validity is established in terms of correlations between test scores and criterion scores or by regressing criterion scores on test scores. However, a validity argument based on students' scores on a criterion measure is only as strong as the validity argument for the criterion measure, and although criterion validity appears to be objective and purely quantitative, it depends on the subjective value judgment of what criterion to specify (Kane, 2006).

The model of content validity also developed in these years as a means of validating the criterion measures. The content model of validity uses the idea of domain sampling: test scores represent a sample of performance in the domain of interest. A

content valid test should elicit a broad and representative sample of the examinee's performance in the domain. This representative sample is used to estimate the examinee's overall level of skill or achievement in that area (Kane, 2006). Content validity is established solely on the basis of expert judgment about the content of the test and does not take into account actual responses. These judgments do not provide any support for inferences to be made from test scores; such interpretations of the meaning of test scores lack any justification (Messick, 1989).

Construct Validity

By the 1950s, criterion-related validity was broadly accepted, as was content validity to help justify the use of the criterion measures (Kane, 2006). Construct validity emerged as a third type of validity in the mid-1950s. Rather than supplant other views of validity, construct validity came to sit alongside them (Messick, 1989). Construct validity indicated the extent to which the instrument's scores allowed meaningful inferences about some psychological construct (Crocker & Algina, 1986). Construct validity originated out of personality testing, where no obvious criterion existed. In this model of validity, the test developer begins with a theory about the existence of a construct, rather than a criterion, and uses that theory to devise measures. Validation of a test under this model "is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (Messick, 1989, p. 17).

The Unified Model of Validity

Although the construct model of validity began as an alternative model when no suitable criterion was available, it was soon recognized as the fundamental idea of validity. Loevinger (1957, as cited in Kane, 2006) is credited as the first psychometrician

to put forth the idea of what would become the unified model of validity when she stated that the criterion and content models were means to get at the goal of construct validity. Through the 1960s and 1970s, validity continued to be widely viewed as a set of methods to choose among depending on the nature of the test. At the same time, validity theorists continued to develop and argue for a unified approach, in which different "types of validity" are viewed as types of evidence of construct validity. By the early 1980s, the unified point of view was gaining greater acceptance. With Messick's 1989 chapter on validity in the third edition of *Educational Measurement*, the construct validity model was authoritatively put forth as the unifying concept of all test validation.

From 1989 to the present, this unified approach to validity has been promulgated and extended but rarely contradicted (cf., Borsboom, Mellenbergh, & van Heerden, 2004; Lissetz & Samuelsen, 2007). Messick's (1989) definition of 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; emphasis in original) stresses that validity applies to inferences, not to tests or even test scores; that validity is a judgment; and that validity arguments rely on multiple sources of evidence. Although modern validity theorists nearly unanimously accept the unified approach to validity, practitioners of testing have "implicitly rejected" important aspects modern validity theory by continuing to present validity studies more in keeping with the view that validity is a property of the test and that different *kinds* of validity can be used to support the validity of a test (Cizek et al., 2008, p. 409).

Modern validity theory emphasizes test *validation* rather than test *validity*. The generally accepted view is that one validates *interpretations* or *uses* of tests. Validation is a process; it is "the development of evidence to support the proposed interpretations and uses," that is, "to show that [the proposed interpretation or use] is justified" (Kane, 2006, p. 17). The *Standards for Educational and Psychological Testing* (AERA et al., 1999) propose five sources of validity evidence, including evidence based on: (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. The authors of the *Standards* stress that while the different sources highlight different aspects of test validity, they do not represent different *types* of validity. The kind of evidence collected should depend on the proposed interpretation. For example, if scores on a subtest are interpreted as a unidimensional measure of arithmetic achievement, validation should include collecting evidence as to the internal structure (i.e., the dimensionality) of the subtest as well as evidence that the content of the subtest is representative of the arithmetic content the examinees have had an opportunity to learn.

Evidence of the internal structure of the test is especially relevant for a collection of items (or subtests) that is purported to allow measurement of a few broad constructs. For example, a battery of achievement tests may include ten or more subtests yet report scores on such broad constructs as reading achievement, mathematics achievement, and science achievement. In such cases, composite scores are often reported, meaning that some combination of subtest scores are used to produce a composite score for a broad content area (i.e., construct). The broad composites can be interpreted as *latent variables* or *factors*, that is, they can be conceived of as unobservable abilities or traits that can

only be measured indirectly by means of observable indicators (the subtest scores) (Thompson, 2004). When a particular latent structure is proposed for a set of scores, that structure is one interpretation of the scores and should be subject to validation. Evidence to support (or refute) such a structure can be collected via factor analysis, which is described below.

Data Analysis to Support Validation

Confirmatory Factor Analysis

The factor structure of an instrument can be examined with confirmatory factor analysis (CFA), a theory-driven analysis requiring specification of the relationship of indicators to underlying traits. In the context of validation studies, CFA can be used to assist in understanding the internal structure of a test and to provide evidence in support of a proposed interpretation of the score structure. Rival hypotheses (i.e., alternative models) can be tested within the CFA framework, which can lead to stronger evidence of validity (Thompson & Daniel, 1996). By testing rival models one can investigate alternative interpretations of test scores. If a posited model is found to be a more plausible interpretation than rival models, then that model gains credence (Kane, 2006).

A history of the association between validity studies and factor analysis has been reported in some detail (e.g., Thompson, 1997; Thompson, 2004; Thompson & Daniel, 1996). Factor analysis is intended to model the relationship between latent constructs, or factors, and observed variables, or indicators. Latent constructs are unobserved and thus cannot be measured directly. But observed variables are influenced by the latent constructs, and thus indicate something about the number and nature of the latent constructs (Brown, 2006). More specifically, factor analysis techniques allow researchers

to analyze the covariance between indicators and separate out common variance, that part of the variance that is influenced by a common factor (or factors), versus unique variance.

In a CFA study, the researcher posits a theory-based model and investigates how well the data fit that model. The fit of data to the model can support, disconfirm, or suggest changes in a theory. Although exploratory factor analysis (EFA) and CFA both are concerned with how observed variables are linked to latent variables, EFA takes an exploratory approach to generate possible models when the links are unknown. CFA, however, is appropriate when a theoretical model is suggested *a priori* (Byrne, 1998).

It is possible to use CFA in an exploratory manner by respecifying models without regard to underlying theory; however, such an approach can lead to capitalization on chance (Keith, 2005). Rather than attempting to interpret a structure implied by the data, CFA is intended to test the fit of data to a structure that follows from theoretical considerations. CFA requires that constructs are defined *before* testing a model (Graham, Guthrie, & Thompson, 2003). Whereas exploratory methods (e.g., EFA) use the data to *create* a model, CFA requires the researcher to explicitly define how indicators are hypothetically linked to underlying constructs. These hypothetical links can then be supported or disconfirmed—partly in an absolute sense (i.e., does the model fit?) and, even more so, in a relative sense (i.e., does the model fit better than other defensible models?).

Analysis of the fit of a model can lead a researcher to consider alternative interpretations of the scores. Such investigations can be a valuable part of developing an understanding of an instrument's structure. However, interpretations derived from such

an analysis should be validated by fitting an independent sample of test scores to the new proposed model (MacCallum, 1995).

How CFA Works

The CFA model relates observed (x) variables to latent constructs (ξ) using a linear model:

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}\,. \tag{1}$$

In this equation, $\mathbf{x}$ is the vector of observed variables, $\mathbf{\Lambda}$ is the matrix of factor loadings, (or, more precisely, factor pattern coefficients), $\boldsymbol{\xi}$ is the vector of factors, and $\boldsymbol{\delta}$ is the vector of error terms for the indicator scores (or, more precisely, unique components). Estimation of parameters in the CFA model is implemented using the covariance structure of the data. That is, the covariance matrix of the observed variables, $\mathbf{\Sigma}$, is modeled as

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}\,, \tag{2}$$

where $\mathbf{\Phi}$ is the matrix of factor variances and covariances and $\mathbf{\Theta}$ is the matrix of indicator error variance and covariance.

The CFA model also can be represented graphically as shown in Figure 1. In such representations, referred to as path diagrams, latent constructs are represented with ellipses or circles, and observed variables are represented with rectangles or squares. An arrow leading from a latent construct to an observed variable indicates that the construct is presumed to influence the variable. Latent constructs are unobserved and cannot directly be measured, so the observed variables are often referred to as indicators, to convey the notion that it is through the observed variables that we indirectly measure the

latent construct. Double headed arrows between two features in the diagram indicate that the covariance between those terms is estimated as part of the model.

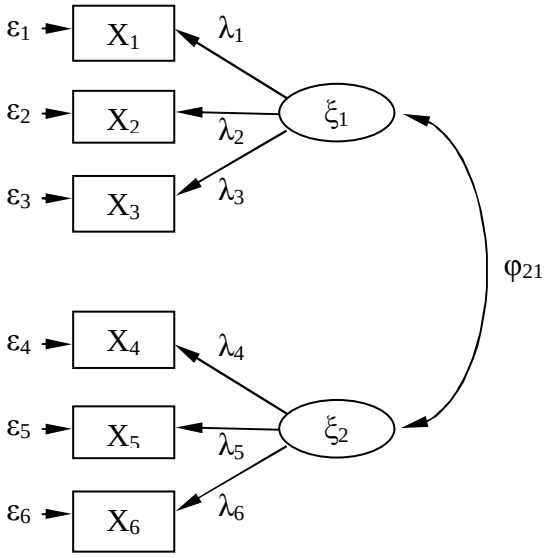

Figure 1. A simple path diagram for a two-factor model.

Usually, error terms are assumed to be independent, since common variance between variables is reflected in the latent construct. However, if two variables share variance that is not reflected in the model (i.e., that is not presumed to be reflected in the latent construct), then including the covariance of the error terms for those variables may improve the fit of the model (Kline, 2005).

Most CFA models assume simple structure, in which each indicator is associated with exactly one factor. Having a link between an observed variable and more than one factor is referred to as a cross-loading. Although most CFA models are used to test theories in which a model is specified to have simple structure, it is often misleading to regard that simple structure as removing all relationship between indicators and the

factors to which they are not linked. Some of the common variation between indicators of differing factors is captured in the covariance between factors (Brown, 2006), and factor structure coefficients should be calculated to measure the correlation of the indicators with the factors (Thompson, 1997). When factors are not correlated, structure coefficients are simply the pattern coefficients, but when factors are correlated, the structure coefficients reveal the association between factors and the indicators to which they are *not* linked. Analysis of structure coefficients in addition to the pattern coefficients ($\mathbf{\Lambda}$) can also be illuminating in examining relationships (Graham et al., 2003).

When data are fit to a properly specified model and parameters are estimated (estimation methods will be discussed in the Methods section), the output includes parameter estimates and standard errors, which can be used to derive a *t*-value to test whether a parameter is significantly different from zero, as well as model-fit statistics, which are used to judge how well the model fits the data. Good fit lends support to the theory that led to the model. However, as discussed earlier, models are best judged by comparison with theoretically defensible competing models (Thompson, 2004).

The estimated loadings can be interpreted as the strength of association between a factor and an indicator. High loadings of a set of indicators on a factor provide evidence that the indicators are associated with the factor. In addition, a collection of fit indices is produced when a model is estimated. Fit indices are produced by comparing the covariance matrix of the data to the covariance matrix implied by the specified model. Good fit of the data to the model is further evidence that the proposed model is a plausible interpretation of the test scores (Kline, 2005). Model fit statistics will be discussed in the Methods section.

Higher-order factor analysis can be used to account for correlations among first-order factors and should be investigated under such correlated solutions (Thompson, 2004). If latent constructs are strongly correlated, then perhaps a second-order factor can be specified as a common influence on the latent constructs (Brown, 2006). As with all model specification, such relationships between constructs should be theory based, as opposed to purely data-driven.

Higher-order factor analysis in the LISREL notational scheme (Jöreskog & Sörbom, 1996) requires that the model be written in terms of *y*-variables. The observed indicators (**Y**) are influenced by the first-order factors (**η**) as reflected in the equation:

$$\mathbf{Y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}. \quad (3)$$

But the first-order factors (**η**) are influenced by second-order factors (ξ) as reflected in the equation:

$$\boldsymbol{\eta} = \mathbf{\Gamma}\xi + \zeta, \quad (4)$$

where **Γ** is the matrix of second-order factor loadings and **ζ** is the vector of "error" in **η**. This "error" vector is more properly conceived of as unique variance in the first-order factors. Higher-order factor analysis is common in intelligence testing, with a second-order *g*, or "general intelligence," factor that is influenced by multiple first-order factors (e.g., Keith, Fine, Taub, Reynolds, & Kranzler, 2006).

Dimensionality and DIMTEST

A test that is designed to measure exactly one trait or ability is said to be *unidimensional*. The dimensionality of a test is closely related to its internal structure: a test designed to allow inferences about a certain number of traits or abilities should measure the same number of dimensions. If a proposed interpretation of scores involves

assumptions about the dimensional structure of the scores, then the assessment of the test scores' dimensional structure is an important part of the validation of that interpretation. Confirming the unidimensionality of test scores is important in validation efforts for at least three reasons: (a) to assess whether the measurement of a trait is being contaminated by the measurement of a second trait, (b) to help determine whether a test is measuring multiple traits and should be divided into separate subtests for interpretation, and (c) to evaluate the suitability of the scores for analysis that rely on the assumption of unidimensional scores (e.g., CFA) (Stout, 1987).

DIMTEST is a nonparametric procedure to test the hypothesis that a test is *essentially unidimensional.* The concept of essential unidimensionality recognizes that it is exceedingly rare for a test to truly measure one and only one dimension, but that it is possible for only one dimension to be seen as important or interpretable (Nandakumar, 1991). DIMTEST works by examining two partitions of the test items: AT, the assessment subtest, and PT, the partitioning subtest. The AT includes items that are known, or hypothesized, to be dimensionally distinct from the items in PT. The DIMTEST procedure then calculates a statistic to test the null hypothesis that the AT set is dimensionally similar to PT versus the alternative hypothesis that the subtests are dimensionally distinct (Stout, 2006). Because the DIMTEST statistic is known to be statistically biased (Stout, 1987), a nonparametric IRT bootstrap based bias correction for the DIMTEST statistic has been incorporated into the DIMTEST procedure (Froelich & Stout, 2003). The corrected DIMTEST statistic has a standard normal distribution, and a statistically significant result is evidence that the test is not essentially unidimensional

(Stout, 2006). The DIMTEST procedure is implemented as a component of the DIMPACK nonparametric dimensionality analysis software package (Stout, 2006).

The DIMTEST procedure rests on the notion that the covariance of pairs of items, conditioned on estimated examinee ability, should be small when a test is essentially unidimensional. To calculate the DIMTEST statistic, the PT is used to estimate the examinee ability vector for the test scores. The conditional covariance for each pair of items in AT is calculated, conditioning on PT subtest score, and these covariances are combined in the DIMTEST statistic, first presented by Stout (1987) and summarized as follows by Finch and Habing (2007):

$$T^* = \sum_{i<l \in AT} \int_{-\infty}^{\infty} \mathrm{Cov}(U_i, U_l \mid \hat{\theta}_{PT}) d\hat{\theta}_{PT} \ . \qquad (5)$$

Finally, the positive bias in T^* is corrected with a nonparametric IRT bootstrap procedure described in Froelich and Stout (2003). The final DIMTEST statistic provides a statistical test for the null hypothesis that the conditional covariances of AT items are small enough to conclude the test is unidimensional.

The reader may wonder why linear exploratory factor analysis (EFA) is not used to assess dimensionality compared to a more complex and time-intensive analysis. Historically, linear EFA has been used to assess the dimensionality; however, factor analysis is problematic as a method of determining dimensionality. Item difficulty can be confounded with dimensionality. If the relationship between item performance and latent ability is nonlinear, poor model fit can result, which can prevent the analyst from drawing conclusions. In addition, factor analysis with dichotomous data is complicated, and models can be difficult to estimate (Ackerman, Gierl, & Walker, 2003).

Dimensionality studies are not common among validity studies, but they have been used in a variety of ways to investigate the internal structure of scores. Prior to the development of DIMTEST, dimensionality was investigated using a principal component analysis and multidimensional scaling software to assess the effect of alternative scoring methods on the psychometric properties of computation items (Birenbaum & Tatsuoka, 1983). DIMTEST was used to provide evidence for the construct validity of a set of items intended to assess international students' speaking anxiety by showing that scores on two sets of items were not dimensionally distinct (Yang, 2006). Dimensionality studies using DIMTEST have provided evidence of differential item functioning (DIF) (Metcalf, 2002) and have provided statistical, as well as substantive, corroboration of DIF findings from other methods (Gierl, Bisanz, Bisanz, & Boughton, 2003).

DIMTEST was developed to fill the need in IRT analysis for a statistical significance test of the unidimensionality of test items (Stout, 1987). However, this procedure may also prove useful in assessing reasons for lack of fit in factor models of test scores. In most CFA studies, the scores on indicators linked to a single factor are assumed to be unidimensional. Indeed, simple structure specifically depends on the assumption of unidimensional measurement, and relaxing that specification (i.e., allowing some indicators to "cross load" on two or more factors) has been controversial in measurement literature (Kline, 2005). Lack of fit in any given factor model can stem from multiple sources, but the unidimensionality of indicators is rarely explicitly checked. Instead, researchers analyze models and the details of misfit to guide their respecification of models (Brown, 2006; Kline 2005).

CIBS-II

The focus of the present study is the Brigance Comprehensive Inventory of Basic Skills-II (CIBS-II; Brigance, 2009), the newest revision in the Brigance inventories series. The previous version of the CIBS was commonly used as a screening tool and for monitoring the progress of students, particularly in special education programs. This study took place as the technical manual and full testing materials were under development; thus, no previous validation studies of the CIBS-II or its scores exist. A review of the precursors of the CIBS-II is relevant, as such a review can shed light on the new instrument and its development. In particular, research into the validity of scores from precursors to the CIBS-II might guide new validity studies. Studies of previous versions of the instrument could, for example, suggest a starting point for new exploratory or confirmatory studies.

The Brigance CIBS-II began as an effort by A. H. Brigance to develop a criterion-referenced system of assessments for special education students (Brigance, 1998). In his work as a special-education teacher and administrator, Brigance found that norm-referenced scores from typical achievement tests did not yield information useful for planning individualized instructional programs for students. Brigance's first published instrument was the *Inventory of Basic Skills* (Brigance, 1976). This first inventory soon led to the *Diagnostic Inventory of Basic Skills* (DIBS; Brigance, 1977), which was expanded and modified to become the *Comprehensive Inventory of Basic Skills* in 1983. The *Comprehensive Inventory of Basic Skills-Revised* (CIBS-R) was published in 1998. This major revision included an update of many assessment items to reference then-current textbooks and the introduction of standardized, norm-referenced, score reporting.

The CIBS-R test materials include a bibliography listing the student textbooks and professional publications used in the development of the CIBS-R (Brigance, 1998).

The introduction of norm-referenced score reporting extended the proposed uses of the CIBS-R. The CIBS-R was presented as "a valuable resource in school programs emphasizing individualized instruction. The CIBS-R will be especially helpful in programs serving students with special needs" (Brigance, 1998, p. ix). The test materials state that the CIBS-R components may be used for identification of skills mastered and not mastered, as a diagnostic instrument to identify strengths and weaknesses, as a part of a testing regimen to identify students with special needs, and "as a standardized testing instrument when needed" (p. ix). The claim was made that "the assessments are based on curriculum content and objectives" (p. x) and tied to the content and sequence of common elementary school textbooks. Skill sequences and grade level expectations were reportedly based on what was found in researching multiple textbooks from different publishers (Brigance, 1998).

From the earliest development of the inventory, the subtests were written to reflect the grade-level content that appears in textbooks used in elementary schools (Connelly, 1985; Brigance 1998). No information was given in the CIBS-R test manual to indicate how the author ensured adequate content domain representation; however, in its original form, the CIBS was purported to be useful "as a scope and sequence, and [educators] may consider or choose specific objectives from it" (Connelly, 1985, p. 4). Other studies of the original CIBS include a mention of "field testing and the jury system" being used to establish the content validity of the instrument (Linkoas, Enright, Messer, & Thomas, 1986; p. 6).

Few studies exist to investigate the validity of scores from any of the precursors to the CIBS or early versions of the CIBS. As early as 1983, this lack of validity evidence was noted: "The test author, while explaining how grade levels were derived, provided no statistical data justifying the procedure or verifying the validity of the test scores" (Krawiec & Spadafore, 1983, p. 230). The work completed in those few early studies focused on content validity (e.g., Ferguson & Kersting, 1988). The CIBS was intended only for instructional decision making, as opposed to being used for prediction or educational placement. Since no claim was made as to a construct being measured by the CIBS subtests, no apparent need existed for construct validity studies.

With the CIBS-R, a norm-referenced interpretation of some subtest scores was added. The combination of norm-referenced and criterion-referenced interpretations was meant to facilitate the movement from interpreting scores in a normative fashion (e.g., for determining eligibility for special-education programs) to interpreting scores in a skill-based manner (e.g., for determining objectives to include in an individualized education program) (Glascoe, 1999a). Norm-referenced interpretation of scores facilitates comparisons of students. Norm-referenced scores are directed toward a student's relative standing. Criterion-referenced score interpretation can complement information about relative standing by providing information about exactly what skills students have or have not been achieved (Popham, 1978).

At the same time that norm-referenced interpretation of scores was introduced, the test included a composite score structure (see Table 1). Such a structure implies that a group of constructs are being measured by the subtests. The technical manual for the CIBS-R (Glascoe, 1999b) uses correlations between subtests and assessments as evidence

of construct validity; however, this evidence is inadequate, as correlations do not necessarily support the intended interpretation of the scores (Cizek, 2001). Further, relying only on reliabilities of subscales and correlations among subscales without investigating the dimensionality of the scales can lead to erroneous conclusions about the structure of a test (Green, 2007).

Table 1

Composite Score Structure of the CIBS-II Subtests

Subtest	Composite
Word Recognition Grade Placement Word Analysis Survey	Basic Reading
Reading Vocabulary Comprehension Grade Placement Comprehends Passage	Reading Comprehension
Computational Skills Grade Placement Problem Solving Grade Placement	Math
Spelling Grade Placement Sentence Writing Grade Placement	Written Expression
Listening Vocabulary Comprehension Grade Placement	Listening Comprehension

Aside from the CIBS-R technical manual published literature shows a complete lack of studies into the validity of the composite scores for this instrument or even of its subtest scores. Recall that evidence based on the internal structure of the test is one of the five main sources of validity evidence recommended in the *Standards* (AERA et al.,

1999). Indeed, Standard 1.11 addresses this point directly: "If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided" (p. 20). Continuing, the comment section for the standard recommends that interrelationships of a test's subtest scores "should be shown to be consistent with the construct(s) being assessed" (p. 20).

Overview of the Present Study

Validity studies of new or revised instruments are common. Even small changes in an instrument may have unpredictable impact on its psychometric properties, which necessitates validation of the revised instrument (AERA et al., 1999). Achievement tests are subjected to scrutiny since inferences made from the scores on achievement tests can have high-stakes impact on examinees. As explained previously, inferences made on the basis of composite scores are in particular need of validation. The paucity of studies providing evidence of the validity of composite scores from earlier versions of the CIBS emphasizes the need for validity studies of CIBS-II scores. Results from a CFA study of the internal structure of scores from the CIBS-R standardization sample did not support the theoretical structure proposed by the test's author (Breidenbach & French, 2008). The composite score structure of the CIBS-II is very similar to that of the CIBS-R; however, the CIBS-II standardization sample is larger and more representative of the intended audience than that of the CIBS-R. A careful validation study of the internal structure of the CIBS-II will provide important evidence to support interpretations of the test's subscores, both individually and combined in composite scores.

Factor analytic studies of achievement instruments are uncommon. Validity studies for achievement tests typically do not seek to support the factorial structures of the tests (e.g., Daub & Colarusso, 1996; Connolly, 1998; Erford & Dutton, 2005). Even though achievement test results are commonly used to make high stakes decisions, "there is surprisingly little published evidence that supports the structure of such instruments and the validity of their intended use and interpretation" (Stevens & Zvoch, 2007, p. 977). In two exceptions to this general rule, researchers used factor analysis to investigate the structure of achievement tests and could not find support for the structure described by the tests' authors (Erford & Klein, 2007; Williams, Fall, Eaves, Darch, & Woods-Groves, 2007). In addition, a confirmatory factor analysis (CFA) investigation of part of the TerraNova assessment system (CTB/McGraw Hill, 1997) found that the internal structure was not as clearly defined as the publisher suggested (Stevens & Zvoch, 2007). Comparison of two- and three-factor models found little difference in fit, suggesting that some content areas are not well represented in the test structure. Such studies emphasize the point that if composite scores are to be reported and used in interpreting students' results, factorial validity evidence must exist to support such use (Williams et al., 2007).

The current study uses the CIBS-II national standardization sample to investigate the factor structure of CIBS-II subtest scores. The dimensionality of the subtest scores are investigated with DIMTEST, and the CIBS-II's theoretical structure is examined via confirmatory factor analysis. The theoretical structure is tested along with competing models that were derived from substantive examination of the content of subtests and from results of the DIMTEST dimensionality analysis.

CHAPTER THREE

METHODOLOGY

Participants

Data for this study comes from the national standardization study for the Comprehensive Inventory of Basic Skills-II (CIBS-II; French & Glascoe, 2009). The goals of the national standardization study were to develop a sample representative of the United States population of school children in grades K–6, administer nine of the CIBS-II subtests to the children in the sample, and use their scores to (a) investigate the psychometric properties of the test items and subtest scores and (b) develop standard scores and normative tables.

Teachers from the four geographic regions (Northeast, Midwest, South, and West) in the United States were recruited to administer the assessments to small numbers of students. Teacher selection was guided by geographic region, grade level taught, and accessibility. That is, teachers were selected based on ability of the study organizers to access (usually via e-mail) the teachers. Participating teachers were given directions to select students with a wide variety of achievement levels. The specific instructions guided teachers via an example: “If, for example, you select 6 children please select: 1 child whom you believe is performing above average, 4 who are performing averagely, and 1 who is performing below average.” Participating teachers were paid $30 for each completed test form.

The data collection for this project was implemented at Purdue University. The university’s Institutional Review Board ruled that the study was exempt from informed

consent requirements because administering the CIBS-II is considered standard educational practice and because data was collected with no personally identifiable information. Nonetheless, parents were given the chance to opt their child out of the study with a brief permission letter that was included with an optional parent survey. The parent survey asked for demographic information about the child and family and invited parents to share concerns over a variety of topics related to the child's educational development.

A full description of the standardization sample is given in the test manual (French & Glascoe, 2009), and a brief summary is presented here. The sample (N = 1,411) matches closely the U.S. population on a number of important demographic variables (e.g., age, race/ethnicity, geographic region) as reported in the U.S. Bureau of the Census projections for 2007 and the U.S. Department of Education's National Center for Education Statistics (Hussar & Bailey, 2006). Geographic distribution, gender, age, and racial/ethnic categories are reported in Tables 2–5. To allow for model cross-validation, records were randomly assigned to two subgroups of 706 and 705 participants. No effort was made to match the subgroups on any variable in order to more closely approximate independent random samples.

Table 2

Distribution of Examinees by Region, Compared to U.S. Population

Region	Sample		U.S.
	N	%	%
Midwest	502	36	22
Northeast	164	12	17
South	445	32	37
West	300	21	24
Total	1411		

Note. U.S. distribution is based on Hussar & Bailey, 2006.

Table 3

Distribution of Examinees by Gender and Region

Region	Female	Male	No report	Total
Midwest	226	251	25	502
Northeast	82	82		164
South	213	232		445
West	159	139	2	300
Total	680	704	27	1411

Table 4

Distribution of Examinees by Race/Ethnicity, Compared to U.S. population

Region	White	African-American	Hispanic	Asian	Other	Multiple	Missing
Midwest	42%	26%	15%	3%	2%	4%	7%
Northeast	88%	4%	2%	3%	2%	1%	1%
South	42%	24%	25%	1%	2%	4%	1%
West	45%	10%	25%	3%	6%	8%	3%
Total sample	48%	20%	19%	2%	3%	4%	4%
U.S. Population	60%	15%	18%	4%	1%	2%	

Note. U.S. distribution is based on Hussar & Bailey, 2006.

Table 5

Distribution of Examinees by Age and Region

	Region				
Grade	Midwest	Northeast	South	West	Total
1	82	61	98	74	315
2	79	42	60	42	223
3	50	14	31	78	173
4	63	13	102	37	215
5	87	28	104	21	240
6	141	6	50	48	245
Total	502	164	445	300	1411

Instrument

The grades 1–6 form of the CIBS-II includes more than 150 subtests, which range from teacher checklists to performance tasks to subtests composed of dichotomously scored items. Nine dichotomously scored subtests (described below) were selected by the instrument's publisher to be standardized. Throughout this study, the general name "CIBS-II" refers to the nine subtests selected by the publisher. The CIBS-II scales are intended to be administered individually but do not require specialized training. Children respond orally or on student response pages, and scoring is marked on the corresponding teacher pages, which includes the answer key.

The nine subtests are designed to cluster into four composites and one indicator. A brief description of each composite and its associated subtests follows:

Basic Reading Composite

1. *Word Recognition Grade Placement Subtest:* Children are asked to quickly read aloud words arranged into lists by grade level, from preprimer to grade eight.

2. *Word Analysis Survey*: Children are asked to respond to auditory discrimination items (i.e., respond "yes" or "no" to indicate whether two words read by the test administrator sound exactly the same), to identify sounds heard in words read aloud by the test administrator, to read aloud words and nonsense words to sample phonemic awareness, and to divide words into syllables.

Reading Comprehension Composite

3. *Reading Vocabulary Comprehension Grade Placement Test*: Children indicate single-word vocabulary comprehension of printed words by choosing the one word with a different meaning from groups of five words each.

4. *Comprehends Passages Subtest*: Children read a short passage between primer and grade nine levels and answer five oral-response, multiple-choice questions about the passage.

Math Composite

5. *Computational Skills Grade Placement Test*: Children solve arithmetic problems involving addition, subtraction, multiplication, division, fractions, and percentages. Problems are arranged by grade level (grades one through eight).

6. *Problem Solving Grade Placement Test*: Children solve arithmetic word problems. Problems are read aloud to grades 1–3 students while the student reads a printed version. Grades 4–6 children read the printed problem with assistance if needed.

Written Expression Composite

7. *Spelling Grade Placement Test*: This subtest is a standard written spelling test with words arranged in first-grade through eighth-grade lists.

8. *Sentence Writing Grade Placement Test*: Children are given three (grades 1–3 level) to four (grades 4–6 level) words and attempt to compose a single sentence using the words.

Listening Comprehension Indicator

9. *Listening Vocabulary Comprehension Grade Placement Test*: This subtest is not designed to correspond to any of the composites. Children indicate single-word vocabulary comprehension of words read aloud by the test administrator by choosing the one word with a different meaning from groups of four words each.

All subtests are presented in two forms. The forms are intended to be parallel and allow for pre- and post-testing of students without score inflation due to practice effects.

The parallel forms also allow test administrators to re-test a student if an initial test administration is invalidated for any reason.

Variables

Subtest raw scores are simple sums of the number of correct responses. Many subtests have entry and discontinue rules (i.e., basal and ceiling rules). Items before the entry rule are scored as correct, and items after the discontinue rule are scored as incorrect. Composite scores are simple sums of the subtests scaled scores corresponding to the composite.

Statistical equating (Kolen & Brennan, 2004) was used to adjust scores for differences in the difficulty of the two forms of the subtests (French & Glascoe, 2009). All Form B raw scores were transformed to Form A equivalents. After equating, the adjusted raw subtest scores were normalized by age category and scaled to have a mean of 10 and a standard deviation of 3. Composite scores were created by summing the subtest scaled scores associated with the respective composite and then reported as standard scores, with mean 100 and standard deviation 15. Descriptive statistics for the subtests, including reliability estimates, are presented in Table 6.

Analysis

Dimensionality of Subtests

DIMTEST (Stout, 2006) was used to investigate the dimensionality of each subtest. Since it is exceedingly rare for real scores to show true unidimensionality, DIMTEST provides a statistical test of *essential unidimensionality* (Nandakumar, 1991). Scores that are essentially unidimensional may have one or more minor dimensions, but these minor dimensions are unimportant and not interpretable.

Table 6

Descriptive Statistics for CIBS-II Subtest Scaled Scores

	M	*SD*	Reliability	Skew	Kurtosis
Comprehends Passages	10.372	2.670	. 974	-0.028	-0.303
Computation	9.709	2.197	.925	-0.019	-0.305
Listening Vocabulary	10.002	2.268	.894	0.024	-0.357
Problem Solving	9.536	2.136	.859	0.030	-0.275
Reading Vocabulary	9.627	2.397	.921	0.079	-0.286
Sentence Writing	9.860	2.510	.807	0.058	-0.351
Spelling	10.035	2.121	.960	0.116	-0.535
Word Analysis	10.372	2.670	.955	-0.028	-0.303
Word Recognition	9.709	2.197	.987	-0.019	-0.305

DIMTEST is a nonparametric procedure to test the hypothesis that a test is essentially unidimensional. The DIMTEST procedure uses scores from a two-group partition of the test items. The assessment subtest (AT) consists of items that are presumed to be dimensionally distinct from other items in the test. The partitioning subtest (PT) consists of all test items not in AT. The DIMTEST procedure then calculates a statistic to test the null hypothesis that the conditional covariances of AT items are small, which indicates that the AT set is dimensionally similar to PT, versus the alternative hypothesis that the AT set is dimensionally distinct from PT. A statistically significant result is evidence that the test is not essentially unidimensional (Stout, 2006).

DIMTEST can be performed in a confirmatory sense if a set of items are suspected to be dimensionally distinct from the remainder of the test. Alternatively, DIMTEST can be used in an exploratory manner by using a statistically based

partitioning method. Performance of DIMTEST to detect multidimensionality depends greatly on the choice of AT. Early implementation of DIMTEST recommended the use of linear factor analysis or expert content analysis to choose dimensionally distinct items for the AT (Stout, 1987); however, the method of using factor analysis has been shown to perform poorly in many situations, and content analysis can sometimes fail to detect statistical dimensions (e.g., a dimension related to speededness) that are not apparent in the items' content (Froelich & Habing, 2008).

For the present study, a method developed by Froelich and Habing (2008) that relies on conditional covariance-based cluster analysis (CCPROX/HAC; Roussos, Stout, & Marden, 1998) to identify potential AT partitions coupled with the DETECT procedure (Kim, 1994; Zhang & Stout, 1999) was used to select the "best" of the potential AT sets (i.e., the AT set with the greatest DETECT index was used to calculate the DIMTEST statistic).

CCPROX/HAC is a two-step procedure to produce a hierarchical cluster analysis based on conditional covariance. In the analysis, each item is considered a separate cluster, and the conditional-covariance-based proximity of clusters is used to combine clusters. The clustering continues until the entire test is joined into a single cluster (Stout, 2006). For the present study, this procedure was used with each subtest to generate potential partitions into AT/PT sets, where a potential AT set must contain at least 4 items but not more than half the subtest's items.

DETECT can stand alone as an exploratory procedure to determine the number of dimensions in a test and identify which dimension is measured by each item. Included in that exploratory process is an effect size for multidimensionality, the DETECT index. In

a confirmatory mode, the DETECT index can be calculated for a collection of different groupings of items. For the present study, the DETECT index was calculated for each partition identified in the previous step, and the "best" partition of items for DIMTEST analysis was the partition with the highest DETECT index, since that partition showed the best evidence of representing distinct dimensions.

To avoid capitalizing on chance, one-third of the student responses were used to select the AT set, and the remaining two-thirds were used to calculate the DIMTEST statistic (Froelich & Habing, 2008). Each examinee was assigned a random number between 0 and 1. Participants were used to select the AT set if their assigned numbers were less than the 33rd percentile of all the assigned random numbers. This method of group assignment resulted in a random sample of one-third of the respondents being assigned to the AT selection group.

Factor Structure

The factor structure of the nine subtests was examined with confirmatory factor analysis (CFA), which allowed for analysis of the structure of the scores as well as validation of proposed structures. To allow for model cross-validation, the sample was randomly split into two subgroups. Each participant was assigned a random number between 0 and 1. Participants were assigned to subgroup 1 if their assigned random numbers were less than the median of all assigned numbers. This method of assignment resulted in random assignment to subgroups with 706 and 705 subjects. Separating the standardization sample into two random subgroups allowed for the option to cross-validate the best-fitting model.

The standardization sample appears to be a multilevel structure, that is, students are nested within classrooms. With multilevel data, observations are not completely independent, which violates the assumptions of many statistical models, including factor analysis. Nonindependence in such models can bias the results: model parameters tend to be overestimated, and their standard errors are underestimated, thus increasing Type I error rates (Bliese & Hanges, 2004; Julian, 2001). However, multilevel analysis with severely unbalanced groups can result in models that fail to converge or in an inability to estimate some parameters (Muthén, 1989; Raudenbusch & Bryk, 2002; Singer & Willet, 2003).

In the standardization sample, at least 193 different teachers submitted data (57 observations do not include teacher name). Teachers submitted numbers of test booklets ranging from 1 to 42. Sixty teachers submitted only 1 test booklet each, the mean number submitted was 7, and the median number submitted was 3.5. Examination of test booklets and parent surveys indicate that some participating teachers used “teacher name” to indicate the child’s classroom teacher, while others interpreted “teacher name” to indicate the name of the examination administrator. Thus, the multilevel structure of the sample is quite ill-defined and unbalanced. Consequently, it was decided that multilevel analysis was inappropriate.

Model specification

The hallmark of CFA is the use of competing models. In the present study, four plausible models (Figures 2–5) were posited prior to analysis. These are presented on the following pages.

Model 1 is a one-factor model to test the hypothesis that the subtests are simply facets of a single “achievement” trait.

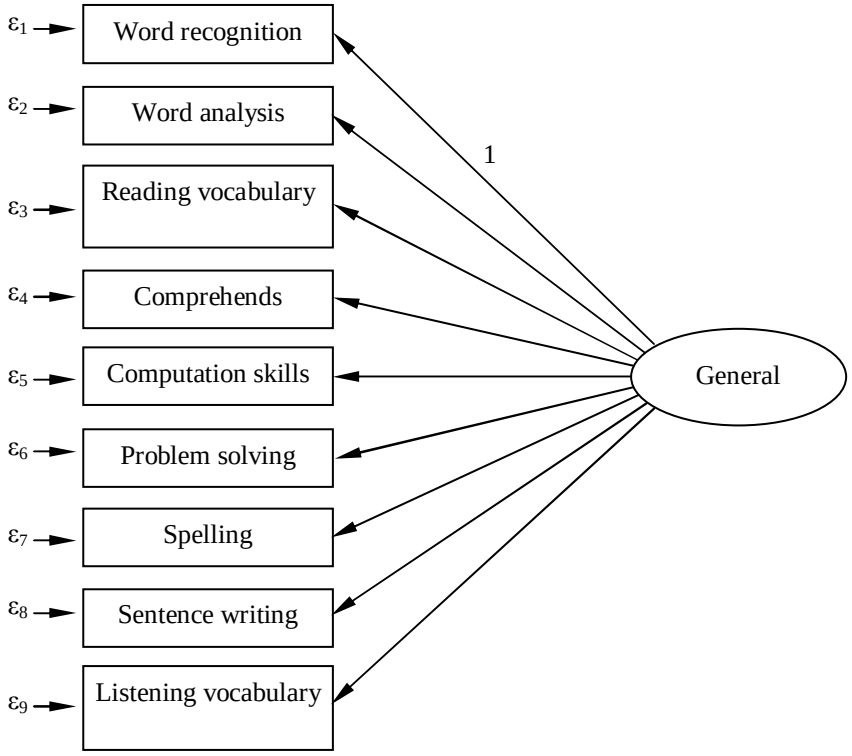

Figure 2. Model 1, a one-factor model.

Model 2 is the test author's model, which was inferred from the composite score structure for the CIBS-II (see Figure 3).

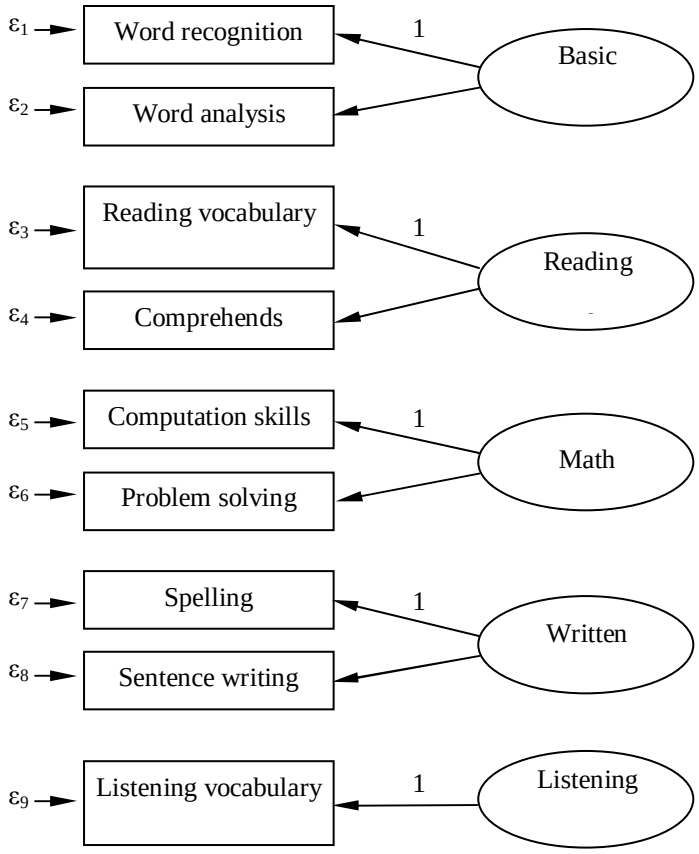

Figure 3. Model 2, a five-factor model. Covariances between the five latent variables will be estimated freely. Paths are not shown in the figure for the sake of clarity.

A potential Model 3 (see Figure 4) would extend Model 2 to test the hypothesis that the constructs influencing subtest scores are all related to a higher-order "general achievement factor." This model should be estimated only if the correlations between factors in Model 2 support the existence of a higher-order factor.

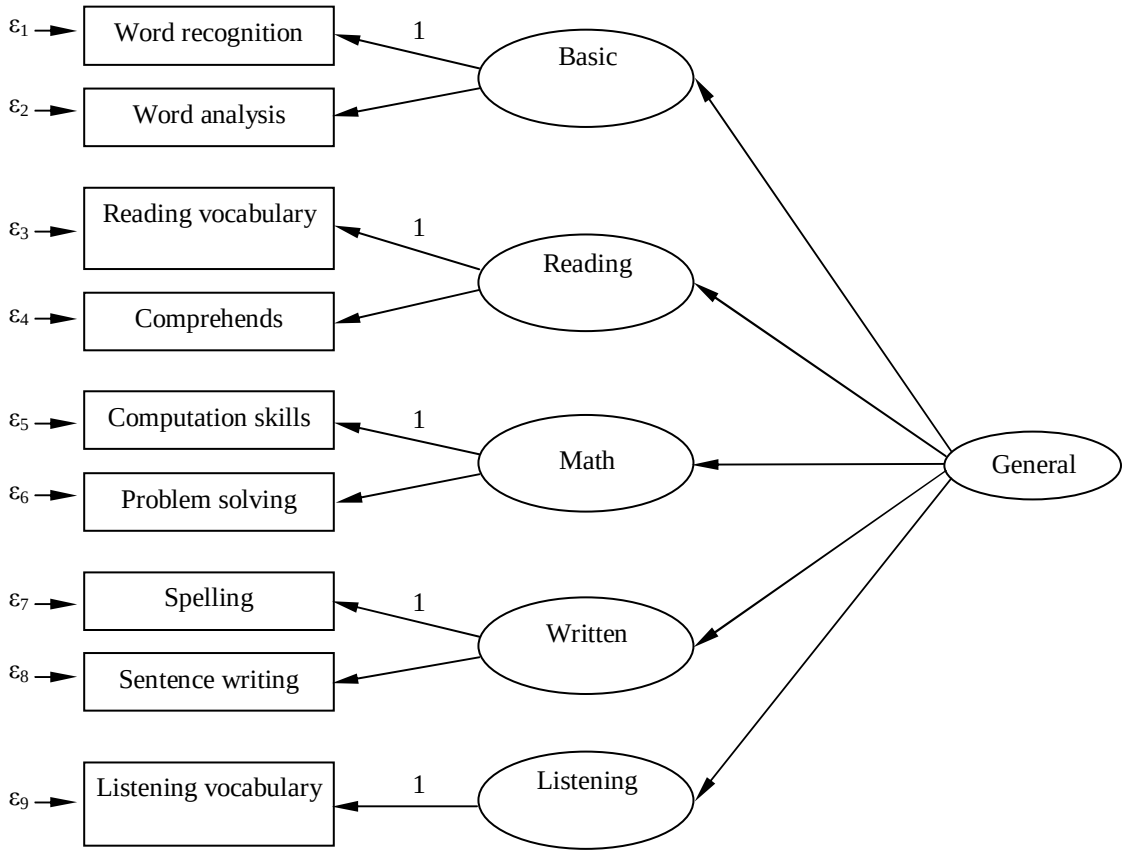

Figure 4. Model 3, a five-factor model with one second-order general factor.

Model 4 (see Figure 5) is a three-factor model based on the supposition that all reading-related subtests should be linked to a general reading factor and that the Listening Comprehension subtest taps understandings of word meanings exclusive of reading, which is more akin to a writing-related skill.

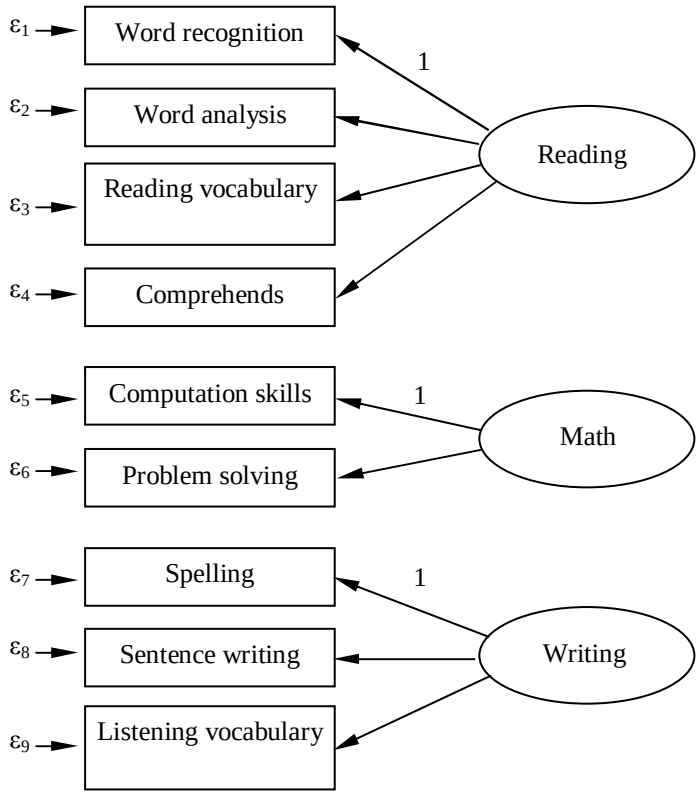

Figure 5. Model 4, a three-factor model. Covariances between the three latent variables will be estimated freely. Paths are not shown in the figure for the sake of clarity.

The listening vocabulary subtest is included in the composite score structure as the Listening Comprehension Indicator. However, this clearly is not a composite score, since no other scores are combined with it. In latent variable modeling, a latent variable is

unobservable. Manifest variables, the observable indicators, are used to infer information about the unobservable latent variables. In using Model 2 to represent the composite score structure, the Listening Comprehension Indicator is included as a latent factor with a single indicator. Single-indicator factors pose technical and substantive problems in latent variable models.

The technical problem arises because a latent factor with a single indicator creates an indeterminate, or under-identified, model. A solution cannot be estimated. This problem is resolved by setting the pattern coefficient to a specific value, in this case 1, and setting the error variance for the indicator to a specific value, usually 0 or some other estimate of error. In this case, an estimate based on the reliability and the variance of the subtest is used.

Substantively, there is disagreement over whether a latent factor with a single indicator should be considered a factor at all, because a factor is meant to account for shared variance among a set of observed indicators (Brown, 2006). Indicators have error, which create inaccuracy in measuring a latent variable. Using multiple indicators helps reduce the effect of the error because the error is part of the model. In addition, it is possible that not all of an indicator's variance reflects the latent construct so the score on the single indicator does not perfectly assess the construct (Kline, 2005). Again, multiple indicators resolve this problem, since their common variance is used to make inferences about the latent variable. Controversy aside, the composite score structure for the CIBS-II was developed by the test's author, and the model based on that structure was tested.

CFA modeling can be used in an exploratory manner by using modification indices to guide model respecification. A modification index can suggest adding a

parameter to the model; for example, the index might indicate that allowing an indicator to load on two factors would improve the fit of the model. Following such a suggestion without substantive reason is likely to lead to capitalization on chance. Modification indices should only prompt respecification if a theoretically defensible argument in support of the change can be made (Brown, 2006; Jöreskog & Sörbom, 1996).

Respecifying models on the basis of results from a particular sample can lead to results that are peculiar to that sample. A model is examined, adjustments are made to the model, and then the modified model is refit using the same data. Cross-validating the respecified model with data from an independent sample provides evidence that the new model is not merely capitalizing on chance features of the original sample (MacCallum, 1995).

Estimation

Maximum likelihood (ML) estimation is the most common method of fitting data to CFA models (Jöreskog & Sörbom, 1996; Kline, 2005). This method is appropriate only for continuous multivariate normal data, although it is robust to minor departures from normality. Use of ML when data show pronounced departure from normality is known to produce unreliable results, including inaccurate fit statistics and under-estimation of standard errors (Brown, 2006).

Means, standard deviations, skew, and kurtosis for the nine subtests' scale scores are reported in Table 6. The individual subtest scores do not show dramatic departures from normal distributions; however, univariate normality of variables is a necessary but not sufficient condition for MVN. PRELIS 2.8, a data preprocessing program for LISREL 8.80 (Jöreskog & Sörbom, 2006), was used assess the multivariate normality

(MVN) of the subtest scores. The relative multivariate kurtosis of the scores is 1.064. Bentler (1998) recommends that MVN can be assumed when this index is below 3.0. Based on this evidence of MVN, LISREL 8.80 with ML estimation was used to estimate the models in this analysis.

Evaluation of model fit

Model fit was evaluated using a combination of fit indices, following the recommendations of Hu and Bentler (1999) and Brown (2006). Brown categorizes fit indices as measures of absolute fit, measures adjusted for parsimony, and measures of comparative or incremental fit. The χ^2 statistic is a measure of absolute model fit. The sensitivity of χ^2 to sample size is widely reported. For moderate to large sample sizes, the χ^2 statistic is sensitive to even small differences between the input covariance matrix and the model implied covariance matrix, causing the model to be spuriously rejected. However, χ^2 statistics are useful in comparing nested models, so they are routinely reported.

The standardized root mean square residual index (SRMR) is another measure of absolute fit. Its value is not dependent on sample size, making it a better index of absolute fit. Brown describes the SRMR as "the average discrepancy between the correlations observed in the input matrix and the correlations predicted by the model" (2006, p. 82). SRMR varies between 0 and 1, with small values reflecting better fit. Hu and Bentler (1999) suggest that SRMR < .08 implies reasonable fit, although this index should not be used in isolation.

The root mean square error of approximation (RMSEA) index assesses absolute fit, but it also accounts for parsimony in that it can be interpreted as average discrepancy

between the input matrix and model-implied matrix *per each degree of freedom.* All else held equal, a complex model with many degrees of freedom (i.e., fewer freely estimated parameters) will have a lower RMSEA than a model with few degrees of freedom. RMSEA is positive number, and values near zero imply good fit. Following the recommendation of Hu and Bentler (1999) RMSEA < .06 was used as an indication of reasonable fit.

The comparative fit index (CFI; Bentler, 1990) compares the fit of the tested model against the fit of an independence model implying no relationship among the variables. The Tucker-Lewis index (TLI; Tucker & Lewis, 1973), also known as the non-normed fit index (NNFI), also compares the tested model to the independence model, but it penalizes models with an excessive number of freely estimated parameters. CFI and TLI both approach 1 for well-fitted models. Hu and Bentler (1999) suggest that values of CFI and TLI above about .95 imply reasonable fit.

Models were judged as having acceptable fit only if *all* the selected fit statistics fell within acceptable range. In particular, RMSEA was given careful consideration since it gives preference to more parsimonious models (i.e., more degrees of freedom). In sum, then, the collection of fit indices used to judge model fit in this study were: SRMR < .08, RMSEA < .06, CFI > .95, and TLI > .95. In addition to reviewing fit indices, parameter estimates were also used in judging models. Parameter admissibility, significance, and interpretability helped guide model choice. Models were also evaluated by examining modification indices and residuals. Residuals for good fitting models should be approximately normally distributed, and standardized residuals with magnitude greater than 2 may indicate localized poor fit.

Conclusion

The instrument under examination in this study is a collection of nine subtests from the CIBS-II, which in its full form contains more than 150 subtests and checklists. The nine subtests under study were chosen to be standardized with a large representative norming sample. Validation of an instrument can be viewed as building an argument in support of interpreting test scores in a particular way for a particular purpose (Kane, 2006). The CIBS-II standardized scoring sheet produces four composite scores (basic reading, reading comprehension, math, and written expression) and a listening comprehension indicator score. A strong argument in support of reporting the nine CIBS-II subtest scores according to these five scores would be the existence of a five-factor model corresponding to the scoring structure (i.e., Model 2).

This study investigated the proposed internal structure of the test scores by examining the dimensionality of the subtests using the DIMTEST procedure (Stout, 1987; Froelich & Stout, 2003; Froelich & Habing, 2008). Results of the dimensionality study may also help interpret CFA models. The study also investigated the fit of CIBS-II subtest scores from the standardization sample to the model implied by the score structure as well as to other theoretically plausible models. The sample was randomly split in half to allow one sample to be used to test and respecify models while holding the second sample in reserve for cross-validation of the selected model.

CHAPTER FOUR

RESULTS

In this chapter, results are reported for the DIMTEST analysis of dimensionality, and then results are reported for the confirmatory factor analysis (CFA) study of the factor structure of the CIBS-II subtest scores.

Dimensionality/DIMTEST

The DIMTEST procedure produces a statistical test of the null hypothesis that a test is essentially unidimensional. The procedure uses a subset of the items called the assessment test (AT), which is suspected of being dimensionally distinct from the remaining items in the test (the partitioning test, or PT). DIMTEST estimates the conditional covariances of all pairs of items in AT, conditional on PT scores. The average of these conditional covariances over all possible PT scores is used to produce the DIMTEST statistic. After a bootstrap bias correction is applied, the final statistic has an asymptotically normal distribution.

The success of DIMTEST in detecting a lack of essential unidimensionality depends on the choice of AT. If multidimensionality is suspected based on the content of some items, DIMTEST can be applied in a confirmatory manner by choosing AT items based on their content. For this study, DIMTEST was used in an exploratory mode. Following a method described by Froelich and Habing (2008), conditional covariance-based cluster analysis (CCPROX/HAC, Stout, 2006; Roussos, Stout, & Marden, 1998) was used to identify potential AT sets. The DETECT index (Kim, 1994; Zhang & Stout,

1999; Stout, 2006) was calculated for each of the potential AT sets. The set with the highest DETECT index was selected as the target AT set.

For the present study, responses from one-third of the examinees (487) were randomly selected to use in selecting the AT sets for each subtest (Froelich & Habing, 2008). The remaining two-thirds were used in the DIMTEST analysis.

Using the 24-item listening vocabulary subtest as an example, of the 24 hierarchical clusters generated by CCPROX/HAC, ten met the criteria for potential AT sets: the set must contain at least 4 items but not more than half the items. The DETECT procedure was applied to each potential AT set, and the index was recorded. The potential AT sets for the listening vocabulary subtest and their DETECT index values are shown in Table 7. AT set number 7 was tentatively selected since its DETECT value was greatest.

The DIMTEST program (Stout, 2006) includes a routine called ATFIND, which applies a different combination of CCPROX/HAC and DETECT analyses to identify a candidate AT set. Before selecting the final AT set, ATFIND was used, and its results were compared to the results from the procedure described above. For example, for the listening vocabulary subtest, ATFIND identified potential AT set 9 from Table 7. Since the DETECT value for set 9 is lower than that for set 7, the final AT set chosen for DIMTEST analysis of the listening vocabulary subtest was set number 7.

For the DIMTEST analyses of the nine CIBS-II subtests in this study, five used the AT set identified by ATFIND, and four used the AT set selected from the two-step procedure described by Froelich and Habing (2008). A summary of the number of items

in the final AT set for each subtest is presented in Table 8. A full listing of items selected for each subtest is presented in the Appendix.

Table 7

Potential AT Sets for the Listening Vocabulary Subtest

AT Set Number	1	2	3	4	5	6	7	8	9	10
Items	4	4	4	4	4	10	6	6	8	8
	10	10	10	10	10	14	7	8	9	9
	13	13	13	14	14	18	8	9	11	11
	14	14	14	18	18	21	9	11	12	12
	17	17	17	21	21		11	12	15	15
	18	18	18	22			12	15	16	
	19	19	21				15	16		
	20	21	22				16			
	21	22								
	22									
DETECT Index	.386	.357	.308	.245	.182	.143	.409	.392	.343	.274

Table 8

Number of Items in Each Final AT Set

wordrec	wordanly	readvoc	compass	compute	probsol	spell	sentwrit	listnvoc
35*	17*	8*	18*	12*	8	13	4	8

Note. For subtests marked with an asterisk, ATFIND identified the AT set. For the others, the two step procedure described identified the AT set. Subtest names are abbreviated in all tables as follows: wordrec = word recognition; wordanly = word analysis; readvoc = reading vocabulary; compass = comprehends passages; compute = computation; probsol = problem solving; spell = spelling; sentwrit = sentence writing; listnvoc = listening vocabulary.

Results of the DIMTEST analyses are presented in Table 9. For each subtest, the DIMTEST statistic *T*, the *p*-value for the statistic under the null hypothesis of essential unidimensionality, and the conclusion drawn about the subtest's dimensional structure are given. For the comprehends passages subtest, computation subtest, problem solving subtest, and spelling subtest, the null hypothesis of essential unidimensionality could not be rejected, so it was concluded that these subtests are essentially unidimensional. The null hypothesis was rejected for the five other subtests, so it was concluded that the assumption of essential unidimensionality of these subtests is untenable. The implications of these results are discussed in Chapter Five.

Table 9

DIMTEST Results for Each Subtest

	wordrec	wordanly	readvoc	compass	compute	probsol	spell	sentwrit	listnvoc
T	4.771	6.696	5.505	-0.715	0.414	-0.520	-0.789	6.868	5.696
p	<.00001	<.00001	<.00001	0.763	0.340	0.699	0.785	<.00001	<.00001
Dim	Multi	Multi	Multi	Uni	Uni	Uni	Uni	Multi	Multi

Note. T is the DIMTEST statistic. Dim = dimensionality; Multi = multidimensional; Uni = Unidimensional.

Internal Structure/CFA

To assess the extent to which the standardization data conform to the factor structure suggested by the test's author, the data were fit to a series of factor models using confirmatory factor analysis (CFA). As described previously, CFA can provide evidence in support of a particular model in two ways: (1) by providing fit indices that describe the extent to which the data fit the hypothesized model, and (2) by fitting the

data to competing models to show whether the hypothesized model fits better than other plausible models (Thompson, 2004).

LISREL 8.80 (Jöreskog & Sörbom, 2006) was used to fit the data to the various models. The estimation and iteration procedure is based on calculations with the covariance matrix. As described earlier, the data were randomly split into two groups. Subgroup 1 was used for model fitting and respecification (a quasi-exploratory process) followed by cross-validation of the favored model with subgroup 2. The covariance matrices for each subgroup are presented in Tables 10 and 11.

Four models were posited prior to analysis. Path diagrams for the four models are shown on pages 45–48. Model 1 was a single factor model. This model is the simplest (i.e., most parsimonious) and provides a check of the possibility that the nine subtests are all indicators of a general achievement factor. Model 2 is a five-factor model that corresponds to the composite score structure presented by the test's author. Model 3 is an extension of Model 2 wherein the common variance among the five factors from Model 2 is modeled as being influenced by a second order factor. Model 4 was specified based on subjective content analysis of the subtests. This model posits that the nine subtests are indicators of three factors called Reading, Math, and Writing. In Model 4 the Listening Vocabulary Subtest is assumed to assess understandings of word meanings exclusive of reading, which is more akin to a writing-related skill, so it is assumed to be an indicator for the Writing factor.

Table 10

Covariance Matrix of Subgroup 1 Standardization Data

	wordrec	wordanly	readvoc	compass	compute	probsolv	spell	sentwrit	listnvoc
wordrec	9.700								
wordanly	6.050	8.036							
readvoc	5.070	3.875	5.646						
compass	5.673	4.333	4.103	7.311					
compute	3.260	2.643	2.432	2.870	4.647				
probsolv	3.542	2.948	2.813	3.241	2.690	4.515			
spell	6.133	4.434	3.633	4.071	2.823	2.837	6.323		
sentwrit	3.780	2.983	2.578	2.984	2.199	2.185	3.230	4.418	
listnvoc	4.017	3.143	3.331	3.603	1.955	2.401	2.706	2.056	5.298

Table 11

Covariance Matrix of Subgroup2 Standardization Data

	wordrec	wordanly	readvoc	compass	compute	probsolv	spell	sentwrit	listnvoc
wordrec	9.489								
wordanly	6.535	8.820							
readvoc	5.152	4.318	5.8349						
compass	5.728	4.861	4.238	6.926					
compute	3.676	3.123	2.744	3.278	5.006				
probsolv	3.748	3.507	2.786	3.364	3.105	4.608			
spell	6.176	5.144	3.787	4.335	2.986	2.945	6.245		
sentwrit	3.908	3.224	2.618	3.066	2.301	2.029	3.163	4.580	
listnvoc	3.933	2.913	3.176	3.438	2.227	2.320	3.018	2.067	4.992

In the following sections, fit indices of these models are presented and compared. The fit indices for all four models are summarized in Table 12. Interpretation and subsequent model specification is described following the initial results for Models 1–4.

Table 12

Model Fit Indices for Models 1–4

Model	χ^2	*df*	*p*	RMSEA	SRMR	CFI	TLI
1	271.904	27	< .0001	.113	.041	.972	.972
2	47.470	18	.0002	.048	.021	.996	.996
3	159.833	23	< .0001	.096	.032	.983	.983
4	188.171	24	< .0001	.099	.033	.981	.981

Model 1

Model 1, the single-factor model, represents an extreme of parsimony and may be considered less plausible than a multi-factor model. Comparing a target model to an implausible model is not considered good practice; results on the basis of such comparison lack strength, since the competing model is a "straw man" (Brown, 2006; Kline, 2005). However, analysis that rules out this simple model provides support for the existence of a multifactor model (Thompson, 2004).

Model fit indices in Table 12 indicate that Model 1 has poor fit, with a very low *p*-value and RMSEA well above the cutoff. Given the complex theoretical structure suggested by the author, Model 1 was not examined further.

Model 2

Model 2 represents the five-factor structure implied by the composite score structure of the CIBS-II subtests. The listening vocabulary subtest is the only indicator on the Listening Comprehension factor. In factor analysis, a "factor" accounts for common variance among a set of indicators. Therefore, the Listening Comprehension factor is treated as a "pseudofactor" (Brown, 2006, p. 141). In practice, this distinction has little impact on estimation of the model, with the exception of consideration of measurement error.

In CFA, the measurement error (i.e., ε_9 in Figure 3) is conceived of as the amount of variance in the indicator that is not accounted for by the factor. With a single indicator linked to a factor, attempting to estimate the measurement error can cause serious problems with overall model estimation (Kline, 2005). Fixing measurement error of the indicator to a set value resolves this problem. A reasonable estimate for error variance comes from using a reliability index and the variance of the indicator (Kline, 2005):

$$\varepsilon_9 = s_9^2(1 - r_9).$$

The reliability of the listening vocabulary subtest as measured by Cronbach's alpha is .894, which suggests that $1 - .894 = .106$ represents the proportion of variance in listening vocabulary scores due to error. The variance for this subtest is 5.145, so $(.106)(5.145) = .545$ was used as the estimate of error variance for this subtest.

Model 2 showed good fit (see Table 12). As an extra check on RMSEA, the 90% confidence interval (0.0181, 0.0763) was examined. The upper bound of this interval is higher than the acceptable cutoff for RMSEA, which suggests possible room for

improvement to the model. However, RMSEA values between .05 and .08 have been suggested as indicating reasonable fit (Browne & Cudeck, 1993).

The largest modification index was 22.985 for the error covariance between the spelling subtest and the word recognition subtest. The next highest was 17.314 for the error covariance between the sentence writing subtest and the word recognition subtest. This pattern of common variance between a Basic Reading indicator and the Written Expression indicators could also be seen in the factor correlation matrix. The correlation between the Basic Reading factor and the Written Expression factor was quite large (.959). In addition, the correlation between Basic Reading and Reading Comprehension was also large (.936). Parameter estimates and ad hoc respecification of Model 2 are reported later in this chapter.

Model 3

Model 3 adds a second-order factor to Model 2 to account for the interfactor correlations. Since this model uses five estimated parameters (the pattern coefficients of the first-order factors on the second-order factor) to account for the ten estimated interfactor correlations in Model 2, Model 3 is more parsimonious. In reducing the number of freely estimated parameters in a model, model fit is necessarily degraded. If, however, the more parsimonious solution still has acceptable fit, it is preferred over the more complex model.

Fit for Model 3 (see Table 12) was substantially degraded as compared to Model 2, with a much lower *p*-value and a marked increase in RMSEA. Even the lower bound of the 90% confidence interval for RMSEA (0.0826, 0.1093) was above the specified cutoff for adequate fit. Model 3 is nested under Model 2, so the χ^2-difference test can be used to

judge whether Model 2 provides significantly better fit. The χ^2-difference between Model 3 and Model 2 was 112.363 with 5 degrees of freedom, which indicates that Model 2 yields significantly better fit ($p < .0001$). It was decided that exploring improvements to Model 2 fit was a higher priority than making respecifications to Model 3.

Model 4

Model 4 is a three-factor model based on the assumption that the four reading-related subtests measure one Reading factor and that the listening vocabulary subtest could be linked to the Written Expression factor since the subtest dealt with word meanings without accounting explicitly for reading (i.e., the test was administered verbally). As was the case for Model 3, this model has fewer estimated parameters than Model 2, so model fit is necessarily degraded. Nonetheless, the model warrants consideration because a more parsimonious solution is preferred if it has acceptable fit.

Model 4 produced an inadmissible solution: the completely standardized correlation between the Reading and Writing factors was estimated to be 1.023. Inadmissible solutions generally result from misspecified models (Brown, 2006; Jöreskog & Sörbom, 1996). This model also produced many large residuals (differences between the input covariance matrix and the model-implied covariance matrix) and had fit indices that were substantially worse than those for Model 2.

Model 2 Parameters

After Model 2 was identified as the best fitting model of the initial models, its parameters were investigated. LISREL calculates a standard error estimate for all estimated parameters, which allows the statistical significance of the parameters to be evaluated. In Model 2, all parameters were found to be statistically significant (i.e., *p*-

values < 0.01). The completely standardized solution standardizes all parameters and allows them to be compared directly; therefore, the completely standardized solution is presented here. Tables 13, 14, and 15 present the model's pattern coefficients, structure coefficients, and factor correlations, respectively.

The structure coefficients are interpreted as the correlation between indicators and factors. Although the theoretical basis of Model 2 assumes, for example, that the Basic Reading factor directly influences only the word analysis subtest scores and word recognition subtest scores, the structure coefficients show that for subgroup 1 of the standardization sample, the Basic Reading factor and the spelling subtest had a correlation coefficient of .843. The Basic Reading, Reading Comprehension, and Written Expression factors each had at least one subtest with a structure coefficient higher than one of its estimated pattern coefficients.

Table 13

Pattern Coefficients for Model 2

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	0.912	- -	- -	- -	- -
wordanly	0.752	- -	- -	- -	- -
readvoc	- -	0.806	- -	- -	- -
compass	- -	0.792	- -	- -	- -
compute	- -	- -	0.728	- -	- -
probsolv	- -	- -	0.806	- -	- -
spell	- -	- -	- -	0.879	- -
sentwrit	- -	- -	- -	0.695	- -
listnvoc	- -	- -	- -	- -	0.947

Note. Parameters fixed at zero are represented with dashes.

Table 14

Structure Coefficients for Model 2

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	**0.912**	0.853	0.679	0.875	0.597
wordanly	**0.751**	0.703	0.559	0.721	0.492
readvoc	0.754	**0.806**	0.689	0.706	0.633
compass	0.742	**0.792**	0.677	0.694	0.622
compute	0.542	0.622	**0.728**	0.584	0.448
probsolv	0.600	0.689	**0.806**	0.647	0.496
spell	0.843	0.770	0.705	**0.879**	0.510
sentwrit	0.667	0.609	0.558	**0.695**	0.403
listnvoc	0.620	0.744	0.583	0.549	**0.947**

Note. Coefficients in bold print are equal to the estimated pattern coefficients.

Table 15

Factor Correlations for Model 2

	Basic_Rd	Rd_Comp	Math	Write	Listen
Basic_Rd	1.000				
Rd_Comp	0.936	1.000			
Math	0.744	0.855	1.000		
Write	0.959	0.876	0.802	1.000	
Listen	0.655	0.785	0.616	0.580	1.000

The factor correlation matrix in Table 15 shows that the Basic Reading Subtest was very highly correlated with both the Reading Comprehension and the Written Comprehension factors. Factor correlations approaching unity are often an indication that a model has too many factors (Brown, 2006). High correlation between factors weakens the argument that the factors are measuring distinct constructs.

The pattern of relationships revealed in the structure coefficients and factor correlations from Model 2 suggested two additional models. Model 5 is a variation on Model 3 in which the second order factor is a "verbal" construct that influences only the Basic Reading, Reading Comprehension, and Written Expression factors. The two remaining factors are allowed to correlate. Model 6 is a very simple two-factor (Verbal and Math) model, which comes from collapsing the three highly correlated factors into one factor.

Neither of the two models approached acceptable fit (see Table 16). Model 5 results included very high modification indices suggesting that the Math and Listening Comprehension factors should load on the second-order factor; in other words, the modification indices suggested Model 3. Model 6 results included several large modification indices suggesting correlated error variance of indicators. Such a pattern may be indicative of missing factors in the model (Brown, 2006).

Table 16

Model Fit Indices for Models 5 & 6

Model	χ^2	*df*	*p*	RMSEA	SRMR	CFI	TLI
5	494.799	24	< .0001	.167	.324	.910	.866
6	201.793	26	< .0001	.098	.034	.980	.973

Though Model 2 had good fit, the model showed areas of local strain, with four particularly large residuals that were well outside the desired normal distribution and several large modification indices. In particular, the largest modification index in the results from Model 2 was 22.985 for the covariance of errors for the spelling and word recognition subtests. The content of the two subtests are clearly linked: spelling asks

children to recall or "sound out" how to spell words relatively quickly, and word recognition asks children to quickly sight-read lists of words with no context. Both subtests require familiarity and fluency with phonics as well as knowledge of many of the peculiar spelling and pronunciation rules of the English language. Thus, it was decided to specify Model 2A as being identical to Model 2 with the addition of estimating the covariance of errors between the two subtests.

Model 2A had very good fit: $\chi^2(17) = 23.227$ ($p = .113$); RMSEA = .024 with a 90% confidence interval of (0.0, 0.043); SRMR = .014; CFI = .999; TLI = .998. Model 2 is nested under Model 2A, and the χ^2-difference test indicated that Model 2A provides significantly better fit than Model 2 ($\chi^2(1) = 24.243$, $p < .0001$). Tables 17, 18, and 19 present the pattern coefficients, structure coefficients, and factor correlations for Model 2A. The solution's residuals were within reasonable limits and approximately normally distributed. No large modification indices were included in the solution. Some interfactor correlations, however, were still large (see Table 19), especially among reading and writing related factors. Likewise, structure coefficients (see Table 18) showed high correlations between some indicators and factors to which they are not presumed to be linked. For example, the correlation between the word recognition subtest and the Reading Comprehension factor is greater than the pattern coefficients between the factor and its indicators.

Table 17

Pattern coefficients for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	0.888	- -	- -	- -	- -
wordanly	0.766	- -	- -	- -	- -
readvoc	- -	0.805	- -	- -	- -
compass	- -	0.793	- -	- -	- -
compute	- -	- -	0.729	- -	- -
probsolv	- -	- -	0.805	- -	- -
spell	- -	- -	- -	0.848	- -
sentwrit	- -	- -	- -	0.725	- -
listnvoc	- -	- -	- -	- -	0.947

Note. Parameters fixed at zero are represented with dashes.

Table 18

Structure Coefficients for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	**0.887**	0.844	0.678	0.817	0.593
wordanly	**0.766**	0.729	0.585	0.705	0.512
readvoc	0.767	**0.805**	0.689	0.717	0.633
compass	0.754	**0.793**	0.678	0.705	0.623
compute	0.557	0.623	**0.729**	0.601	0.449
probsolv	0.615	0.688	**0.805**	0.664	0.496
spell	0.781	0.755	0.699	**0.849**	0.503
sentwrit	0.668	0.645	0.598	**0.725**	0.430
listnvoc	0.633	0.744	0.583	0.561	**0.947**

Note. Coefficients in bold print are equal to the estimated pattern coefficients.

Table 19

Completely standardized factor correlations for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
Basic_Rd	1.000				
Rd_Comp	0.952	1.000			
Math	0.764	0.855	1.000		
Write	0.921	0.890	0.824	1.000	
Listen	0.669	0.785	0.616	0.592	1.000

In an attempt to achieve slightly more parsimonious model, and to help explain the high interfactor correlations, Model 3A was specified as Model 2A with a second-order factor rather than correlated factors. Since Model 3A is more parsimonious, slightly degraded model fit is to be expected, but since RMSEA rewards parsimony, acceptable fit could still be achieved.

The resulting model had marginal fit: $\chi^2(22) = 86.817$; $p < .0001$, RMSEA = .064; SRMR = .026; CFI = .992; TLI = .987. Although the fit statistics for this model appeared to be acceptable, the model showed a great deal of localized strain (Brown, 2006), with many large residuals and large modification indices that lacked substantive interpretations. In addition, the χ^2-difference test indicated Model 2A fits significantly better than Model 3A ($\chi^2(5) = 63.590$, $p < .0001$).

Model 4 was similar to Model 2 in that it resulted in a modification index suggesting that the error of the spelling subtest and word recognition subtest should be allowed to covary. Model 4A was specified in a manner analogous to Model 2A. Fit for Model 4A was significantly better than for Model 4, although the fit was still rather

marginal: $\chi^2(23) = 117.053$; $p < .0001$; RMSEA = .076; SRMR = .028; CFI = .989; TLI = .983. The solution had many large residuals and included large modification indices that could not be supported by substantive reasoning. Model 4A also resulted in very high correlation between factors, with correlation of .988 between Reading and Writing and correlation of .891 between Writing and Math. Model 4A is nested under Model 2A, and the χ^2-difference test again showed that Model 2A has significantly better fit ($\chi^2(6) = 93.826$, $p < .0001$).

Cross-Validation

Model 2A was selected as the final model that best fit the covariance matrix for subgroup 1. A cross-validation with subgroup 2 was performed to ensure that Model 2A was not merely replicating sample specific variation and to estimate final model parameters. The cross-validation study essentially is an investigation into whether the factor structure of the CIBS-II is invariant across the two independent subgroups by means of a multisample CFA. (For a detailed review of measurement invariance from the CFA perspective, see Vandenberg & Lance, 2000.) To fully cross-validate Model 2A, the model is simultaneously, but independently, fit to the covariance matrices of both subgroups. If good fit is achieved, then the estimation is repeated with a series of increasingly restrictive constraints. The process is illustrated here with Model 2A.

As described above, the model was fit to both subgroups' covariance matrices simultaneously. In this situation, fitting the data to the same model means that the pattern of factor loadings was the same for both samples as was the specification of correlated error variance between the word recognition and spelling subtests; however, the estimated parameters could differ between the subgroups' solutions. The solution was a

good fit to the data (see Table 20), so invariance testing continued. Next the process was repeated subject to the constraint that the pattern coefficients were invariant (i.e., the Λ matrices were identical). The fit for this model is shown in "step 2" in Table 20. Next the process was repeated subject to the constraint that the pattern coefficients and error variances were invariant across groups (i.e., the Λ and Θ matrices were identical). Finally, the process was repeated subject to the constraint that the pattern coefficients, error variances, and factor covariances were invariant across groups (i.e., the Λ, Θ, and Φ matrices were identical). Fit indices at each step are listed in Table 20.

Table 20

Global Model Fit Indices for Cross-Validation

Step	Invariance Constraint	χ^2	*df*	*p*	RMSEA	SRMR	CFI	TLI
1	None	67.532	34	.0005	.037	.017	.998	.996
2	Λ	71.689	38	.0008	.035	.019	.998	.996
3	Λ & Θ	85.920	47	.0005	.034	.021	.998	.996
4	Λ, Θ, & Φ	102.868	62	.0009	.031	.028	.997	.997

Evaluating the extent to which these results show invariance involves the χ^2-difference test, which has been supported as an accurate method to evaluate invariance (French & Finch, 2006). This test checks for statistically significant decline in fit as the models are constrained by checking for a significant χ^2-difference. Table 21 summarizes the results of this test across the invariance testing of Model 2A. Comparing the first, unconstrained model to the final, fully constrained model the χ^2-difference was 35.336 with 28 degrees of freedom, with a *p*-value of .1603. Thus the conclusion that Model 2A fits the entire CIBS-II standardization sample is supported.

Table 21

χ^2-Difference Tests for Cross-Validation

Step	χ^2	*df*	χ^2-difference	*df*- difference	*p*
1	67.532	34			
2	71.689	38	4.157	4	0.385
3	85.920	47	14.231	9	0.114
4	102.868	62	16.948	15	0.322

Final parameters for the model are taken from the simultaneous fit of both subgroups to the model in the final step of cross-validation (see Tables 22–24). The final model parameters suffer the same interpretive problems as do the initial parameters for Model 2A (Tables 17–19). In particular, correlations are large among the reading and writing factors (.952) and moderately large among the math, reading, and writing factors (.770, .805, and .846). In addition, the structure coefficients (see Table 23) show high correlation of the word recognition subtest with factors to which it is presumably not linked. The word recognition subtest is modeled as linked only to the Basic Reading factor, but its correlation (i.e., structure coefficient) with the Reading Comprehension factor is higher than the two indicators which are modeled as linked to that factor. The Written Expression factor also has multiple indicators with high structure coefficients, including the word recognition subtest. However, in specifying a CFA model in which the word recognition subtest was linked to more than one factor (i.e., it was cross-loaded), the estimated factor coefficients were not significantly different from zero.

Table 22

Pattern coefficients for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	0.893	- -	- -	- -	- -
wordanly	0.780	- -	- -	- -	- -
readvoc	- -	0.804	- -	- -	- -
compass	- -	0.812	- -	- -	- -
compute	- -	- -	0.757	- -	- -
probsolv	- -	- -	0.816	- -	- -
spell	- -	- -	- -	0.859	- -
sentwrit	- -	- -	- -	0.703	- -
listnvoc	- -	- -	- -	- -	0.946

Note. Parameters fixed at zero are represented with dashes.

Table 23

Structure Coefficients for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
wordrec	**0.891**	*0.848*	0.686	*0.848*	0.587
wordanly	**0.780**	*0.742*	0.600	*0.742*	0.513
readvoc	0.765	**0.804**	0.680	0.725	0.621
compass	0.772	**0.811**	0.686	0.732	0.627
compute	0.583	0.640	**0.756**	0.609	0.464
probsolv	0.628	0.690	**0.816**	0.657	0.500
spell	*0.817*	*0.775*	0.691	**0.858**	0.534
sentwrit	0.669	0.634	0.566	**0.703**	0.437
listnvoc	0.623	*0.730*	0.580	0.589	**0.946**

Note. Coefficients in bold print are equal (within rounding error) to the estimated pattern coefficients. Coefficients in italic print are discussed in Chapter 5.

Table 24

Completely standardized factor correlations for Model 2A

	Basic_Rd	Rd_Comp	Math	Write	Listen
Basic_Rd	1.000				
Rd_Comp	0.952	1.000			
Math	0.770	0.846	1.000		
Write	0.952	0.902	0.805	1.000	
Listen	0.659	0.772	0.613	0.623	1.000

Summary

The dimensionality study used an exploratory approach to assess the essential unidimensionality of the CIBS-II subtests. The DIMTEST results led to the rejection of the null hypothesis of essential unidimensionality for five of the nine subtests. The null hypothesis could not be rejected for the other four subtests.

The CFA study showed support for the composite score structure proposed by the test's author. By allowing the error covariance between the word recognition and spelling subtests to be estimated, very good model fit was established. This good fit was cross-validated across an independent random subgroup of the data. At the same time, structure coefficients and interfactor correlations hinder the interpretation of the model and may indicate the presence of complex interactions among the subtests and latent constructs.

Although the dimensionality analysis and the confirmatory factor analysis of these subtest scores both involved evidence regarding the internal structure of the CIBS-II scores, no formal link between the two forms of analyses was made. The relationship between these two analyses and implications for interpretation of CIBS-II scores is explored in the following chapter.

CHAPTER FIVE

DISCUSSION

In this chapter, results from the dimensionality study and the confirmatory factor analysis (CFA) study are discussed separately, and then the results are woven together to draw overall conclusions from the study.

Dimensionality/DIMTEST

DIMTEST uses conditional covariances to test the null hypothesis of essential unidimensionality of a set of test scores. The concept of essential unidimensionality was developed to acknowledge the fact that pure unidimensionality is an extremely strong assumption. An essentially unidimensional test can be considered as measuring one major dimension even though other unimportant and uninterpretable minor dimensions may be present (Nandakumar, 1991).

In expressing the composite score structure of the CIBS-II, the test's author implies that the subtests have a particular dimensional structure. That is, each subtest contributes to exactly one composite score, with the implication that each subtest measures one dimension. The CFA study of the composite score structure can provide evidence to support or refute this interpretation of the subtest scores, but such a study fails to investigate the underlying dimensionality of the subtests.

The present study used DIMTEST to assess the dimensionality of the CIBS-II subtests. A randomly selected subgroup of one-third of the respondents was used to select the assessment test (AT) for each subtest. The AT sets were selected via a combination of conditional covariance-based cluster analysis and the DETECT index. This procedure is

designed to select a set of items that are most likely to display dimensional distinctiveness. The other two-thirds of the respondents were used in the actual DIMTEST analysis of the test scores, in which the covariances of pairs of AT items, conditioned on examinee ability as measured via the items not included in AT, are used to calculate the DIMTEST statistic.

The null hypothesis of essential unidimensionality was rejected for five subtests: word recognition, word analysis, sentence writing, reading vocabulary comprehension, and listening vocabulary comprehension. The null hypothesis was not rejected for the four remaining subtests: comprehends passages, computation, problem solving, and spelling. The standardization sample yields evidence that these four subtests are essentially unidimensional. For the other five subtests, the DIMTEST analysis does not yield specific information about their dimensional structure. DIMTEST simply provides a statistical test to indicate that these subtests should not be assumed to be essentially unidimensional. These results indicate that each of these subtests measures at least two dominant dimensions; the exact number of dimensions cannot be determined from this information.

It is worth noting that two subtests with content commonly used to exemplify multidimensionality were among the four subtests that appear to be essentially unidimensional. Tests of mathematical problem solving are often used as examples of multidimensional tests (e.g., Zhang & Stout, 1999) because such tests measure can measure multiple dimensions, for example: examinees' ability to read and comprehend a problem, reason mathematically about that problem, and successfully perform computations necessary to correctly solve the problem. Likewise, tests based on

comprehension of short text passages (e.g., the comprehends passages subtest; Zhang & Stout, 1999) are common among examples of multidimensional tests because the content of a text passage can affect how an examinee interacts with the passage. Thus, a passage based on historical events may measure reading comprehension and understanding of (or interest in) history, while a passage based on nature may measure reading comprehension and understanding of (or interest in) environmental science.

Among the subtests for which the null hypothesis of unidimensionality was rejected are some rather complex subtests. For example, the word analysis subtest includes items that require the examinee to indicate whether two words read by the test administrator sound exactly the same (e.g., "Listen carefully to these words: *boy-toy.* Are they the same?"); identify sounds heard in words read aloud by the test administrator (e.g., "I want you to listen carefully and then tell me the first letter you hear in the word."); read aloud words and nonsense words to sample phonemic awareness (e.g., the examinee is asked to read aloud such lists of words as "bush, push, fush"); and divide written words into syllables. The reading and listening vocabulary comprehension subtests involve reading or listening to sets of words and identifying the word that does not belong in each list (e.g., "Tell me the word that does not belong: circulate, orbit, rotate, *recover.*"). The sentence writing subtest requires examinees to understand word meanings and to apply them within the context of English grammar and sentence structure when, for example, they are asked to use the words "captain, complain, terrible, and dangerous" in one complete sentence. Although scoring of the sentences was intentionally generous to examinees (e.g., correct spelling was not required, and rules of

grammar and punctuation were not stringently enforced), constructing sentences is a complex activity requiring lexical and syntactical knowledge of the English language.

It is much more difficult, however, to substantively interpret the results for the word recognition subtest, which was identified as multidimensional. In this subtest children are asked to sight-read lists of words. The lists are devoid of any context and were arranged by increasing difficulty (i.e., grade level). Each word was considered an item and was scored correct if the child correctly pronounced the word and incorrect if the child mispronounced, misread, or was excessively slow in "sounding out" the word. One possible source of multidimensionality could arise from words of the same formal difficulty level having a different level of familiarity for children. For example, "play," "me," and "small" were at the same grade level yet easier to sound out than "two" and "what." At a higher grade level, "attitude" and "diminish" are likely more familiar to most sixth graders, and easier to sound out, compared to "plateau."

The lack of essential unidimensionality in five of the nine subtests is a threat to the interpretability of the composite score structure of the CIBS-II. The five subtests in question appear to measure more than one important dimension, but in the score structure, they are each interpreted as contributing to exactly one composite score. Since these subtests appear to be multidimensional, the meanings of the subtest scores are not clear. Is an individual's high score on a given subtest a result of the individual's ability with respect to the targeted dimension or another unknown dimension? What can be inferred about the ability of two individuals with the same scores on a subtest that is not unidimensional? These questions cannot be answered without much deeper analysis of

the dimensional structure of the subtest in question, but they could emphasize the difficulties associated with interpreting scores from multidimensional subtests.

Internal Structure/CFA

This study sought evidence in support of the composite score interpretation of the nine CIBS-II subtests. Using the CFA framework, the CIBS-II subtest scores from the standardization sample were fit to the model implied by the composite score structure as well as to other theoretically plausible rival models. The sample was randomly split in half to allow one subgroup to be used to test and respecify models while holding the second subgroup in reserve for cross-validation of the best-fitting model.

Model 1 was a one-factor model. Support for this model would imply that all CIBS-II subtest scores contribute to a single "general achievement" factor. Such a model would serve as support for the use of a single composite score from the nine subtests. Model 1, however, did not fit the data, which suggests that the nine subtests measure more than one general construct.

Model 2 was a reflection of the composite score structure advocated by the test's author, and it fit the data well. By modifying the model to estimate the covariance of errors in the word recognition and spelling subtests (which created Model 2A), the model fit was improved to a point that further modifications could no longer be reasonably proposed.

Statistical and substantive justification can be made for allowing estimation of error covariance. The error covariance adds to the model common variance between two indicators that could otherwise only be accounted for in the covariance of their respective factors. In Model 2, the structure coefficient between the spelling subtest and the Basic

Reading factor was .843, which indicates a relationship between the spelling subtest and the Basic Reading factor. In addition, the highest modification index from Model 2 suggested estimating the error covariance rather than constraining it to zero.

In addition to these statistical indications that the error covariance should be estimated, an argument can be made on substantive grounds that these two subtests share variance that should be considered in the model. Although being able to spell words may be influenced primarily by a Written Expression factor and reading lists of words out of context may be influenced primarily by a Basic Reading factor, the two skills are both related to familiarity with common words and both require fluency with phonics concepts.

Other models considered as rivals to Model 2 did not result in acceptable fit. Of most interest was modeling the interfactor correlations in Models 2 and 2A as a second order factor (Models 3 and 3A, respectively). However, both of these models degraded fit significantly leaving Model 2A as the favored model.

In Models 2 and 2A, the listening vocabulary subtest is the only indicator linked to the Listening Comprehension factor. Single-indicator factors create technical and interpretive challenges. The technical challenge can be surmounted through fixing parameters rather than estimating them. In Model 2 and 2A, the parameter coefficient was fixed to 1.0 and the error variance was fixed to 0.545.

The interpretive challenge centers on considering an individual variable as an indicator of a latent variable; latent variables are normally assessed conceived of as reflecting the common variance of multiple indicators (Thompson, 2004). The alternatives to interpreting the listening vocabulary subtest as the single indicator for a

latent factor are to (a) add more indicators for the latent factor in question, (b) use the variable as an indicator for one of the other factors, or (c) drop the variable and factor from the model. The first alternative is not an option because the present study can only involve the existing subtests. The second and third alternatives are attractive because they would result in models that involve a more familiar interpretation, but the solution for Model 2A supports the option of rejecting both alternatives.

It is not known how the test's author derived the composite score structure for the subtests, but the structure coefficients provide support for keeping the listening vocabulary subtest separate from other factors. The listening vocabulary subtest had a high structure coefficient on the Reading Comprehension factor, which can be explained by the importance of an examinee's vocabulary to both the Reading Comprehension factor and the listening vocabulary subtest. However, the Listening Comprehension factor had moderate structure coefficients with the other subtests, which suggests that this factor stands apart from the other subtests. The listening vocabulary subtest seems to be assessing something distinct from the other factors, its correlation with Reading Comprehension notwithstanding.

The listening vocabulary subtest was found to be multidimensional in the DIMTEST analysis, yet this subtest does not seem to be as intertwined with other subtests and factors as other subtests identified as multidimensional. If the subtest could be clustered into a small number of unidimensional parcels, these parcels might be useful as multiple indicators of the latent Listening Comprehension factor. However, it is possible that such parcels would measure dimensions too disparate to be common indicators or that unidimensional parcels do not exist.

Linking Dimensionality and Internal Structure

The dimensionality study yields information that can be helpful in understanding the parameters in the model identified as the favored model from the CFA investigation (Model 2A). Five subtests (word recognition, word analysis, sentence writing, reading vocabulary comprehension, and listening vocabulary comprehension) were identified as being not essentially unidimensional. All the factor models specified in this study involved an assumption of unidimensionality: each subtest was specified as being influenced by exactly one factor. Multidimensional measurement is specified in a factor model by estimating pattern coefficients (i.e., λs) from two (or more) factors to one indicator. That is, a single indicator would be linked to (or more properly, influenced by) multiple factors. Specifying multidimensional measurement in this manner is referred to as cross-loading indicators.

Fitting data that is not unidimensional to a model that is based on the assumption of unidimensional indicators should manifest in lack of fit. Modification indices should suggest cross-loading indicators to multiple factors or allowing error covariances to be estimated. Given what is known about the dimensionality of the CIBS-II subtests, the fit of Model 2A is remarkably good. The multidimensional nature of the five subtests did not cause severe enough misfit to cause the proposed composite score structure to be rejected. However, the effect of the multidimensional nature of some subtests is revealed in the interfactor correlations and structure coefficients in Model 2A.

High correlations between factors can suggest overlap in the latent constructs measured in the model. A common rule of thumb states that factor correlation exceeding .85 may indicate the presence of too many factors in the model (Brown, 2006). However,

as the results in Chapter 4 indicate, the five factor model for CIBS-II model appears to be optimal; models with fewer factors had unacceptable fit indices.

The structure coefficients can help explain this apparent paradox. In a model with unidimensional measurement and correlated factors, structure coefficients represent the correlation between an indicator and a factor. The word recognition subtest had very high structure coefficients with the Basic Reading, Reading Comprehension, and Written Expression factors (see the italicized coefficients in Table 23). The same pattern of very high structure coefficients was seen with the word analysis, reading vocabulary, and spelling subtests (see Table 24). The listening vocabulary subtest had a high structure coefficient with the Reading Comprehension factor. Four of the subtests mentioned here are among those identified as being not essentially unidimensional. The multidimensional nature of these subtests is apparently evident in the high correlations with other factors. The nature of these results suggests that the composite score structure put forth by the test's author is a valid interpretation of the scores (i.e., the internal structure supports such an interpretation); however, the details of the model hint that this composite score structure may be a simpler model than what one would find were a full exploratory study mounted, including full consideration of the content of the subtests and a deeper investigation of the dimensionality of the subtests (e.g., what is the extent of the multidimensionality among the five identified subtests?).

A strikingly similar pattern of results was discovered in a CFA validity study of the TerraNova achievement test system (CTB/McGraw Hill, 1997): although model fit was adequate, very high interfactor correlations and high structure coefficients raised questions about the interpretability of the three-factor structure (Stevens & Zvoch, 1997).

A CFA study of the KeyMath Revised Normative Update (Connolly, 1998) also showed reasonable model fit for the three-factor structure advocated by the test's author, but high interfactor correlations led the researchers to perform an exploratory study that gave support for a one-factor model (Williams et al., 2007). Analyses of the dimensionality of the individual subtests were not performed in these other studies, but researchers speculate about the presence of "common, nonachievement features of performance such as decoding or problem solving" (Stevens & Zvock, 1997, p. 987). Such a common construct would be very likely to manifest in each subtest as a secondary dimension. The discovery of a similar pattern of results in the CIBS-II subtests may suggest a need for a broader investigation into the latent structures of achievement test results.

It is not the purpose of the current study to deeply examine the content of the subtests. Likewise, it not the purpose of this study to propose a restructuring of the score structure. This study intends to describe the nature of the scores and investigate whether evidence supports the proposed interpretation (i.e., the nine subtests can be interpreted as measuring five broad areas). The proposed structure, as interpreted in Model 2A fits better than any other plausible model that assumes unidimensional subtests.

Conclusion

The results of this study show support for the composite score structure for CIBS-II subtest scores. Questions remain as to whether a more complex model, taking into account the multidimensional structure of individual subtests, would produce a more meaningful interpretation, but the present study produced evidence that the composite score structure is a good fit to the standardization sample in an absolute sense (i.e., the

model fits) and that the composite score structure is a better fit than other plausible models.

The support for the composite score model was weakened slightly by evidence that several subtests are multidimensional. No attempt was made to identify the dimensional structure of the individual subtests, but the high structure coefficients and high interfactor correlations may indicate that the content of the subtests overlaps to a higher degree than intended or that the subtests measure some unidentified common construct.

In the unified view of validity, validation is an ongoing process. In this model of validity, it is not a test or a test's scores that are validated; it is a proposed interpretation of the scores that is validated (AERA et al., 1999; Kane, 2006). The *Standards for Educational and Psychological Testing* (AERA et al., 1999) list five common sources of evidence: evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Validating one proposed interpretation of the scores might involve gathering multiple sources of evidence. Other interpretations might demand different sources of evidence.

The present study provides evidence based on the internal structure of the test, which supports the composite score structure. However, the evidence from this study could also be used to support a more complex interpretation involving other as yet unidentified constructs. Evidence based on test content and response processes might be used to help build a case for such an interpretation.

Other evidence could be collected to add to the present validation investigation. The construct model of validity relies on the statement of an underlying theory (Messick,

1989). An explication of the theoretical underpinnings of the individual subtests and the manner in which they were combined into composite scores would amplify the evidence provided in the present study. Such an analysis would also provide content-based evidence of the validity of the composite score structure. Similarly, a confirmatory study to more thoroughly investigate the dimensionality of the subtests, including content-based analysis of clusters identified by CCPROX/HAC, would provide additional evidence of the dimensional structure of the subtests individually and of the CIBS-II as a whole.

The present study addresses a small portion of what should be an ongoing process to validate the intended uses of the CIBS-II. The overall goal of validation is to evaluate "the proposed interpretations and uses of measurements" (Kane, 2006, p. 59). Inherent is such an evaluation is a consideration of the consequences of testing and of the proposed interpretations of scores. A proposed interpretation or use of scores has potential consequences. The nature of those potential consequences (e.g., high stakes vs. low stakes) affects the evaluation of whether the scores support such a use or interpretation.

The explicit consideration of the consequences of test use as a part of a validation is generally accepted as part of the unified view of validity (e.g., AERA et al., 1999; Messick, 1989; Kane, 2006), but it is not a universally held position (e.g., Cizek et al., 2008). The argument seems to center on whether test developers should (or indeed whether they *are able to*) anticipate the possible uses of test scores. In addition, consequential validity evidence cannot be gathered until a test is in use, which means no test can be adequately validated before it is published if the explicit consideration of consequential validity is required as a part of validation (Cizek et al., 2008). A middle

ground might be to recognize that consideration of the consequences of proposed interpretations should be involved in the development of any instrument (i.e., it is a responsibility of the test developer), and consequences of novel uses of an instrument should be considered by test users (Nichols & Williams, 2009).

Such arguments will likely occupy validity theorists for many years to come. In the meantime, practitioners study what they can. Consequences of the CIBS-II cannot currently be known since the test is planned for release as this study is being completed. However, some potential uses can be considered, and the evidence from this study can help evaluate the suitability of CIBS-II scores for those uses. For example, composite scores from the CIBS-II appear to be suitable for such low-stakes uses as monitoring student progress, identifying areas of strength and weakness, or setting learning goals. However, questions about the possible existence of more complex interpretations of subtest scores render the composite score structure unsuitable for such high-stakes uses as qualifying students for placement in special education courses or for accountability reporting.

Historically, the CIBS test series has received little attention from researchers investigating the validity of its scores. The research reported here represents a change from that historical pattern. This research, follow-up studies to this work, and the studies reported in the test manual (French & Glascoe, 2009) represent the kind of accumulation of evidence that characterizes modern notions of test validation.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*, 37-53.

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

American Psychological Association. (2001). *Appropriate use of high-stakes testing in our nation's schools.* Retrieved December 6, 2008, from http://www.apa.org/pubinfo/testing.html

Bentler, P. (1998, March 10). Kurtosis, residuals, fit indices. Message posted to SEMNET discussion list. Available from http://bama.ua.edu/cgi-bin/wa?A2=ind9803&L=semnet&T =0&O=D&P=20612

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238-246.

Berry, C. R., & Howell, W. G. (2008). Accountability lost. *Education Next, 8,* 66-72.

Birenbaum, M., & Tatsuoka, K. K. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement, 20,* 17-26.

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7,* 400-417.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Breidenbach, D. H., & French, B. F. (2008, April). *Construct validity of the comprehensive Inventory of Basic Skills-Revised via confirmatory factor analysis.* Paper presented at the National Council on Measurement in Education, New York, NY.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th, ed., pp. 1-16). Westport, CT: Praeger.

Brigance, A. H. (1976). I*nventory of Basic Skills.* Wolburn, MA: Curriculum Associates.

Brigance, A. H. (1977). *Diagnostic inventory of basic skills.* Wolburn, MA: Curriculum Associates.

Brigance, A. H. (1983). *Comprehensive Inventory of Basic Skills.* North Billerica, MA: Curriculum Associates.

Brigance, A. H. (1998). *Comprehensive Inventory of Basic Skills-Revised.* North Billerica, MA: Curriculum Associates.

Brigance, A. H. (2009). *Comprehensive Inventory of Basic Skills-II.* North Billerica, MA: Curriculum Associates.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Buros Institute of Mental Measurements. (n.d.). Complete index. Retrieved December 7, 2008, from http://www.unl.edu/buros/bimm/html/index00.html

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

Cizek, G. J. (2001). Test review of the CIBS-R. From B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* [Electronic version]. Retrieved September 8, 2007, from the Buros Institute's *Test Reviews Online* web site: http://www.unl.edu/buros

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

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

Cohen, M. (2002). Unruly crew. *Education Next, 2.* Retrieved December 6, 2008, from http://media.hoover.org/documents/ednext20023_42.pdf

Connelly, J. B. (1985, April). *The Brigance inventories and the CST process.* Paper presented at the Annual Convention of the Council for Exceptional Children, Anaheim, CA. (ERIC Document Reproduction Service No. ED258404)

Connolly, A. J. (1998). *KeyMath-Revised Normative Update: A diagnostic inventory of essential mathematics.* Circle Pines, MN: American Guidance Service.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Philadelphia: Harcourt.

Croom, L. (1997). Mathematics for all students: Access, excellence, and equity. In J. Trentacosta, & M. J. Kenney (Eds.) *Multicultural and gender equity in the mathematics classroom: The gift of diversity* (pp. 1-9). Reston, VA: National Council of Teachers of Mathematics.

CTB/McGraw-Hill. (1997). *TerraNova CTBS Multiple Assessments*. Monterey, CA: Author.

Daub, D., & Colarusso, R. P. (1996). The validity of the WJ-R, PIAT-R, and DAB-2 reading subtests with students with learning disabilities. *Learning Disabilities Research & Practice, 11,* 90-95.

Dwyer, C. A. (Ed.). (2005). *Measurement and research in the accountability era.* Mahwah, NJ: Lawrence Erlbaum Associates.

Erford, B., & Dutton, J. L. (2005). Technical analysis of the Slosson Phonics and Structural Analysis Test. *Educational and Psychological Measurement, 65,* 1011-1025.

Erford, B., & Klein, L. (2007). Technical analysis of the Slosson–Diagnostic Math Screener (S-DMS). *Educational and Psychological Measurement, 67,* 132-153.

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and U.S. Department of Education reviews under the No Child Left Behind Act of 2001. Washington, DC: Council of Chief State School Officers.

Retrieved on December 5, 2008, from http://www.ccsso.org/content/pdfs/StatewideEducationalAccountabilityUnderNCLB.pdf

Ferguson, J., & Kersting, F. (1988). Comparison of diagnostic inventories used in special education with state-approved essential skills tests (KEST). *Journal of Human Behavior & Learning, 5,* 39-42.

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

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement, 31,* 292-307.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13,* 378-402.

French, B. F., & Glascoe, F. P. (2009). *Comprehensive Inventory of Basic Skills-II standardization and validation manual.* North Billerica, MA: Curriculum Associates.

Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement, 32,* 138-155.

Froelich, A. G., & Stout, W. F. (2003). A new bias correction method for the DIMTEST procedure. Manuscript submitted for publication. Retrieved September 9, 2008, from http://www.stat.iastate.edu/preprint//articles/2001-16.pdf

Geisinger, K. F., Spies, R. A., Carlson, J. F., & Plake, B. S. (Eds.). (2007). *The seventeenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality DIF analysis paradigm. *Journal of Educational Measurement, 40,* 281-306.

Glascoe, F. P. (1999a). The Brigance Comprehensive Inventory of Basic Skills-Revised. *Diagnostique, 24,* 41-51.

Glascoe, F. P. (1999b). *Comprehensive Inventory of Basic Skills-Revised standardization and validation manual.* North Billerica, MA: Curriculum Associates.

Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling, 10,* 142-153.

Green, J. P., Jr. (2007). Determining the reliability and validity of service quality scores in a public library context: A confirmatory approach. Ph.D. dissertation, Capella University, United States -- Minnesota. Retrieved February 5, 2008, from ProQuest Digital Dissertations database. (Publication No. AAT 3241793)

Hart, P. D., & Teeter, R. M. (2002). *A national priority: Americans speak on teacher quality.* Princeton, NJ: Educational Testing Service. Retrieved December 6, 2008, from ftp://ftp.ets.org/pub/corp/survey2002.pdf

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., Murray, H., Olivier, A., & Human, P. (1997). Equity and accessibility. In *Making sense: Teaching and*

learning mathematics with understanding (pp. 65-74). Portsmouth, NH: Heinemann.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64,* 802-812.

Howell, W. G., West, M. R., & Peterson, P. E. (2007). What Americans think about their schools. *Education Next, 7,* 12-26.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Hussar, W. J., & Bailey, T. M. (2006). *Projections of education statistics to 2015* (NCES 2006-084). U. S. Department of Education, National Center for Education Statistics. Washington, DC: U. S. Government Printing Office. Retrieved on November 7, 2006 from http://nces.ed.gov/pubs2006/2006084.pdf

Improving America's Schools Act (1994). Retrieved December 13, 2008, from http://www.ed.gov/legislation/ESEA/index.html

Individuals with Disabilities Education Act. (2004). Retrieved December 7, 2008, from http://idea.ed.gov/download/statute.html

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8. Chicago, IL: Scientific Software International, Inc.

Joshua, M. T., Joshua, A. M., & Kritsonas, W. A. (2006). Use of student achievement scores as basis for assessing teachers' instructional effectiveness: Issues and

research results. *National Forum of Teacher Education Journal, 17.* Retrieved December 6, 2008, from http://www.nationalforum.com/Electronic%20Journal%20Volumes/Joshua,%20Monday%20Use%20of%20Student%20Achievement.pdf

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8,* 325-352.

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

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 581-614). New York: Guilford Press.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the *Wechsler Intelligence Scale for Children—Fourth Edition*: What does it measure? *School Psychology Review, 35,* 108-127.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. Ph.D. dissertation, University of Illinois at Urbana-Champaign. Retrieved December 6, 2008, from ProQuest Digital Dissertations database. (Publication No. AAT 9512427)

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Kohn, A. (2000). Sell schools, not test scores. *Realtor: The Business Tool for Real Estate Professionals.* Retrieved December 6, 2008, from http://www.realtor.org/archives/sellschoolalfarchive2000jan

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

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37,* 752-777.

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

Krawiec, R. M., & Spadafore, G. J. (1983). Comparing the Brigance Diagnostic Inventory of Basic Skills and the Wide Range Achievement Test. *Reading Improvement, 20,* 230-232.

Lewis, S. (2005). Issues related to disaggregating data in a new accountability era. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 31-39). Mahwah, NJ: Lawrence Erlbaum Associates.

Linkoas, L. W., Enright, B. E., Messer, P., & Thomas, P. J. (1986, March). A reliability study of the Brigance CIBS, Comprehensive Inventory of Basic Skills. Paper presented at the Annual Convention of the Council for Exceptional Children, New Orleans, LA. (ERIC Document Reproduction Service No. ED271502)

Linn, R. L. (2005). Scientific evidence and inference in educational policy and practice: Implications for evaluating adequate yearly progress. In C. A. Dwyer (Ed.),

Measurement and research in the accountability era (pp. 21-30). Mahwah, NJ: Lawrence Erlbaum Associates.

Lissetz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Education Researcher, 36,* 437-448.

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16-36). Thousand Oaks, CA: Sage.

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

Metcalf, L. A. (2002). Curriculum-sensitive assessment: A psychometric study of tracking as a distributor of opportunity to learn high school mathematics. Ph.D. dissertation, University of Illinois at Urbana-Champaign. Retrieved December 6, 2008, from ProQuest Digital Dissertations database. (Publication No. AAT 3070030)

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

National Center for Education Statistics. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance* (NCES 2000-469). Retrieved August 13, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2000469

National Center for Education Statistics. (2008). *The condition of education 2008* (NCES 2008-031). Retrieved December 6, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008031

National Commission of Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28,* 3-9.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

Peterson, P. E. (2007). A lens that distorts: NCLB's faulty way of measuring school quality. *Education Next, 7,* 46-51.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68,* 679-634.

Porter, A. C. (2005). Prospects for school reform and closing the achievement gap. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 59-98). Mahwah, NJ: Lawrence Erlbaum Associates.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Thousand Oaks, CA: Sage.

Resnick, D. P. (1982). History of educational testing. In A. L. Wigdor & W. R. Garner (Eds)., *Ability testing: Uses, consequences, and controversies, Part II* (pp. 173-194). Washington, DC: National Academy Press.

Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1-30.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Stevens, J. J., & Zvoch, K. (2007). Confirmatory factor analysis of the TerraNova Comprehensive Tests of Basic Skills/5. *Educational and Psychological Measurement, 67,* 976-989.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Stout, W. (2006). DIMPACK 1.0. Chicago, IL: Assessment Systems Corporation.

Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement, 57,* 5-19.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Thompson, B. & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56,* 197-208.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70.

Walberg, H. (2003). Accountability unplugged. *Education Next, 3.* Retrieved December 2, 2008, from http://media.hoover.org/documents/ednext20032_76.pdf

Williams, T. O., Jr., Fall, A. M., Eaves, R. C., Darch, C., & Woods-Groves, S. (2007). Factor analysis of the KeyMath-Revised Normative Update Form A. *Assessment for Effective Intervention, 32,* 113-120.

Yang, T. (2006). Measurement of Korean EFL college students' foreign language classroom speaking anxiety: Evidence of psychometric properties and accuracy of a computerized adaptive test (CAT) with dichotomously scored items using a CAT simulation. The University of Texas at Austin. (Publication No. AAT 3204228)

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

APPENDIX

Table of Items Chosen for Final AT Sets Used in DIMTEST Analysis

Subtest	Items in Final AT Set
Word Recognition	1 9 10 16 17 22 23 27 32 33 44 47 48 52 54 56 57 58 59 61 65 67 68 70 72 75 77 78 79 84 85 90 93 94 96
Word Analysis	8 32 33 35 36 37 38 39 40 41 42 43 44 45 46 47 48
Reading Vocabulary	7 8 10 12 14 18 19 24
Comprehends Passages	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Computation	5 6 7 8 9 10 11 12 13 14 16 18
Problem Solving	3 4 5 6 7 8 9 10
Spelling	16 17 18 21 24 25 26 27 29 30 32 34 38
Sentence Writing	1 2 3 4
Listening Vocabulary	6 7 8 9 11 12 15 16