THE ROLE OF EXTENDED TIME ON THE SAT® REASONING TEST FOR STUDENTS WITH DISABILITES

by

Jennifer Hartwig Lindstrom

Under the Direction of Noël Gregg

ABSTRACT

A great deal of controversy surrounds the question of whether valid inferences can be made from scores obtained from accommodated test administrations for students with disabilities. This study was designed to examine the latent structure of the newly revised Scholastic Aptitude Reasoning Test (SAT®, 2005) across groups of examinees without disabilities tested under standard time conditions and examinees with disabilities tested with extended time to determine whether the test measures the same construct for both groups. The impact of the recent changes in item type, test length, and response format on test scores of students with disabilities is not clear. An assessment of measurement invariance was conducted to determine the extent to which test scores across the two groups of examinees are comparable.

Data from the initial administration of the new SAT Reasoning Test (administered March 17, 2005) was used for the analyses in a sample of 4,952 examinees. First, confirmatory factor analysis was used to assess the fit of a single-factor structure model for the Critical Reading, Math, and Writing sections to each of the two groups. Next, a study of factorial invariance examined whether a common factor model for the Critical Reading, Math, and Writing sections holds across the two groups at increasingly restrictive levels of constraint. Invariance across the

two groups was supported for factor loadings, thresholds, and factor variances. Thus, there was no real evidence to suggest that the scores on the Critical Reading, Math, and Writing sections of the SAT Reasoning Test have different interpretations when examinees have an extended time administration as opposed to the standard time administration.

INDEX WORDS: Learning disabilities, Accommodations, High-stakes tests, Postsecondary students, Scholastic Aptitude Test (SAT®), Measurement invariance

THE ROLE OF EXTENDED TIME ON THE SAT® REASONING TEST FOR STUDENTS WITH DISABILITES

by

JENNIFER HARTWIG LINDSTROM

B.S.Ed. The University of Missouri, 1996

M.Ed. Central Missouri State University, 1999

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GA

2006

THE ROLE OF EXTENDED TIME ON THE SAT® REASONING TEST FOR STUDENTS WITH DISABILITES

by

JENNIFER HARTWIG LINDSTROM

Major Professor: Noël Gregg

Committee: Allan Cohen
Deborah Bandalos
John Langone
J. Mark Davis

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2006

DEDICATION

For my father, James C. Hartwig, in loving memory.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Noël Gregg, my major professor, for her ongoing encouragement and inspiration. Her support and investment in me as a professional and as a person has greatly contributed to who I am today and to who I will continue to become. Dr. Gregg's guidance and collaboration on various projects have been at worst challenging, and at best, truly enlightening. What I have gained from her mentorship, professionalism, and expertise is immeasurable - I can only hope to provide others the same.

I would also like to thank the members of my dissertation committee, Allan Cohen, Deborah Bandalos, John Langone, and Mark Davis for their time and commitment to my success. I greatly appreciate all of their support and guidance. I am particularly grateful to Drs. Cohen and Bandalos for their patience and statistical expertise; their council has truly made this challenge an incredible learning experience. I am also thankful to the individuals at the College Board who are responsible for making my dissertation become a reality. Without their assistance, this project could not have happened.

Also, I would like to thank my fellow colleagues and mentors at the Regents' Center for Learning Disorders at the University of Georgia – Mark, Chris, Scott, Christopher, Lisa, Melissa, and Vicki. Their guidance, support, thoughtfulness, and sense of humor provided a truly enriching environment in which to learn and work. Probably the greatest gift I learned as a graduate assistant at the RCLD is the importance of teaching by example. Never before have I worked alongside a group of individuals whom I equally admire, respect, and whose company I enjoy. I look forward to many more years of collaboration with all of them.

Lastly, I would like to thank my family, especially my mother, for their support, guidance, and strength bestowed upon me during my doctoral career, the completion of this dissertation, and my professional and personal growth. And most importantly, to my boyfriend/fiancé/husband (need I say more?), Will, who has always believed in me and, at times, trusted me more than I trusted myself. I will be forever grateful for his eternal love, kindness, and patience. I look forward to beginning our journey together, and know that wherever it may lead, our love will endure.

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION

Purpose of the Study

The purpose of this study was to examine whether the Scholastic Aptitude Reasoning Test (SAT®, 2005) measures the same construct across two groups of examinees. Specifically, the extent to which test scores of examinees without disabilities tested under standard (e.g., timed) conditions of the SAT are comparable to the scores of students with disabilities tested with extended time was examined. The topic of interest concerns measurement invariance, which is whether a set of indicators assesses the same constructs in different groups (Kline, 2005). In other words, does the provision of extended time change the construct of the test for students with disabilities?

The SAT Reasoning Test is one of the most widely used college admissions tests in the United States (Kobrin & Schmidt, 2005). It is one of several assessment programs (including the PSAT/NMSQT®, and the Advanced Placement Program® [AP®]) developed by *The College Board*, a not-for-profit membership association assisting students in the transition to higher education (College Board, 2005f). The SAT is a three-hour-and-45-minute test that measures critical reading, mathematical reasoning, and writing skills that students have developed over time and need in order to be successful in college (College Board, 2005b).

In 2002, the College Board announced that a new SAT would be introduced in March 2005. The most notable changes to the test (formerly known as the SAT I: Reasoning Test)

included the following: a separate writing section (including multiple-choice questions and a student-written essay) was added, verbal analogies were eliminated from the critical reading section, short reading passages were added to existing long reading passages, math content was expanded to include topics from third-year college preparatory math, and quantitative comparisons were eliminated (College Board, 2005d). In addition, the time limits for the entire test were extended by 45 minutes. See Appendix A for a more detailed description of the changes made to the SAT.

The current study is particularly important in light of the significant changes that have been made to the test. The impact of these changes in item type, test length, and response format on test scores of students with disabilities is not clear. In addition, no evidence has yet been gathered regarding score comparability across regular and extended time administrations for the new SAT Reasoning Test.

Nature of the Problem

Along with the increased focus on high-stakes testing throughout public education is the changing complexion of students enrolling in postsecondary institutions (Scott, McGuire, & Shaw, 2003). Growing numbers of students with disabilities are enrolling in some type of postsecondary institution (American Council on Education, 2000). According to a survey conducted by the U.S. Department of Education's Office of Special Education and Rehabilitative Services (OSERS), almost half (46%) of the students with disabilities enrolled in 2-year and 4-year postsecondary institutions have specific learning disabilities (Ward & Berry, 2005). Similarly, students with learning disabilities currently account for about half of all students with disabilities between the ages of 3 and 21 years (Cahalan, Mandanich, & Camara, 2002).

The increasing number of students with learning disabilities in K-12 and postsecondary educational settings has led to a rise in the number of requests for accommodations, particularly requests for extended time. Recent reviews of the K-12 literature on test accommodations for students with disabilities identified extended time as one of the most frequently used and investigated accommodations (Chiu & Pearson, 1999; Sireci, Li, & Scarpati, 2003; Thompson, Blount, & Thurlow, 2002; Tindal & Fuchs, 2000). Studies have also revealed that the most common accommodation requested by students with learning disabilities on college admissions tests is extended time (Cahalan et al., 2002). In fact, the number of SAT examinees requesting extra time grew by about 26 percent between 1998 and 2003 (Bridgeman, Trapani, & Curley, 2003). Between 1990 and 1995, the percentage of students with learning disabilities who took an accommodated SAT Reasoning Test increased by an average of 14 percent per year, and has since stabilized to approximately two percent of all SAT test-takers (Cahalan et al., 2002; D. Lazarus, personal communication, December 9, 2005). This rise in the use of accommodations, particularly extended time, on standardized admissions tests has led to a greater interest in the comparability of test scores from accommodated administrations.

The reason extended time is a frequently provided accommodation is likely a result of both its theoretical and applied appropriateness. Learning disabilities stem from neurological differences in brain structure, and can dramatically impact the manner and duration in which persons with learning disabilities read, write, learn and take tests (Disability Rights Advocates, 2001). The vast majority of students with learning disabilities are those with reading disabilities or dyslexia. There is strong evidence that individuals do not outgrow a reading disability; it is a persistent and chronic problem (Shaywitz, 2003). Accumulating neurobiologic evidence demonstrates a functional disruption in children, adolescents, and adults with reading disabilities

in those specific neural systems responsible for fast, automatic reading (Gregg, Mather, Shaywitz, & Sireci, 2002). Thus, the need for extended time is supported by evidence of the persistence of reading disabilities and the lack of fluency for individuals with reading disabilities.

Testing Accommodations

The purpose of using a test accommodation is to adjust conditions with the goal of equalizing the opportunity to demonstrate knowledge (Gregg, Morgan, Hartwig, & Coleman, in press). In other words, test accommodations are designed to promote fairness in testing and lead to more accurate interpretations of examinees' tests scores (Sireci et al., 2003). Although accommodations (e.g., extended time) are intended to provide equal access by removing unnecessary challenges (e.g., construct irrelevant variance), some types of accommodations may change the test's construct, thus altering the comparability of scores derived from the accommodated test. Preservation of construct validity allows for score comparability across individuals with and without accommodations.

Research examining the role of accommodations (e.g., extended time) as a potential threat to construct validity is critical. If an accommodation is shown to change the construct of the test for specific groups of examinees, scores from accommodated tests may not be considered comparable. Thus, their test their test scores can no longer be used to determine qualifications for admission, employment, certification, or licensure (Cahalan et al., 2002). Phillips argues that any changes to testing conditions should be avoided if the change (a) alters the skill being measured, (b) precludes the comparison of scores between examinees who received the extended time and those who did not, or (c) allows examinees without disabilities to benefit if they were granted the same accommodation. This last criterion is contentious and recently several researchers have argued that accommodations should only be provided if they offer a "differential" boost to

students with disabilities (see Sireci, 2005; Elliot, McKevitt & Kettler, 2002; Fuchs & Fuchs, 1999; Pitoniak & Royer, 2001 for a discussion of this research).

For the purpose of this study, Phillips' first and second criteria (altering the construct of the test; score comparison) as it pertains to extended time was examined. This study isolated a single, albeit varied, disability (learning) and measure (SAT Reasoning Test) in order to examine whether extended time as an accommodation changes the construct of the test or the comparability of scores obtained from extended time administrations for students with disabilities. In the context of this study, score comparability ensures that the meaning and interpretation of the test score is the same for all groups of students (Pomplun & Omar, 2001).

Measurement Issues in Test Accommodations

Construct Irrelevant Variance

To better understand the issues surrounding the use of accommodations on standardized tests for students with disabilities, a discussion of the measurement issues (e.g., psychometric properties of accommodated tests) associated with test accommodations is necessary. The primary purpose of using test accommodations is to remove construct irrelevant barriers to evaluate performance while maintaining the integrity of the construct being measured by the test (Sireci et al., 2003). For example, a student who learned Spanish as her first language may do worse on a math test administered in English. In this case, English proficiency may be considered extraneous to the math construct targeted by the test, but it would certainly affect her test performance on the English language version of the test. Removing these barriers, which is analogous to accommodating the administration, is thus seen as removing construct irrelevant variance and strengthening the validity of test scores.

Paradoxically, accommodations may also introduce construct irrelevant variance if the accommodation itself changes the construct being measured. If the construct intended to be measured by a test changes, and the new characteristics measured represent a different and unintended construct, then construct irrelevant variance is also present (Sireci, 2005). Furthermore, if the accommodation removes or replaces portions of the test content, construct underrepresentation may result. Therefore, although accommodations are designed to promote fairness in testing, the degree to which the accommodation(s) strengthens validity is directly related to the degree to which the accommodation alters the construct measured (Sireci).

Score Comparability

Despite its common use in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999, Standard 10.11) and elsewhere, the term *score comparability* is not defined anywhere in the standards. The lack of a clear definition for score comparability has led many researchers to define it in the limited framework of differences in mean scores across groups. In fact, for many years, researchers assumed that to appropriately measure these differences one must simply administer a measure across different testing situations and/or different groups and compute the difference between the two (or more) observed scores (Cronbach & Furby, 1970). However, a number of potentially problematic issues in the use of difference scores have been identified (see Cronbach, 1992, or Edwards, 1994). To get an adequate assessment of these differences when comparing mean scores across groups, it is essential that the measure is perceived to be used in the same way by individuals. In other words, it is necessary to show that the two measurements are psychometrically equivalent to make valid comparisons across groups of respondents (Horn & McArdle, 1992).

A test fulfills *measurement equivalence/invariance* when it is shown to measure the same attribute under different conditions (Meade & Lautenschlager, 2004). These different conditions may include the stability of measurement across different populations (e.g. individuals with and without disabilities) and/or different methods of test administration (e.g., extended time administration vs. standard time administration). Under such conditions, tests of equivalence/invariance are typically conducted via confirmatory factor analysis (CFA) methods (Meade & Lautenschlager), thus allowing researchers to determine the extent to which test scores across groups and/or conditions are comparable.

Speededness

Another important measurement issue that warrants further investigation is the concept of *speededness*. The appropriateness of time limits is a critical validity issue: the degree to which educational tests are speeded has a direct bearing on the issue of score comparability because the accommodation of extended time changes the construct measured on a speeded test, but not on a test that is not speeded (e.g., a power test). According to the *Standards*, speededness is "a test characteristic, dictated by the test's time limits, that results in a test taker's score being dependent on the rate at which work is performed as well as the correctness of the responses…Speededness is often an undesirable characteristic" (p. 182). In general, on a pure *speed* test, individual differences depend entirely upon the speed of performance, and the items are relatively easy; on *power* tests, the differences are not contingent on speed and the items increase in difficulty (Ofiesh, Mather, & Russell, 2005). If speed of responding plays a significant role in determining scores on power tests, and speed is not part of the intended construct, then the validity of the assessments is threatened (Bridgeman et al., 2003).

Tests used for college admissions at the undergraduate or graduate level, such as the SAT and the Graduate Record Examination (GRE), are generally designed to minimize the importance of speed. According to the technical handbook for the SAT, the speed with which students can answer the questions should play at most a minor role in determining scores (Donlon, 1984). Similarly, the GRE Technical Manual (Briel, O'Neill, & Scheuneman, 1993) states that the purpose of the GRE General Test is to assess "reasoning skills considered fundamental in graduate study: verbal reasoning, quantitative reasoning, and analytical reasoning" (p. 7), and the "GRE General and Subject Tests are not intended to be speeded" (p. 32). On these tests, most examinees are expected to have enough time to reach all test items. However, when Bridgeman, Curley, and Trapani (2001) examined the extent to which the SAT I: Reasoning Test is speeded, they found that increasing the time limits on the Verbal section of the SAT I: Reasoning Test resulted in a 5-to-10 point standard score increase, and a 20-point standard score increase on the Math section, on average, for examinees without disabilities. These results suggested that there may have been a small degree of speededness on the SAT I Reasoning Test (Bridgeman et al., 2001).

Until recently, test scores obtained from extended time administrations of the test were flagged with an asterisk indicating that the test was taken under nonstandard conditions. However, many testing agencies (e.g., the College Board, Educational Testing Services [ETS], American College Testing [ACT]) no longer flag test scores when an examinee receives extended time due to a disability. Since the removal of the flag, it is essential that testing organizations provide evidence regarding the validity of tests scores taken with accommodations. However, there are a number of challenges associated with conducting this type of research. Such challenges include: (a) multiple types of accommodations that can be provided (e.g., oral

presentation, scribe, extended time, multiple testing sessions, etc.), (b) a high degree of heterogeneity among individuals with disabilities (e.g., variety and severity of disabilities), and (c) controversy regarding how each accommodation changes the construct of the test. These challenges are further compounded by the fact that accommodations are often provided jointly (e.g., extended time and private setting), thus making it extremely difficult to analyze the effects of a single accommodation (Sireci et al., 2003).

Current Study

Examining whether the new SAT Reasoning Test administered with and without the accommodation of extended time measures the same construct across two groups of examinees is particularly important in light of the significant changes that have been made to the test. The changes in item type, test length, and response format on the test scores of students with and without disabilities has not yet been established. In addition, no evidence has been gathered regarding score comparability across regular and modified administrations for the new SAT Reasoning Test.

The purpose of this study was to examine whether the new SAT measures the same construct(s) for examinees with disabilities who received an extended time accommodation and examinees without disabilities tested under standard time conditions. Although previous evidence suggests that the test scores have the same meaning for examinees who took the test with and without accommodations (Rock, Bennett, Kaplan, & Jirele, 1988; Morgan & Huff, 2002), this evidence was based on data obtained from the original SAT and the SAT I: Reasoning Test. Since these studies were conducted, the SAT Reasoning Test scales have been re-centered (in 1990; Dorans, 2002) and additional changes have been made to the question

formats used on the test (in 2005; College Board, 2005a). It is important, therefore, that score comparability be revisited in the context of the newly revised SAT.

Research Questions

Specifically, the following two research questions were addressed in this study:

Research Question 1: Does the SAT Reasoning Test measure the same construct(s) for examinees with disabilities who received an extended time accommodation and examinees without disabilities tested under standard time conditions?

Research Question 2: To what extent are test scores for examinees without disabilities who took a standard (e.g., timed) administration of the SAT comparable to the scores for students with disabilities tested with extended time?

CHAPTER II

REVIEW OF THE LITERATURE

This study examined whether the provision of extended time changes the construct of the SAT Reasoning Test for students with disabilities. In addition, this study investigated the meaning of SAT scores for students with disabilities who were tested with extended time in comparison to students without disabilities tested under standard (e.g., timed) conditions. The topic of interest concerns construct irrelevant variance, which, if present, weakens the validity of interpretations and use of test scores (Haladyna & Downing, 2004).

The aim of this review is to provide both historical and legal perspectives on the provision of extended time in standardized testing for students with disabilities. Following this discussion, the measurement issues thought to affect the use of extended time as well as a review of the research on this topic will be presented. Finally, the purpose and research questions of the current study will be addressed.

Historical Perspective

Understanding the issues surrounding accommodations on high-stakes tests begins with recognition of the consequences for individuals with disabilities who are not provided equal opportunities to demonstrate knowledge. The percentage of students with disabilities going on to postsecondary education and later to professional schools is still less than half that of their peers in the general population (Wagner, Newman, Cameto, Garza, & Levine, 2005). The findings from the National Longitudinal Transition Study-2 (Wagner et al.) indicate that approximately

one out of five "out-of-secondary-school youth with disabilities" (19%) currently attends postsecondary school, a rate that is less than half that of their peers without disabilities (40%; p. 4-8). The rate of enrollment of adolescents with disabilities in 2-year community colleges is not significantly different from that of their peers in the general population (10% vs. 12%). However, similar-age youth without disabilities are more than four and one-half times as likely as youth with disabilities to be currently taking courses at a 4-year college (28% vs. 6%; Wagner et al.). Unfortunately, one contributing factor to these discouraging statistics is lack of access for many students with learning disabilities to appropriate accommodations.

Legislation Governing the Provision of Test Accommodations

The passage of legislation in the 1970s through 1990s focused much needed attention on the need to provide accommodations to individuals whose disabilities interfered with the accurate measurement of their skills and abilities. For instance, the 1990 Americans with Disabilities Act (ADA) mandated the availability of testing accommodations for individuals enrolled in postsecondary settings receiving federal funding (Pitoniak & Royer, 2001). The ADA refers to an accommodation as any variation in the specified assessment environment or process that does not alter in any significant way what the test measures or the comparability of scores (Morgan & Cahalan, 2003). The goal of such accommodations is to provide equal access, or, attempt to "level the playing field" for the test-taker (Tindal & Fuchs, 2000, pg. 9).

Section 504 of the Vocational Rehabilitation Act of 1973 is part of a civil rights law that also contains regulations associated with test accommodations. This law states that "no otherwise qualified handicapped individual in the United States...shall, solely by reason of his handicap, be excluded from the participation in, be denied the benefits of, or be subjected to discrimination under any programs receiving Federal financial assistance" (29 U.S.C. § 794). Section 504

regulations further state that admissions tests must "accurately reflect the test-taker's aptitude or achievement and not the examinee's lack of skill related to the disability except where that skill is the factor the test purports to measure" (Vocational Rehabilitation Act, 1973, Section 84.42[b][3]).

In addition to federal legislation governing the testing of individuals with disabilities, there are also standards that have been developed by the professional communities involved in educational measurement, psychology, and educational research. These standards, while not legally binding, are a widely respected guide to the best practices that have evolved over several decades (Koenig & Bachman, 2004). According to the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), the purpose of an accommodation is "to minimize the impact of the test-taker attributes that are not relevant to the construct that is the primary focus of the assessment" (p.101). These accommodations are provided because the standard procedures in some way interfere with or impede test-takers from performing up to their ability.

Postsecondary Students and Test Accommodations

The composition of the pool of examinees with disabilities has changed over the years, particularly among those who are applying to and enrolling in postsecondary institutions (Scott, McGuire, & Shaw, 2003). Many of the earliest testing accommodations were provided to examinees with physical impairments, hearing impairments, or visual impairments. Over the past decade, however, there has been a significant increase in the number of students with psychological processing disabilities requesting accommodations. In fact, students with learning disabilities currently account for the largest percentage of college freshman with disabilities

(approximately 40 percent at four-year colleges and universities; Cahalan, Mandinach, & Camara, 2002).

Nationally, numerous requests for test accommodations are made in higher education each year. Recent data from the report, "Who Took the GED?" indicated that the number of overall requests for accommodations from individuals with specific learning disabilities in 2001 increased 162 percent over the requests made in 2000 (GED Testing Service, 2002). Likewise, Camara, Copeland, and Rothschild (1998) reported that the number of examinees requesting accommodations on the SAT I: Reasoning Test doubled between 1992 and 1997. Of the requests studied, approximately 90 percent were from students with specific learning disabilities, and approximately two-thirds of the requests were for extended time (Morgan & Huff, 2002).

In general, the accommodations often approved for postsecondary students with disabilities on large-scale assessments fall under the following categories: presentation, response, scheduling/timing, and setting accommodations. Large-scale assessments such as the SAT, the American College Test (ACT), the Graduate Record Examinations (GRE), the Graduate Management Admission Test (GMAT), and the Law School Admissions Test (LSAT) commonly involve one or more of the aforementioned accommodations for examinees with disabilities. However, the number and combination of potential and allowable accommodations are quite large and varied. The number of examinees seeking accommodations over the last ten years has increased substantially, with the most notable increase in requests for extended time (Pitoniak & Royer, 2001).

Given the steady increase in the number of requests for extended time by students with learning disabilities, further research is needed to help guide college admissions officers and testing/licensing agencies in the interpretation and use of test scores from accommodated tests. In

particular, examining the role of extended time on the construct of the test is critical. If an accommodation has been shown to change the construct of the test for some individuals, the users of the test (e.g., college admissions officers; testing and licensing agencies) may question the test's ability to determine qualifications for admission, certification, or licensure (Cahalan et al., 2002).

Measurement Issues in Test Accommodations

Reliability

Reliability concerns the degree to which test scores are free from random measurement error. Because there are different types of random error, it is often necessary to evaluate different aspects of score reliability. Measuring the internal consistency of a test, which is the degree to which responses are consistent across items within a single measure, is particularly relevant to questions pertaining to the current study (Kline, 2005). If internal consistency is low, the content of the items may be so heterogeneous that the total score is not the best possible unit of analysis for the measure. Estimating the internal consistency across different groups of examinees (e.g., examinees with and without disabilities tested under standard time and extended time conditions) using the same measure provides essential psychometric information (e.g., such as the degree of variation within a population and the extent to which test scores are free from measurement error within a particular group). It is also important to note that score reliability is a necessary but not sufficient requirement for validity. That is, reliable scores may also be valid, but unreliable scores cannot be valid (Kline).

The implications of increased measurement error (thus resulting in low internal consistency/poor reliability) have been articulated by Bennett, Rock, Kaplan, and Jirele (1988), who noted that "differences in the precision of measurement across groups can have negative

effects for the group less accurately measured, thereby affecting score comparability" (p. 84). As an example, Bennett et al. referenced a situation in which an admissions officer's decision to admit a student with a disability could be more prone to error than a decision to admit a student without a disability if the score on a test used for admission was administered with accommodations to the student with a disability and thus possibly introduced measurement error.

One examination of the impact of providing accommodations on the reliability of test scores is contained in a series of studies conducted by Educational Testing Service (ETS), The College Board, and the Graduate Record Examinations Board (hereafter referred to as ETS/CB/GREB). Bennett et al. (1988) examined the reliability of the SAT and the GRE General Test by analyzing the internal consistency standard errors of measurement (SEMs). Examinees were divided into five groups based on their disability status; the reference group consisted of students without disabilities who took the SAT under standard (timed) conditions (Bennett et al.). Results indicated that the SEMs were virtually identical among all of the groups, even when SEMs were computed from parallel-forms reliability to take into account any confounding by the factor of speed. Bennett et al. concluded that the reliability of test forms administered to examinees with accommodations versus those who took a standard form of the test should not be of significant concern for psychometricians.

In 2002, Morgan and Huff conducted a similar study in which they calculated reliability estimates and estimates of the standard error of measurement (SEM) to compare the internal consistency of each SAT I test section (e.g., Verbal and Math) for students testing under standard time conditions compared to those testing with extended time. Results indicated that changes in reliability estimates were negligible across test sections, groups, and administrations; slight SEM

differences were found between the two groups of examinees (in three of the four cases the difference was approximately ten percent).

Construct Validity

Validity is indisputably a major concern of any testing program. It is in the interest of the user that a test measures what it is purported to measure, that it does not measure what it is not supposed to measure, and that it bears a reasonable relationship to the criteria it is intended to predict (Willingham, 1976). It is the responsibility of the test developer to insure that these qualities prevail in the testing program. For national assessment programs such as the SAT, which affect large numbers of individuals (approximately 1.5 million students per year register for the SAT Reasoning Test; College Board, 2005e), the principle of responsibility is particularly important.

According to the 1999 *Standards* (AERA/APA/ NCME), validity is defined as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184). Similar to the view stressed in the 1999 *Standards*, Messick (1995) asserts that validity is actually a unitary concept and that there are not different types of validity, only different types of validity evidence (see also Benson, 1998). Most forms of score validity are subsumed under the concept of construct validity, which concerns whether the scores measure the hypothetical construct the researcher (and/or test developer) believes they do. There is not a single, definitive test of construct validity, nor is it typically established in a single study (Kline, 2005).

One facet of construct validity is criterion-related validity (commonly referred to as predictive validity), which concerns whether a measure relates to an external standard (criterion) against which the measure can be evaluated (Kline, 2005). For many test uses, such as college

admissions decisions, predictive validity is considered an appropriate model for evaluating the use of a test or test battery (AERA/APA/NCME, 1999). However, while single studies of predictive validity are common practice for establishing the criterion validity of a particular measure, this method alone does not provide sufficient evidence of score validity, particularly across different groups (e.g., examinees from different ethnic groups, individuals with disabilities, and/or students with limited English proficiency).

In the following section, a number of studies examining various aspects of score validity, including criterion-related validity and construct validity, are considered to examine the degree to which tests administered with accommodations continue to assess the intended construct(s) for students with disabilities. While these facets are presented separately, it is important to stress that validity information should be viewed as an accumulation of evidence from multiple sources considered in relation to the interpretation that will be made from a given set of test scores (Haladyna & Downing, 2004).

Predictive Validity Research

In the area of admissions testing, in which there is some agreement on the appropriate criterion variable, some research examining the predictive validity of entrance exams on students' college performance has been conducted. Research into criterion validity has shown that generally, scores from accommodated administrations of entrance exams have less association (e.g., lower correlations, smaller effect sizes) with criterion measures. The criterion to which scores are typically compared is students' first-year college grade point average (GPA). Table 2.1 reviews the results from some of these research studies.

Table 2.1: Studies Examining Criterion Validity of Entrance Exams with Extended Time

Study	Measure	Results
Willingham, Ragosta, Bennett, Braun, Rock, & Powers (1988)	GRE	Scores for students with disabilities who received an extended time accommodation overpredicted first-year GPA; the greater the extended time, the greater the discrepancy.
Wightman (1993)	LSAT	LSAT scores for students with learning disabilities from tests administered with accommodations had a lower correlation with first-year law school GPA than the scores of students without disabilities who took a standard version of the LSAT.
Zurcher & Bryant (2001)	MAT	Scores for examinees with LD from extended time administration of the MAT had a weaker correlation with college GPA than the scores for students without LD who were administered the test under standard conditions.
Cahalan, Mandinach, & Camara (2002)	SAT	The correlation between first-year college GPA and GPA predicted from SAT scores was noticeably lower for students with LD who received extra time ($r = .35$, $p < .001$) than for students without LD who took the test under standard time conditions ($r = .48$, $p < .001$).

Note: GPA = grade point average; GRE = Graduate Record Examination; LD = learning disabilities; LSAT = Law School Admission Test; MAT = Miller Analogies Test; SAT = Scholastic Aptitude Test I: Reasoning Test

Based on the results of the studies noted in Table 2.1, it appears that scores from entrance examinations are less valid as predictors of postsecondary education GPA for students with learning disabilities when extended time accommodations are provided. Specifically, in all of the studies noted in Table 2.1, when accommodations were provided, the scores from entrance examinations overpredicted GPAs for students with learning disabilities.

However, what is not depicted in Table 2.1 is the common finding that when first-year GPA was predicted using both entrance exam (e.g., SAT) scores and high school GPA (combined), the reduction in predictive accuracy for the accommodated group essentially disappeared (Willingham et al. 1988; Cahalan et al., 2002; see also Sireci, Zanetti, & Berger, 2003). Furthermore, when results from the Cahalan et al. study were broken down by gender, the results for females exhibited underprediction (i.e., their actual grades were higher than predicted by .08 on a four-point scale), and the results for males "showed a trivial overprediction" of .03 (p. 16). It is also worth noting that the differences in predictive validity across standard and extended time administrations were smaller than those found across different ethnic groups who took a standard administration of the SAT I: Reasoning Test (e.g., Caucasian/African American comparisons; Bridgeman, McCamley-Jenkins, & Ervin, 2000).

Studies examining the predictive utility of a specific measure for students with learning disabilities who take an accommodated form of a test can be meaningful and provide good preliminary information about subgroup differences. However, such methods only provide information about the extent to which the measure relates to an external standard (e.g., college grade point average). To more fully understand the complexities of subgroup differences for types of disability and types of accommodations, it is necessary to go beyond descriptive statistics and closely examine the internal structure of a test (Gregg, Morgan, Hartwig, & Coleman, in press).

According to the 1999 *Standards*, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). In addition, the *Standards* have called for research investigating standardized test score validity when accommodations have been provided: "…when a test user makes a substantial change in test format, [or] mode of administration…the user should revalidate the use of the test for the changed conditions" (p. 41). The various methods used to examine the internal structure of a test are reviewed below and include factor analysis, multidimensional scaling, and differential item functioning.

Measurement Invariance Research

Factor Analysis

When test scores are to be directly compared or pooled across populations, valid comparisons require that the test measures the same construct in each population and that the relationship between test scores and scores on the construct be invariant or equivalent across populations (Millsap & Kwok, 2004). A test fulfills *measurement invariance* across populations

when individuals who are identical on the construct being measured, but who are from different populations (e.g. students with and without disabilities), have the same probability of achieving any given score on the test (Meredith & Millsap, 1992).

Among available factor analytic methods, confirmatory factor analysis (CFA) is currently one of the most important tools in the study of measurement invariance across multiple populations (Byrne, Shavelson, & Muthén, 1989; Jöreskog, 1971; Meredith, 1993; Millsap & Everson, 1993; Reise, Widaman, & Pugh, 1993; Steenkamp & Baumgartner, 1998; Vandenberg, 2002; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Factorial invariance signifies a condition in which parameters of the factor model have the same values across different groups of examinees (Millsap & Kwok, 2004). Using CFA procedures, researchers can examine factorial invariance across groups in terms of the number of factors, factor loadings, and intercepts.

Findings from studies examining measurement invariance of college entrance examinations administered with and without accommodations to examinees with and without disabilities are inconsistent (see Table 2.2). In the first two studies reviewed in Table 2.2, the researchers argued that if the relationship between items on the test and the underlying factors to which those items are linked is the same across the two populations, and if the interrelationship among the factors is similarly invariant, this supports the idea that the test scores have the same meaning for examinees who took the test with and without accommodations (Rock et al., 1988).

Table 2.2: Studies Examining Factor Structure of College Entrance Exams

Study	Measure	Hypothesized Model	Results
Rock, Bennett, Kaplan, and Jirele (1988)	SAT	Two-factor	The two factors of verbal and quantitative ability fit the data reasonably well for each of the groups of examinees with disabilities who received accommodations, although the two factors were less correlated with each other for those groups than for the examinees without disabilities, who did not receive accommodations.
Rock, Bennett, Kaplan, and Jirele (1988)	GRE	Three-factor	The analyses revealed problems with the proposed 3-factor structure; for the analytic factor, there appeared to be two factors (logical reasoning and analytical reasoning) for examinees with disabilities testing with accommodations compared to those testing under standard conditions
Morgan & Huff (2002)	SAT I	Two-factor	Results suggested that the number of dimensions estimated using item-level data were somewhat similar for examinees without LD testing under standard time conditions compared to the structure of the test for students with LD who received extended time.

Note: GRE = Graduate Record Examination; LD = learning disabilities;
SAT = Scholastic Aptitude Test; SAT I= Scholastic Aptitude Test I: Reasoning Test

In the Rock et al. (1988) study, the authors suggested the weaker correlation between the two factors for the groups of examinees with disabilities who received accommodations might have been a result of differential achievement growth in different academic areas. Rock et al. further proposed that this differential growth may have been caused by factors extraneous to the test, such as the nature of the disability and the focus of special education programs. For these reasons, they cautioned that SAT scores for examinees who took the test administered under nonstandard conditions should not be aggregated; instead, verbal and quantitative ability scores should be considered separately by college personnel making admission decisions (Rock et al.).

Morgan and Huff's (2002) study is particularly relevant to the current study as it provides a basis upon which decisions regarding research design and implementation were made. In addition, similarities between the two studies exist with regard to the population of examinees (e.g., high school students with and without learning disabilities), the measure (SAT Reasoning Test) and type of accommodation being examined (extended time). It is also worth mentioning that although results from Morgan and Huff's analyses suggested that the number of dimensions of the SAT I: Reasoning Test for examinees tested under standard time conditions were

somewhat similar to those tested with extended time, several cautionary statements were provided. For instance, Morgan and Huff suggested that additional analyses would need to be conducted to reach conclusive results, including the use of item parcels to circumvent problems associated with using dichotomously-scored individual items in principal factor analyses (see Hattie, 1985 for a review; also Carroll, 1983; McDonald & Ahlawat, 1974).

Multidimensional Scaling

A limitation of many factor analytic methods is that they must be done separately for each group (Zumbo, Sireci, & Hambleton, 2003). Therefore, such analyses may be used only as preliminary analyses to determine if there is any observable multidimensionality in the test items. In addition to using factor analysis, Zumbo et al. recommend using a second type of exploratory approach, weighted multidimensional scaling (MDS), to obtain further validity information. Research on the dimensionality of data from large-scale testing programs has shown that results from the two methods are complementary and, when considered together, quite informative (Huff & Sireci, 2001). Previous research has also shown that MDS results are roughly equivalent to conventional factor analytic results after removing the first general factor (Davison, 1985).

The purpose of MDS analyses is to discover the structure of the data simultaneously across groups while also accounting for differences in structure across the groups (Sireci & Gonzalez, 2003). To do this, a weighted MDS procedure can be used. Weighted MDS analyzes several matrices of "dissimilarity" data to derive both a common structure that best represents the data for all groups and individual group weights for adjusting this common structure to best fit the data for each specific group (Sireci & Gonzalez). The weights for each group can be used to compare the relevance of the dimensional structure across groups. The larger a weight on a

dimension, the more the dimension is necessary for accounting for the variation in the data of the specific group (Sireci & Gonzalez).

Morgan and Huff (2002) used MDS (in addition to principal factor analysis; described above) in their examination of the latent structure of the test data on the Verbal and Math sections of the SAT I: Reasoning Test for examinees testing under standard time versus those testing with extended time. Dimensionality results from the MDS analyses did not depart substantially from the principal factor analysis results, and the weighted MDS results provided additional information on the importance of the dimensions for the two groups of interest (Morgan & Huff). However, Morgan and Huff noted that the weighted MDS results were largely inconclusive due to problems with non-positive definite matrices; they recommended that their results be taken as preliminary, and suggested using item parcels to circumvent such problems in future MDS analyses.

Differential Item Functioning

A third procedure that can be used to identify areas of a test that may be inadequate for one or more of the intended groups of examinees is differential item functioning (DIF) analyses, which can be conducted to evaluate test items designed to be used across groups. For dichotomously scored items, an item is said to be functioning differentially when the probability of a correct response to the item is different for examinees at the same ability level but from different groups (Cohen, Gregg, & Deng, 2005). Although DIF analyses are useful for identifying problematic items, an evaluation of the dimensionality (through the use of factor analyses and weighted MDS) of altered (e.g. accommodated) tests is necessary to rule out systematic biases at the total test score level that are not detectable at the item level (Sireci & Gonazalez, 2003).

The presence of DIF items on a test presents a threat to the validity of scores from the test (Thissen, Steinberg, & Wainer, 1988). In this regard, Standard 7.3 in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) specifies the following:

> When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (p. 81)

Bennett et al. (1988) analyzed differential item performance in the ETS/CB/GREB study as a way to examine whether the items on the tests measured the same attribute across different groups. The SAT analysis revealed that some quantitative items on the Braille version of the test were differentially difficult for examinees with visual impairments who received accommodations, as compared with examinees without disabilities who did not receive accommodations (Bennett et al.). The authors reasoned that differential functioning for the examinees with visual impairments could be attributed to several different characteristics of items, including the presence of graphics and the presentation of miscellaneous or novel content. In contrast, some of the clustered quantitative items administered to candidates with learning disabilities and those with hearing impairments, all of whom received accommodations, were differentially easy. According to Bennett et al., this could have been due in part to the fact that, as a result of having received extra testing time, these examinees reached items that examinees without disabilities had not reached. An analysis of individual items revealed that this could be the case for some, but not all, of these items.

For the GRE, differential item functioning was generally not present for the examinees testing under standard time conditions compared to those testing with accommodations (Bennett et al., 1988). The one exception was that some students with visual impairments did better on a cluster of items located at the end of the analytical section, which they may have reached in greater proportions than other examinees. Aside from these specific instances, Bennett et al. concluded that no broad classes of items proved to function differentially for examinees with disabilities and those without disabilities.

Cohen, Gregg, and Deng (2005), using standard DIF analysis with information obtained from an alternative approach involving analysis of latent groups (mixed DIF model), investigated the performance of examinees (with and without learning disabilities) on a statewide mathematics test. The students with learning disabilities were given only extended time as an accommodation. Cohen et al. found that the cause of DIF across items on the mathematics test was less related to use of extended time than to item content (e.g., word problems, algebra). This alternative approach used an exploratory mixture item-response (IRT) model analysis (Cohen & Bolt, 2005) to identify the dimension(s) that caused the DIF and then studied examinee characteristics to support the results from other research (Bolt, Cohen & Wollack, 2002; Cohen & Bolt, 2005).

Summary of Measurement Issues

The psychometric oxymoron of an accommodated standardized test requires test developers to determine whether the accommodation changed the construct measured, and if so, the degree to which the change affected test scores (Sireci, 2005). When such research is conducted, it should include multiple sources of validity evidence including internal (e.g., factor analysis, MDS, DIF)) and external (e.g., predictive validity). Better test designs and statistical

adjustments to scores from nonstandard administrations may eliminate threats to construct validity caused by various sources of construct irrelevant variance. Nevertheless, decisions regarding the use and interpretations of scores from accommodated test administrations must be based on a comprehensive evaluation of reliability, validity, and comparability evidence for a specific testing program.

Historical Perspectives on the SAT

The following section provides an historical account of the evolution of the SAT. The first College Board SAT (the "Scholastic Aptitude Test") was administered to 8,040 students on June 23, 1926 (Lawrence, Rigol, Van Essen & Jackson, 2003). At that time, the SAT consisted of nine subtests: definitions, arithmetical problems, classification, artificial language, antonyms, number series, analogies, logical inference, and paragraph reading. Over the years, the SAT has evolved in the way it measures what is now referred to as critical thinking, reasoning, and writing skills. According to the test developers, a variety of considerations were taken into account with each redesign of the SAT, including fairness issues, scaling issues, cost, public perception, face validity, changes in the test-taking population, changes in patterns of test preparation, and changes in the college admissions process (Lawrence et al.).

A recent debate over admissions test requirements at the University of California sparked a national discussion about what is measured by the various tests—in particular, what is measured by the SAT (Lawrence et al., 2003). The fact that the SAT has been reconfigured several times over the years was frequently downplayed in the news stories. Some of the modifications have involved changes in the types of questions used to measure verbal and mathematical skills (Lawrence et al.). Other modifications focused on liberalizing time limits to ensure that speed of responding to questions has minimal effect on performance. There have

been other changes in the administration of the test, such as allowing students to use calculators on the math sections. Still other revisions have stemmed from a concern that certain types of questions might be more susceptible to coaching (Lawrence et al.).

Early Versions of the SAT (1926–1930)

The 1926 version of the SAT bears little resemblance to the current test. It contained nine subtests: seven with verbal content (definitions, classification, artificial language, antonyms, analogies, logical inference, and paragraph reading) and two with mathematical content (number series and arithmetical problems; Lawrence et al., 2003). The time limits were also quite stringent: 315 questions were administered in 97 minutes. Early versions of the SAT were quite "speeded"; as late as 1943, students were told that they should not expect to finish (Lawrence et al.). Nevertheless, many of the early modifications to the test were made in an attempt to provide more moderate time limits. In 1928, the test was reduced to seven subtests administered in 115 minutes, and in 1929, to six subtests (Lawrence et al.).

In addition to seeking appropriate time limits, developers of the early versions of the SAT were also concerned with the possibility that the test would influence educational practices in negative ways. On the basis of empirical research (Coffman, 1962) that investigated the effects of practice on the various question types, antonyms and analogies were used; research indicated that these types of questions were less responsive to practice than were some of the other question types (cited in Lawrence et al., 2003). In 1930, the SAT was organized into two sections, one segment designed to measure "verbal aptitude" and the other to measure "mathematical aptitude" (Lawrence et al., 2003, pgs. 1-2). Reporting separate verbal and mathematical scores allowed admissions staff to weight the scores differently depending on the type of college and the nature of the college curriculum.

Along with modifications noted in types of questions used to measure verbal and mathematical skills, relaxing time limits, and the way in which the test is administered, SAT score scales have also been realigned over the years. In fact, there have been over 20 different sets of scales used since the SAT exam's inception in 1926 (Dorans, 2002). The score scale is what the test-taker receives, and what the score users use. It provides the framework for the interpretation of scores, and thus the choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as test interpretation (Dorans, 2002).

For a variety of reasons related to score interpretation and psychometric issues, the original SAT scales were replaced in April 1993 by newly re-centered scales (Cook, 1994). The most important reason for this change related to the critical importance of the reference group to the meaning of the SAT score scales. The original SAT scales derived their universal meaning from a *1941 Reference Group* of slightly more than 10,000 test-takers; raw scores on the test were converted to scales scores with a mean of 500 and a standard deviation of 100. Recentering replaced this *1941 Reference Group* with the *1990 Reference Group.* According to Dorans (2002), change has occurred since 1990, but not enough to warrant discarding the *1990 Reference Group.*

Research Behind the New SAT

The College Board announced in June 2002 that a new SAT would be introduced in March 2005. According to the test developers, the content of the new test was inspired by the 1990 blue-ribbon panel report, *Beyond Prediction* (Commission on New Possibilities, 1990). The development of the actual test specifications was informed by expert advice from test development committees in reading, writing, and mathematics and by wide-ranging research studies (Kobrin & Schmidt, 2005). Research on the new SAT can be classified into three major

areas: *pre-field-trial studies*, which were conducted to inform development of the types of items to appear on the test and the overall design of the test; a *large-scale field trial*, which was conducted in March 2003 to evaluate a prototype of the new SAT; and *post-field-trial studies*, which were conducted to follow up on questions that arose during and/or after the field trial, and to examine the validity of a prototype version of the new SAT for predicting college performance (Kobrin & Schmidt).

Pre-Field Trial Studies

Content Alignment

In the spring of 2003, a survey of more than 2,000 English and language arts teachers at both the high school and college level was conducted by the College Board in an effort to better understand the reading and writing curricula in the United States. The primary objective of this large-scale, national reading and writing curriculum survey was to collect data from teachers about the frequency with which specific reading and writing skills were covered in the classrooms and how important that teachers felt these skills were for students entering their first year of college (Milewski, Johnsen, Glazer, and Kubota, 2005). Since the purpose of the new SAT is to reflect current curriculum and institutional practices in high school and college, it was critical for the College Board to examine the nature and extent of the alignment between the tested skills and curricula.

The results showed that the content covered by the critical reading and writing sections of the new SAT were in agreement with the skills rated as important by the survey respondents (Kobrin & Schmidt, 2005; see Milewski et al., 2005 for an in-depth discussion of the survey and findings). The results of the curriculum survey also suggested that the format of the SAT is aligned with the format of tests and quizzes administered b teachers. The survey indicated that

high school teachers and college professors administer multiple-choice tests (the former more than the latter) and that almost all teachers administer essay tests (Milewski et al.).

Simulation of Item Performance

One of the most significant changes to the critical reading section of the new SAT is the elimination of analogy items. Before removing these items, it was necessary for the test developers to ensure that the reliability or measurement precision of the test could be maintained without these items. Using actual SAT data, Liu, Feigenbaum, and Cook (2004) simulated verbal test forms without the analogy items. The results indicated that while it was possible to maintain the high reliability of the verbal test without the analogy items, it would be necessary to modify the distribution of item difficulties in order to obtain adequate precision at the top and bottom of the score scale (Liu et al.). That is, a higher number of very easy and very difficult items in the other verbal item types (sentence completion and reading comprehension) would be needed (Kobrin & Schmidt, 2005; see Liu et al., 2004 for an in-depth discussion of the study).

Simulation of Item Timing

Since the new SAT has different types of questions compared to previous versions, it was necessary for the College Board to determine the amount of time, on average, examinees would need to answer each question so that the number of questions and time limits of the test would be reasonable and fair. The College Board addressed these issues by conducting a series of studies in which item timing data from a computerized version of the SAT, and observation of students as they took new SAT questions under timed conditions, were used to estimate the amount of time needed to answer each type of question (Kobrin & Schmidt, 2005).

Type of Essay Prompt

The College Board also recognized that research was needed to determine the type of essay prompt to include on the writing section. A study investigating the impact of the new type of essay prompt proposed for the new SAT on ethnic, language, and gender groups was conducted by Breland, Kubota, Nickerson, Trapani, and Walker (2004). The prompts that were examined included two regular SAT II: Writing Subject Test prompts and two modifications of these prompts designed to encourage persuasive writing and provide more information to the examinee (Breland et al.). In addition, the study generated estimates of the reliability of scores obtained using the prompts examined. To examine the impact of a new prompt type, random samples of eleventh-grade students in 49 participating high schools were administered writing tests using four different prompts, two of an old type and two of a new type (Breland et al.). To obtain estimates of the reliability of scores, schools were asked to participate in a second round of testing that occurred four months after the initial testing. Results of the impact analyses revealed no significant prompt type effects for ethnic, gender, or language groups, although there were significant differences in mean scores for ethnic and gender groups for all prompts (Breland et al.).

Field Trial

In March 2003, an extensive field trial was conducted by the College Board to gain a better understanding of the proposed changes to the SAT. The purpose of the field trial was to evaluate the content, statistical, and timing specifications for the new SAT and the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT®), as well as whether scores on the new test were comparable to scores on the previous test. More than 45,000 students from 679 high schools participated in the field trial. These students were from both public and private

schools across rural, suburban, and urban areas, and represented every geographic region in the United States (Kobrin & Schmidt, 2005). Each student completed an equivalent new or previous version of the SAT or PSAT/NMSQT. To ensure that the research to address racial/ethnic group differences was based on sufficient numbers of students, a higher proportion of African American and Hispanic/Latino students were included in the field trial sample (Kobrin & Schmidt).

The results showed the following: (1) the range of item difficulties on the new critical reading and math sections were within the same range as on the previous version of these sections of the test; (2) the new test was very similar in reliability to the previous SAT; (3) the correlations between the previous SAT and the new test were very high (between .95–.97) for all three sections, suggesting that the critical reading and math scores on the new SAT could be equated to the verbal and math scores on the previous SAT; and (4) overall, the content and format changes did not appear to exacerbate the score differences between gender and ethnic groups (Liu, Feigenbaum, & Dorans, 2005).

Post-Field Trial Studies

Several studies were conducted after the field trial to evaluate different ways of refining the SAT before deciding on its final form. A separate study of about 3,000 students from the field trial suggested that the proposed time limits for the critical reading and math sections were appropriate and that extra time had virtually no effect on performance (Kobrin & Schmidt, 2005).

The field trial results indicated the need for further research on the number of writing questions to be included and on the placement of the essay for optimal performance by students (Kobrin & Schmidt, 2005). An additional field test of 6,000 students was conducted to evaluate

the effects of varying time limits and number of questions on student performance (Kobrin & Schmidt). The results prompted the College Board's decision to add an additional 10-minute section of writing questions to improve the reliability of the writing score. In a separate study, students indicated a preference for completing the essay first on the test. Performance was also lightly better on the essay when it was placed first on the examination (Kobrin & Schmidt). In its current form, therefore, the essay appears first on the new SAT.

Predictive Validity of the New SAT

One of the goals of adding a writing section to the new SAT was to improve the validity of the test for predicting college success. The American Institutes for Research, in collaboration with the College Board, recently completed a study based on approximately 1,200 first-year students from 13 colleges to examine the predictive and placement validity of the new SAT writing section (Kobrin & Schmidt, 2005). The prototype that was evaluated was 10 minutes shorter and had 12 fewer questions than the operational new writing section.

The results indicated that total scores on the writing section correlated about .46 with first year college grades, and correlated about .32 with English composition grades (Kobrin & Schmidt, 2005). In a review of four studies conducted by the College Board on the utility of a timed writing test score, the largest improvement in predictive validity ranged from .03 to .08; one study showed improvement of zero to .02 (Kobrin, n.d., p. 3). The report concludes, "Based on studies of the predictive validity of the SAT II: Writing Test, the new SAT I writing section may be expected to add modestly to the predictive validity of the SAT I" (Kobrin, n.d., p. 4).

Summary

Despite research indicating that the provision of extended time is among the most common test accommodations (Chiu & Pearson, 1999; Sireci, Li, & Scarpati, 2003; Thompson,

Blount, & Thurlow, 2002), along with the knowledge that the appropriateness of time limits is a crucial validity problem, this issue has not received sufficient attention (Lu & Sireci, 2003). Professional decisions related to selecting specific accommodations are often made on the basis of beliefs about the benefits of accommodations that may not be supported by empirical research, either because the type of empirical evidence needed is not available or because available research is not consulted (Koenig & Bachman, 2004). Although extended time is an appropriate accommodation for some individuals with learning disabilities, a great deal more experimental research involving individuals with learning disabilities is needed to determine whether test scores obtained under accommodated conditions have different meanings than scores obtained without accommodations.

The ultimate purpose of tests, whether they are gatekeeper tests such as those for college admission, licensure, and certification, or assessments given as part of a promotional process in education or employment, may factor into the issue of how extensively accommodations should be used and how test scores should be subsequently interpreted (Pitoniak & Royer, 2001). These questions are particularly relevant in the case of high-stakes tests in which the ranking of individuals can be extremely important. Both legislation and educational initiatives have strengthened the need to afford individuals with disabilities the same testing opportunities, and thus access to the same life experiences in education and employment, as individuals without disabilities (Pitoniak & Royer). Lack of access to educational attainment has an unsettling effect on career development and adult income (Bowen & Bok, 1998; Vogel & Reder, 1999).

It is likely that debate about the provision of testing accommodations to persons with disabilities will persist as long as high-stakes tests continue to be pervasive throughout our educational, professional, and licensure systems. Accommodation policymaking and practice

should be guided by empirical research and informed clinical judgment. Future research can assist in exploring the consequences of test use and provide information to test users about the validity of inferences that can be made from scores obtained from accommodated test administrations for students with disabilities.

CHAPTER III

METHODS AND PROCEDURES

Chapter three explains the methods and procedures that were employed in this study. The purpose of the study was to examine whether the new Scholastic Aptitude Reasoning Test (SAT®) measures the same construct(s) for examinees without disabilities tested under standard (e.g., timed) conditions of the SAT and students with disabilities tested with the accommodation of extended time. Specifically, an assessment of measurement invariance was conducted to determine the extent to which test scores across the two groups of examinees are comparable.

Participants

Data from the initial administration of the new SAT Reasoning Test (administered March 17, 2005) were used for the analyses. The March 2005 data set was comprised of approximately 53,680 total examinees. The volume of this administration provided sufficient numbers of examinees testing with extended time accommodations to obtain stable estimates for both the reliability and dimensionality analyses. Two groups of examinees were used in the analyses and are described below. Subjects who took the SAT Reasoning Test with only the single accommodation of extended time for documented learning disabilities and/or attention-deficit/hyperactivity disorder (AD/HD) will be referred to as "examinees with disabilities," or "students with learning disabilities." It is important to keep in mind that some students with learning disabilities and/or AD/HD may have been excluded from this sample because they (a) opted to take a standard administration, (b) do not require an accommodation, (c) for whatever reason did not take the test with an accommodation (e.g., undiagnosed disability, unaware of test accommodations), (d) reported a disability other than a learning disability or AD/HD, (e.g.,

visually impaired, deaf/hard of hearing, paraplegic, etc.), or (e) received two or more accommodations (e.g. cassette, computer, and/or script). All identifying information was removed from the database so that examinees' anonymity would be retained.

The first sample consisted of 49,504 examinees without disabilities who tested under standard time procedures. From the 49,504 cases, a random sample was drawn for this study comprised of 2,476 examinees. Of the 2,476 examinees without disabilities (EWOD) who took a standard time administration of the SAT, 54.8 percent were female and approximately 55 percent attended public school. Additional demographic information for the EWOD group is presented in Table 3.1.

Table 3.1: Demographic Profile of Examinees

SAT Reasoning Test Takers Who Described Themselves As:	**EWOD** (N = 2,476)	**EWD-T** (N = 2,476)
White	62.2%	73.3%
African American or Black Hispanic or Latino Background	8.2%	4.0%
Asian, Asian American, or Pacific Islander	9.3%	1.7%
Mexican or Mexican American	4.0%	1.1%
Puerto Rican	1.0%	0.8%
Latin American, South American, Central American, or other Hispanic or Latino	3.3%	1.7%
Other	2.9%	2.2%

** Information obtained from the optional SAT Questionnaire® examinees completed when registering for the SAT Reasoning Test*

The second sample consisted of 2,476 examinees from the original sample who had reported having learning disabilities (N = 1,517), AD/HD (N = 588) or both (N = 371), and received only the single accommodation of extended time (see Appendix B for *Eligibility Criteria and Guidelines for Documentation* for accommodations on College Board tests based on

disability). Of the 2,476 examinees with disabilities who received extended time (EWD-T), 1,962 students (79.2%) received time-and-a-half (total testing time allotted was approximately five hours, 37 minutes), and 514 examinees (20.8%) received double time (total testing time allotted was seven hours, 30 minutes, divided over two days). In addition, of the 2,476 examinees with learning disabilities and/or AD/HD who took an extended time administration of the SAT, 41.4 percent were female, and approximately 42 percent attended public school. Table 3.1 displays additional demographic information for the group of examinees with learning disabilities and/or AD/HD who received an extended time accommodation on the SAT reasoning Test (EWD-T).

Table 3.2 presents the means and standard deviations of the Critical Reading, Math, and Writing scores for each group of examinees in this study. Note that for each section, the mean scores for the extended time group (EWD-T) are lower than the mean scores for the standard time group (EWOD). Also note that in both groups, examinees performed better on the Math section compared to the Critical Reading and Writing sections of the SAT. Furthermore, as found in previous research (Morgan & Huff, 2002; Cahalan et al., 2002), there is greater variability in the scores for examinees with learning disabilities than for the group of individuals without disabilities.

Table 3.2: Group Means and Standard Deviations

	Critical Reading		**Math**		**Writing**	
	EWOD	*EWD-T*	*EWOD*	*EWD-T*	*EWOD*	*EWD-T*
Mean	516	499	530	519	515	500
Standard Deviation	105	112	106	123	106	112

Instrumentation

Measure

The current SAT is designed to measure the critical thinking, reasoning, and writing skills needed for success in college. According to the *2005-06 SAT® Program Handbook* (College Board, 2005b), the content and format of the SAT is intended to reflect accepted educational standards and practices. The Critical Reading section emphasizes reading and assesses students' ability to draw inferences, synthesize information, distinguish between main and supporting ideas, and understand vocabulary as it is used in context. The Math section requires students to apply mathematical concepts and to use data literacy skills in interpreting tables, charts, and graphs. The Writing section includes both multiple-choice questions that deal with the mechanics of writing and a direct writing measure in the form of an essay (College Board, 2005a). The SAT consists of 10 sections, including a 25-minute essay, with each section timed separately. The essay always appears first, and the six other 25-minute sections can appear in any order, as can the two sections that are 20 minutes each. In addition, a 10-minute writing multiple-choice section appears at the end of the test. See Table 3.3 for test content and format for each version of the SAT.

Table 3.3: SAT Test Content and Format

Section	*Content*	*Number of Questions*	*Time*
Critical Reading	Extended Reasoning	36-40	70 minutes
	Literal Comprehension	4 to 6	(two 25-minute sections
	Vocabulary in Context	4 to 6	and one 20-minute section)
	Sentence Completion	19	
	Total	**67**	
Math	Number and Operations	11 to 14	70 minutes
	Algebra and Functions	19 - 22	(two 25-minute sections
	Geometry and Measurement	14 - 16	and one 20-minute section)
	Data Analysis, Statistics, and Probability	5 to 8	
	Total	**54**	

Writing	Essay	1	60 minutes
	Improving Sentences	25	(one 25-minute essay,
	Identifying Sentence Errors	18	one 25-minute
	Improving Paragraphs	6	multiple choice section,
	Total	**50**	and one 10-minute
			multiple-choice section)

There is an additional 25-minute section, called an "equating" or variable section, which may be a Critical Reading, Math, or Writing multiple-choice section. The placement of this equating section varies on different editions of the test, and it does not count toward the final score. Its purpose is to try out new questions for future editions of the SAT and to help ensure that scores on new editions of the SAT are comparable to scores on earlier editions of the test (College Board, 2005a).

Scores

Each SAT score is reported on a scale from 200 to 800 points. An examinee's scaled score is computed by first establishing a raw score. For each correct answer, the student earns one point; for a wrong answer to a multiple-choice question, the student loses one-quarter point. No points are deducted for wrong answers to questions that require student-produced responses, or for unanswered questions (College Board, 2005c).

Students receive three scores on the 200 to 800 scale: one for Critical Reading, one for Math, and one for Writing. Students also receive two Writing subscores: a multiple-choice score from 20 to 80 and an essay score from 2 to 12. The total Writing score is a combination of the multiple-choice and essay scores and is reported on the 200 to 800 scale. The essay makes up approximately 30 percent of the total Writing section score (College Board, 2005c).

Essays are scored using a holistic approach by experienced and trained high school and college teachers. Each essay is scored by two people who do not know each other's score, and who do not know the student's identity or his or her school. Each reader takes into account such

aspects as complexity of thought, substantiality of development, and facility with language (College Board, 2005c).

The average score on the SAT is about 500 on the Critical Reading section and 500 on the Math section. Average scores for the Writing section will not be available until 2006, after the test has been administered for a year. The questions on the SAT Reasoning Test range in difficulty from easy to hard, with the majority being of medium difficulty. Medium-difficulty questions are answered correctly by about one-third to two-thirds of students. The SAT is designed so that a student who answers about half the questions correctly will receive an average score (College Board, 2005c).

Considering the significant changes that were made to the SAT in 2005, a common question among many students, parents, and admissions counselors relates to the comparability of scores between the new versions of the test compared to previous versions. The College Board's field trial of more than 45,000 students confirmed that scores on the SAT Critical Reading section are comparable to scores on the former SAT Verbal section, and scores on the SAT Math section are comparable to scores on the former SAT Math section. For 2005 college-bound seniors, the Critical Reading (verbal) mean was 508 and the Math mean was 520. More information about mean scores, including breakdowns by gender and ethnic groups, is available on the College Board web site (www.collegeboard.com/satdata).

Procedures

The current study was designed to compare the following two groups: (1) examinees without disabilities (EWOD) who took the new SAT with a standard time administration and received no accommodation, and (2) examinees with learning disabilities and/or AD/HD who took the SAT with a single accommodation of extended time (EWD-T). Although the advantage

of using subscales in factor analytic studies have been documented (Wainer & Keily, 1987; Dorans & Lawrence, 1987; 1999), specifically in regard to SAT data, subscale analyses can guide the diagnostic process only when the subscales have an interpretable dimensional structure. Based on the lack of evidence of this in the preliminary analyses, formula-scored item-level data were factor analyzed in this study.

Statistical Analyses

To examine the extent to which reliability estimates vary across test sections and groups, reliability estimates for the two groups of interest were calculated to compare the internal consistency of each test section across groups. Since reliability estimates are a function of the variation within a population, differences in measurement error were further assessed by examining the standard error of measurement (SEM) for both groups across each test section.

Drawing from the previous research carried out by Rock et al. (1988) and Morgan and Huff (2002), this study examined the internal consistency and dimensionality of the Critical Reading, Math, and Writing sections of the new SAT. Specifically, the question of whether the factor structure of the Critical Reading, Math, and Writing Sections of the new SAT was invariant across different populations (e.g., examinees who tested under standard conditions and those who received extended time) was examined, since score validity requires the latent structure of an assessment to be the same across varying groups of examinees. Given that the items being analyzed are dichotomous (binary), a tetrachoric correlation matrix (obtained from M*plus* v4.0; Muthén & Muthén, 2006) was used to obtain the input matrices. Six matrices were computed and compared across the variable of interest (i.e., timing condition) for each test section.

Matrix	***Test Section***	***Group***
Matrix 1	Critical Reading	EWOD
Matrix 2	Critical Reading	EWD-T
Matrix 3	Math	EWOD
Matrix 4	Math	EWD-T
Matrix 5	Writing (Multiple Choice)	EWOD
Matrix 6	Writing (Multiple Choice)	EWD-T

The primary analyses were conducted in two different stages, with the first being a prerequisite for the second. In the first stage, a total of six competing models concerning the constructs that are measured by the SAT Reasoning Test were evaluated for model fit and parsimony. The adequacy of fit of two plausible models corresponding to each section of the SAT Reasoning Test were evaluated separately for the two groups of interest. In the second stage, three levels of model invariance for the final three models were tested across the two groups for the Critical Reading, Math, and Writing sections of the SAT Reasoning Test.

A set of procedures has been developed to assess questions of measurement invariance across groups using structural equation modeling (SEM) techniques (Millsap & Yun-Tein, 2004; Muthén & Muthén, 2006). Using these procedures, a series of increasingly restrictive constraints is imposed to force the model parameters to be equal across groups. These procedures result in a series of nested models that can be tested through the use of chi-square difference tests ($\Delta\chi^2$). The significance tests provide a test of whether imposition of the equality constraints in the more constrained model resulted in a significant decrement in the fit of the model across groups (Gregg, Bandalos, Coleman, Davis, Jiménez, Robinson, & Blake, in press). A significant $\Delta\chi^2$ difference implies that values of the parameters held invariant at that step actually differ significantly across groups.

Millsap and Yun-Tein (2004) developed a set of minimal across-group invariance restrictions on thresholds and other parameters intended to provide sufficient conditions for

identification and comparisons of multiple-group model testing using Mplus. When factor indicators are categorical, thresholds are modeled rather than intercepts or means (Muthén & Muthén, 2006). To make meaningful comparisons of factor distributions across groups, a majority of the variables should have both loading and threshold invariance so that the factors not only are in the same metric technically, but so that it is also plausible that the variables measure factors with the same meaning in the different groups (Millsap & Yun-Tein). Then, constraints can be further increased by holding factor variances equal, in addition to the factor loadings and thresholds, across groups. In the current study, this series of tests was used to determine whether the values of these parameters (e.g., factor loadings and thresholds and factor variances) differed significantly across the two groups.

Data Analyses

A categorical data mean-structure model analyzed with the weighted-least-squares parameter estimates with robust standard errors and mean- and variance-adjusted chi-square (χ^2) statistics (WLSMV) was employed using Mplus v4.0 (Muthen & Muthen, 2006). Theta parameterization was used as it is preferred when hypotheses involving residual variances are of interest, as is the case with multiple group analysis (Muthén & Muthén). When examining results of these analyses it is important to note that WLSMV χ^2 statistics and degrees of freedom (*df*) are calculated in a way different to that used for common estimation methods such as maximum-likelihood (Muthén & Muthén). Categorical item responses for the 170 SAT items were the basis of analysis.

Mplus v4.0 with the WLSMV estimator provides several absolute measures of goodness-of-fit, including the adjusted χ^2 and estimates of the Standardized Root Mean Residual (SRMR) and the Weighted Root Mean Square Residual (WRMR) in addition to two incremental fit

indices, the non-normed fit index or Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI). Full descriptions of these various goodness-of-fit statistics are available in a variety of places (e.g., Byrne, 1998; Hu & Bentler, 1999; Muthén & Muthén, 2006). According to current recommendations, a good fit to the data would be indicated by an SRMR value of less than 0.08, TLI and CFI values greater than 0.95, and a WRMR less than 1.0 (Hu & Bentler, 1999; Muthén & Muthén, 2006; Yu, 2002).

CHAPTER IV

RESULTS

The purpose of this study was to investigate the factor structure of the Critical Reading, Math, and Writing sections of the newly revised SAT Reasoning Test. In particular, this study was conducted to explore the relationship between the latent constructs of critical thinking, reasoning, and writing, and items used to measure these constructs across groups of examinees without disabilities tested under standard (e.g., timed) conditions of the SAT and examinees with disabilities tested with the accommodation of extended time.

The objective of this chapter is to present the results of the study. First, the findings from the analyses of internal consistency across the two groups for each of the three sections of the SAT will be considered. Second, the results of the preliminary single group analyses examining the factor structure of the Critical Reading, Math, and Writing sections are presented. Lastly, the results of the assessment of measurement invariance across the two groups of interest will be discussed. A presentation of the statistical data will be provided; this data will help answer the question of whether test scores for examinees without disabilities (EWOD) who took a standard (e.g., timed) administration of the SAT are comparable to the scores for students with disabilities tested with extended time (EWD-T).

Internal Consistency

The reliability and standard error of measurement (SEM) estimates for the two groups of interest for the Critical Reading, Math, and Writing sections of the SAT are presented in Table 4.1. Results indicated that changes in reliability estimates were negligible across test sections and

groups; slight SEM differences were found between the two groups of examinees. These results are similar to those of Morgan and Huff (2002).

Table 4.1: Reliability and SEM Estimates

	Critical Reading		**Math**		**Writing**	
	Standard Time	*Extended Time*	*Standard Time*	*Extended Time*	*Standard Time*	*Extended Time*
Reliability Coefficient	0.927	0.934	0.924	0.94	0.892	0.897
SEM	28.43	28.95	29.33	30.15	34.88	36.21

Factor Models

The factor models used in the preliminary analyses of the current study allowed tests of the hypothesis concerning the possible threat to validity from allowing examinees with disabilities extended time to take the SAT Reasoning Test. A total of six competing models concerning the constructs that are measured by the SAT Reasoning Test were evaluated for model fit and parsimony; these factor models are shown in Table 4.2. Item type (according to the College Board's test specifications) was used as a guiding principle for developing an alternate model for the Critical Reading, Math, and Writing sections; this alternate model was then evaluated and compared to a general factor model.

Table 4.2: Factor Models

SAT Test Section	***Model***	***Description of Model***
Critical Reading	One-factor	General Factor
Critical Reading	Two-factor	Sentence Completion; Reading Passages
Math	One-factor	General Factor
Math	Three-factor	Number & Operations; Algebra; Geometry
Writing	One-factor	General Factor
Writing	Three-factor	Sentence Correction; Usage; Revision in Context

All models were estimated using mean- and variance-adjusted weighted least squares parameter estimates (designated in Mplus as WLSMV; Muthén & Muthén, 2006); tetrachoric

correlation matrices were used as the input matrices. Mplus v4.0 with the WLSMV estimator provides several absolute measures of goodness-of-fit including the adjusted chi-square (χ^2) and estimates of the Standardized Root Mean Residual (SRMR) and the Weighted Root Mean Square Residual (WRMR) and two incremental fit indices, the non-normed fit index or Tucker-Lewis index (TLI) and the Comparative Fit Index (CFI). Full descriptions of these various goodness-of-fit statistics are available in a variety of places (e.g., Byrne, 1998; Hu & Bentler, 1999; Muthén & Muthén, 2006). Of note, the adjusted χ^2 and degrees of freedom (*df*) available with WLSMV method cannot be used for comparisons of nested models in the usual way, but requires a special difference test available with Mplus Version 4.0 (Muthén & Muthén).

According to current recommendations, a good fit to the data would be indicated by an SRMR value of less than 0.08, TLI and CFI values greater than 0.95, and a WRMR less than 1.0 (Hu & Bentler, 1999; Muthén & Muthén, 2004; Yu, 2002). Differences in fit between nested models (i.e., a model with constraints compared to a model without constraints) were assessed by the chi-square difference test ($\Delta\chi^2$) and inspection of changes in other fit indices. The $\Delta\chi^2$difference test was the primary statistic used to assess changes in model fit.

Despite its common use to evaluate model fit in confirmatory factor analyses, the influence of sample size (*N*) on the adjusted χ^2 index warrants a brief discussion. A problem arises because of the statistic's functional dependence on *N*. For large sample sizes (e.g., > 500), the χ^2 statistic provides a highly sensitive statistical test, but not a practical test, of model fit (Cheung & Rensvold, 2002). The value of the χ^2 statistic may lead to rejection of the model(s) even though differences between observed and predicted correlations are slight (Kline, 2005). Its relevance to the current study is also worth mentioning. When conducting confirmatory factor analyses with multiple groups, which test the invariance of estimated parameters of two nested

models across groups, researchers have demonstrated that differences in χ^2 ($\Delta\chi^2$) are also dependent on sample size (Brannick, 1995; Kelloway, 1995). Thus, if sample sizes are large, even a small difference between the nested models may result in a significant value of the $\Delta\chi^2$ test, indicating that the null hypothesis of no difference should be rejected even when the difference is trivial (Brannick; Kelloway). The question then becomes one of statistical significance versus practical significance. Therefore, the focus of the following interpretation of model fit should not be on the magnitude of the indexes (particularly the values of the χ^2 and $\Delta\chi^2$ fit indexes) but rather on the changes in goodness-of-fit indexes (e.g. CFI, TLI, SRMR, WRMR) between models for each group.

Evaluation of Model Fit

Critical Reading

First, the factor structure of the Critical Reading section was examined. The fit of the two-factor model, in which sentence completion items loaded on one factor and reading passages loaded on a second factor, compared with a one-factor model, was first evaluated for each of the two groups, separately. These model comparisons, reported in Table 4.3, showed similar results in model fit for the two groups of interest. Specifically, for the EWOD and EWD-T groups, both models (one-factor and two-factor) met two of the four a priori specified criteria for acceptability (i.e. SRMR $\leq .08$; TLI $\geq .95$). As shown in Table 4.3, the χ^2 value for both models in each group is statistically significant, resulting in a rejection of the null hypotheses. However, the χ^2 fit statistic is affected by sample size and therefore should be interpreted with caution (Kline, 2005).

Table 4.3: Goodness of Fit Statistics by Group and Factor Model: Critical Reading

Group and model	$\chi 2$	*df*	*p*-value	CFI	TLI	SRMR	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*	*p*-value
EWOD										
One-factor model	2588.039	1000	0.000	0.944	0.980	0.045	1.372	-	-	-
Two-factor model	2475.503	1000	0.000	0.948	0.981	0.044	1.341	119.965	1	0.000
EWD-T										
One-factor model	2742.922	1002	0.000	0.941	0.981	0.045	1.393	-	-	-
Two-factor model	2613.137	1003	0.000	0.945	0.983	0.044	1.358	125.284	1	0.000

Note. EWOD = Examinees without disabilities; EWD-T = Examinees with disabilities who received extended time; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Residual; WRMR = Weighted Root Mean Residual; $\Delta\chi 2$ = Chi-square difference test

It is also worth mentioning that for both groups, the two-factor model demonstrated statistically significant improvement over the one-factor model ($\Delta\chi^2_{\text{EWOD}}$ [1, N = 2,476] = 119.965, $p < .001$; $\Delta\chi^2_{\text{EWD-T}}$ [1, N = 2,476] = 125.284, $p < .001$). Again, however, the significant value of the $\Delta\chi^2$ may be a function sample size. In addition, the high correlations among the factors (EWOD: 0.922; EWD-T: 0.921) of the two-factor model comprising the Critical Reading section demonstrated a high degree of shared variance among the factors for both groups. This is not surprising considering the nearly equivalent fit of the one-factor model; the high correlations among the two factors may also suggest that these factors may not be distinct constructs.

When evaluating the model fit of the two-factor model in comparison with a one-factor model for the Critical Reading section of the SAT, it is also important to consider the issue of parsimony. That is, the model that fits best according to the fit indexes may not be the "best" model. According to Kline (2005), if a single-factor model cannot be rejected, there is little reason to evaluate more complex models, even when theory supports a multidimensional model. Furthermore, results of the two-factor model in the current study suggest poor discriminant validity among the two factors, as evidenced by equally high factor correlations (0.92) for both groups. Taken together, findings indicate that the Critical Reading section of the SAT Reasoning Test measures a unidimensional construct for the EWOD and EWD-T groups. Although the $\Delta\chi^2$

difference test for the two-factor model demonstrated statistically significant improvement over the one-factor model ($\Delta\chi^2_{EWOD}$ [1, N = 2,476] = 119.965, $p < .001$; $\Delta\chi^2_{EWD\text{-}T}$ [1, N = 2,476] = 125.284, $p < .001$), the magnitude of change is quite small.

Math

Next, the factor structure of the Math section of the SAT was examined in each group. Specifically, goodness of fit between a general factor model and a three-factor model, in which Number and Operations items loaded on one factor, Algebra items loaded on a second factor, and Geometry items loaded on a third factor, was evaluated for each of the two groups, separately. These model comparisons, reported in Table 4.4, reveal near equivalent model fit for both models across the two groups of interest, with the exception of the χ^2 values. Despite highly discrepant χ^2 values between the two groups for the unidimensional and multidimensional models, all other goodness-of-fit statistics were remarkably similar. In addition, for the EWOD and EWD-T groups, both models (one-factor and three-factor) met two of the four a priori specified criteria for acceptability (i.e. SRMR ≤ .08; TLI ≥ .95), as well as a third criteria for the EWOD group (CFI ≥ .95). The χ^2 value for both models in each group is statistically significant, resulting in a rejection of the null hypotheses. However, as noted above, the χ^2 fit statistic is affected by large sample sizes; the value of the χ^2 may lead to rejection of the model(s) even though differences between observed and predicted correlations are slight (Kline, 2005).

Table 4.4: Goodness of Fit Statistics by Group and Factor Model: Math

Group and model	$\chi 2$	*df*	*p*-value	CFI	TLI	SRMR	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*	*p*-value
EWOD										
One-factor model	1766.656	737	0.000	0.961	0.987	0.05	1.295	-	-	-
Three-factor model	1735.756	736	0.000	0.962	0.987	0.05	1.283	48.341	3	0.000
EWD-T										
One-factor model	2729.187	721	0.000	0.935	0.984	0.057	1.573	-	-	-
Three-factor model	2707.105	720	0.000	0.936	0.984	0.057	1.566	40.163	3	0.000

Note. EWOD = Examinees without disabilities; EWD-T = Examinees with disabilities who received extended time; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Residual; WRMR = Weighted Root Mean Residual; $\Delta\chi 2$ = Chi-square difference test

Also, it is worth noting that in both groups, the three-factor model demonstrated statistically significant improvement over the one-factor model, $\Delta\chi^2_{\text{EWOD}}$ (3, N = 2,476) = 48.341, $p < .001$; $\Delta\chi^2_{\text{EWD-T}}$ (3, N = 2,476) = 40.163, $p < .001$. The significantly high $\Delta\chi^2$ test statistic, however, may be an artifact of large sample sizes. In addition, the high correlations among the factors (EWOD: 0.972, 0.963, 0.961; EWD-T: 0.980, 0.979, 0.972) of the three-factor model comprising the Math section demonstrated a high degree of shared variance among the factors for both groups. Considering the similar fit between the one-factor model and the three-factor model, these results are not surprising. The high correlations among the three factors may also suggest that these factors may not be distinct constructs.

When evaluating the model fit of the three-factor model in comparison with a general factor model for the Math section of the SAT, it is again important to consider the issue of parsimony. Given the negligible differences between the values of the goodness-of-fit indexes between models within each group, there is not strong evidence in favor of one model over another. In addition, results of the three-factor model (in both groups) suggested poor discriminant validity among the factors, as evidenced by very high factor correlations (0.96-0.98) for both groups. Taken together, findings indicate that the Math section of the SAT Reasoning Test measures a unidimensional construct for EWOD and EWD-T groups. Although the $\Delta\chi^2$

difference test for the two-factor model demonstrates statistically significant improvement over the one-factor model ($\Delta\chi^2_{EWOD}$ [3, N = 2,476] = 48.341, $p < .001$; $\Delta\chi^2_{EWD\text{-}T}$ [3, N = 2,476] = 40.163, $p < .001$), the relative degree of change is quite small when considering the remarkably high χ^2 values for both models in each group.

Writing

Finally, the third step in the preliminary set of analyses examined the factor structure of the Writing section of the SAT. Goodness of fit between a general factor model and a three-factor model was evaluated in both groups, separately. For the three-factor model, the item types included Sentence Corrections, Usage, and Revision in Context, all of which loaded on three corresponding factors. The model comparisons reported in Table 4.5 reveal similar patterns in model fit between the two competing models and across the two groups of interest. For instance, similar to the results of model fit of the Critical Reading and Math sections, the TLI and SRMR values for both models and both groups met a priori specified criteria for acceptability (i.e. TLI ≥ .95; SRMR ≤ .08).

Table 4.5: Goodness of Fit Statistics by Group and Factor Model: Writing

Group and model	χ2	*df*	*p*-value	CFI	TLI	SRMR	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*	*p*-value
EWOD										
One-factor	1714.942	657	0.000	0.945	0.974	0.049	1.393	-	-	-
Three-factor	1525.266	657	0.000	0.955	0.979	0.047	1.312	193.563	3	0.000
EWD-T										
One-factor	1896.413	662	0.000	0.938	0.973	0.050	1.448	-	-	-
Three-factor	1788.216	664	0.000	0.943	0.976	0.049	1.402	112.026	3	0.000

Note. EWOD = Examinees without disabilities; EWD-T = Examinees with disabilities who received extended time; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Residual; WRMR = Weighted Root Mean Residual; Δχ2 = Chi-square difference test

In addition, the χ^2 value for both models in each group was statistically significant, resulting in a rejection of the null hypotheses. However, as mentioned above, the χ^2 fit statistic is

affected by large sample size, and likely led to rejection of the model(s) even though differences between observed and predicted correlations were slight (Kline, 2005). Furthermore, for the EWOD and EWD-T groups, the moderate to high correlations among the three factors (EWOD: 0.956, 0.829, 0.838; EWD-T: 0.949, 0.826, 0.848) comprising the Writing section demonstrated a reasonably high degree of shared variance among the factors, though were slightly lower in comparison to the correlations among the factors comprising the Critical Reading and Math sections.

Inspection of Table 4.5 also reveals a notable difference in $\Delta\chi^2$ values that was not observed in the evaluation of model fit for the Critical Reading and Math sections in each of the two groups of interest. Although the three-factor model for both the EWOD and EWD-T groups demonstrated statistically significant improvement over the one-factor model for the Writing section, the degree of improvement is quite inconsistent between groups ($\Delta\chi^2_{EWOD}$ [3] = 193.563, $p < .001$; $\Delta\chi^2_{EWD\text{-}T}$ [3] = 112.026, $p < .001$). However, given the considerably high χ^2 values for all models tested, the degree of discrepancy between the values of the $\Delta\chi^2$ difference test across groups is relatively small and of little practical significance. Similarly, despite $\Delta\chi^2$ difference tests for the three-factor model demonstrating statistically significant improvement over the one-factor model, the magnitude of change is quite small when taking into account the considerably high χ^2 values for both models in the EWOD and EWD-T groups.

Adequacy of the One-Factor Model Across Sections and Groups

As a preliminary step in the current study, the adequacy of fit of the two-factor model (Math) and three-factor model (Critical Reading and Writing) in contrast to a one-factor model had to be established for each of the two groups separately. When the fit of one-factor and multifactor models was evaluated for each group, the results were similar across groups for both

models. Specifically, within each group, the TLI and SRMR values for both models met a priori specified criteria for acceptability (i.e. SRMR ≤ .08; TLI ≥ .95). The CFI and WRMR values consistently fell just below the recommended cutoff value (i.e., ≥ CFI .95) for both models in each group, with the exception of EWD-T group in the Math section ($CFI_{Model\ 1}$: 0.961; $CFI_{Model\ 2}$: 0.962). At the same time, the χ^2 statistics of both competing models for each of the three sections of the SAT Reasoning Test were significant in each group, resulting in a rejection of the null hypotheses. Furthermore, the $\Delta\chi^2$ difference tests indicated that the two-factor model (Math) and three-factor models (Critical Reading and Writing) offered a statistically significant improvement in fit over the one-factor model for each group.

However, with regard to the aforementioned statistically significant χ^2 and $\Delta\chi^2$ values for all models tested, it is important to take into account the influence of sample size on the χ^2 fit statistic. It is likely the large sample sizes used in the current study led to rejection of the model(s) even though differences between observed and predicted correlations were slight. Also, despite statistically significant $\Delta\chi^2$ tests suggesting improvement of fit in the two-and three-factor models over the one-factor models, the magnitude of change is actually quite small when taking into account the considerably high χ^2 values for both models in the EWOD and EWD-T groups.

As a final note, evaluation of the intercorrelations among the factors comprising the Critical Reading, Math, and Writing sections in the two-and three-factor models indicated a high degree of shared variance among the factors (Critical Reading: 96-98%; Math: 96-98%; Writing: 83-96%) for both groups. Considering the negligible differences in fit between the one-factor model and the two-factor (Reading) and three-factor (Math and Writing) models across groups, results suggested poor discriminant validity. Taken together, findings indicated that the Critical

Reading, Math, and Writing sections of the SAT Reasoning Test each measure a unidimensional construct for the two groups of interest. As a result, the one-factor model was considered adequate for the subsequent invariance analyses.

Evaluation of Fit Invariance Across Groups

After the fit of the one-factor model had been established, the comparability of the general factor model for the Critical Reading, Math, and Writing sections of the SAT across the two groups was examined. As explained previously, multiple group analyses involve a series of hierarchically nested models. A set of procedures was followed to assess questions of measurement invariance across groups and included the following: (1) factor loadings and observed variable thresholds were freely estimated in both groups, residual variances were fixed to one in both groups (for identification purposes; Muthén & Muthén, 2006), and factor means were fixed to zero in both groups; (2) in addition to step one, factor loadings and thresholds were constrained to be equal across groups; and (3) in addition to step two, factor variances were constrained to be equal across groups. Also, in steps two and three, the factor means were fixed at zero in the EWD-T group and were free to be estimated in the EWOD group (Muthén & Muthén). This particular set of models is similar to those recommended by Millsap and Yun-Tein (2004) and Muthén and Muthén for measurement invariance of categorical outcomes.

In the invariance analyses described above, each model was evaluated for both groups simultaneously. The purpose of these analyses was to determine the model that best balanced the level of invariance and the fit across the groups. The tenability of the constraints imposed by each of the models in the hierarchy was gauged by a chi-square difference test ($\Delta\chi^2$). In these tests, each model was compared to the previous model to test whether the imposition of the additional constraints resulted in significantly worse fit. Such a finding indicates that the

additional parameters held invariant in that model vary significantly across groups. Differences in the CFI values between the two models were also considered, following Cheung and Rensvold's (2002) suggestion that a difference of .01 or greater is indicative of a significant decrement in fit. If a set of parameters was found to lack invariance, modification indexes were examined in an effort to determine which parameter(s) were causing the lack of fit.

Finally, goodness-of-fit statistics were also examined for each model in the invariance analyses. According to current recommendations, a good fit to the data would be indicated by an RMSEA value of less than 0.06, TLI and CFI values greater than 0.95, and a WRMR less than 1.0 (Hu & Bentler, 1999; Muthén & Muthén, 2006; Reilly, Bowden, Bardenhagen, & Cook, in press).

Invariance Levels of the Critical Reading One-Factor Model Across Groups

Table 4.6 displays the fit indexes for the one-factor multiple group model at three levels of invariance for the Critical Reading section of the SAT Reasoning Test. Again, it is important to point out the influence of sample size on fit statistics. In particular, the χ^2 and $\Delta\chi^2$ fit indexes are sensitive to sample size. Therefore, the focus of this interpretation should not be on the magnitude of the indexes (particularly the values of the χ^2 and $\Delta\chi^2$ fit indexes) but rather on the changes in goodness-of-fit indexes (e.g. CFI, TLI, RMSEA, WRMR) as constraints increase.

For the Critical Reading section, assessment of the baseline model (Table 4.6; Model 1) resulted in a good fit to the data. While the χ^2 was significant at 5,328.89 with 2,002 degrees of freedom, the TLI (0.981), RMSEA (0.026), and WRMR (1.926) all indicated good fit for Model 1; the CFI (0.942) indicated borderline fit. As a means of further verifying the stability of the unconstrained model, the factor loadings for the Critical Reading section were evaluated for each group separately (see Appendix C for standardized factor loadings). Examination of the one-

factor model for the EWOD revealed that 57 of the 67 standardized pattern coefficients had values greater than 0.50, which would indicate that these items measure the unidimensional construct of reading reasonably well. Similarly, in the EWD-T group, 58 of the 67 standardized pattern coefficients had values greater than 0.50, which again suggested that the majority of the items are good measures of the reading factor. It is also worth mentioning that of the nine items (Items 6, 8, 19, 24, 30, 35, 42, 51, and 58) with factor loadings less than 0.50 in the EWD-T group, seven (of 10) of these same items were also found to have factor loadings less than 0.05 in the EWOD group (Items ***6, 8, 19***, 31, ***35***, 38, ***42, 51***, ***58***, and 62). Inspection of these parameter estimates showed that group differences were nominal with regard to factor loading values across groups.

Table 4.6: Goodness of Fit for Invariance Analyses Across Groups: Critical Reading

Invariance Model	χ^{2a}	*df*[a]	CFI	TLI	RMSEA	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*
1. Factor loadings and thresholds freely estimated in both groups; residual variances fixed to one and factor means fixed to zero in both groups	5,328.889	2,002	0.942	0.981	0.026	1.926	-	-
2. Model 1 and all factor loadings and thresholds were constrained to be equal across groups	3,265.278	1,302	0.966	0.982	0.025	2.039	120.499**	71
3. Model 2 and factor variance invariant across groups	2,214.441	868	0.977	0.982	0.025	2.077	4.690*	1

Note. All $\Delta\chi2$ values based on comparison to the previous model.
* $p < .05$; ** $p < .001$
[a] The $\chi2$ value and degrees of freedom (*df*) are calculated differently for the WLSMV estimator and require a special difference test (see Muthén & Muthén, 2006).

Next, the estimates of the factor loadings and thresholds across groups were held invariant. As shown in Table 4.6 (Model 2), the imposition of cross-group constraints on the factor loading and threshold estimates resulted in a decrement in WLSMV difference χ^2 that was statistically significant ($\Delta\chi^2$ [71] = 120.499, $p < .001$) and a change of +0.113 in the WRMR

value, which also suggested a slight loss of fit. However, the RMSEA, CFI, and TLI statistics indicated that there was a slight improvement in fit from the baseline model. Therefore, although the $\Delta\chi^2$ test associated with full restrictions on the factor loadings and observed variable thresholds was statistically significant, the absolute change in other fit statistics was relatively small and inconsequential. Therefore, results support the assumption of strong metric invariance for the Critical Reading section; that is, essentially the same pattern and equality of factor loadings and the numerical equality of variable thresholds were observed across groups.

In Model 3 (Table 4.6), an additional constraint was added to the model; the factor variance was constrained to be equal across groups (in addition to restricting factor loadings and thresholds to be invariant across groups). Again, there was little to no change in the goodness-of-fit statistics (e.g., CFI, TLI, and RMSEA values) when compared to the fit of Model 2. In addition, according to the $\Delta\chi^2$ test, imposing cross-group constraints on the factor variance in addition to the factor loading and threshold estimates resulted in only a slight decrement in fit between Model 2 and Model 3 ($\Delta\chi^2$ [1] = 4.690, $p < .05$). Although a statistically significant $\Delta\chi^2$ difference test implies that values of the parameter (e.g. factor variance) held invariant at this step actually differ significantly across groups, the magnitude of change was relatively small and of little practical significance. Thus, in view of the lack of evidence of loss of fit, these results suggested that for the Critical Reading section, invariance of factor loadings, observed variable thresholds, and factor variances across the two samples is a tenable hypothesis.

Invariance Levels of the Math One-Factor Model Across Groups

Table 4.7 displays the fit indexes for the one-factor multiple group model at three levels of invariance for the Math section of the SAT Reasoning Test. Again, considering the influence of sample size on fit statistics, the focus of this interpretation should not be on the magnitude of

the indexes (particularly the values of the χ^2 and $\Delta\chi^2$ fit indexes) but rather on the changes in goodness-of-fit indexes (e.g. CFI, TLI, RMSEA, WRMR) as constraints increase.

Assessment of the baseline model resulted in a moderately good fit to the data. While the χ^2 was significant at 4,469.396 with 1,324 degrees of freedom, the TLI (0.985) and RMSEA (0.029) both met the a priori specified criteria for acceptability (i.e. TLI ≥ .95; RMSEA < .06) fit for Model 1. The CFI (0.948) and the WRMR (2.0) indicated borderline fit. As a means of further evaluating the unconstrained model, the factor loadings for the Math section were evaluated for each group separately (see Appendix D for standardized factor loadings). Examination of the one-factor model for the EWOD group revealed that 46 of the 54 standardized pattern coefficients had values greater than 0.50, which would suggest that these items measure the unidimensional construct of math reasonably well. In the EWD-T group, 51 of the 54 standardized pattern coefficients had values greater than 0.50, again suggesting that the majority of the items are good measures of the math factor. It is also worth mentioning that all three of the items (Items 20, 47, and 53) with factor loadings less than 0.50 in the EWD-T group were represented in the group of items with factor loadings less than 0.05 in the EWOD group (Items 1, 4, ***20,*** 39, 41, ***47,*** 49, and ***53***). Inspection of these parameter estimates showed that, for the majority of the items, group differences were small with regard to factor loading values across groups.

Table 4.7: Goodness of Fit for Invariance Analyses Across Groups: Math

Invariance Model	χ^{2a}	*df*[a]	CFI	TLI	RMSEA	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*
1. Factor loadings and thresholds freely estimated in both groups; residual variances fixed to one and factor means fixed to zero in both groups	4,469.396	1,457	0.948	0.985	0.029	2.000	-	-
2. Model 1 and all factor loadings and thresholds were constrained to be equal across groups	2,519.785	862	0.971	0.986	0.028	2.160	123.148**	54
3. Model 2 and factor variances invariant across groups	2,081.648	531	0.973	0.979	0.034	2.538	39.195**	1

Note. All Δχ2 values based on comparison to the previous model.
* $p <$ .001
[a] The χ2 value and degrees of freedom (*df*) are calculated differently for the WLSMV estimator and require a special difference test (see Muthén & Muthén, 2006).

In the next step, estimates of the factor loadings and thresholds were held invariant across groups. As shown in Table 4.7 (Model 2), when cross-group restrictions on the factor loading and threshold estimates were imposed, the RMSEA and TLI statistics did not differ appreciably from those observed for Model 1. Despite evidence of a statistically significant decrease in WLSMV difference χ^2 ($\Delta\chi^2$ [54] = 123.48, $p < .001$), which was likely a function of large sample sizes, a slight change of +0.16 in the WRMR value, and a moderate change in CFI values (ΔCFI = 0.023), inspection of the Modification Indexes (MI) did not reveal any items that would suggest these parameters are not invariant across groups.

Inspection of Table 4.7 (Model 3) indicates that the imposition of constraining the factor variance to be equal (in addition to the constraints imposed in Model 2) resulted only in trivial changes in the goodness-of-fit indexes (e.g., ΔCFI = 0.002, ΔTLI = -0.007, and ΔRMSEA = 0.006) compared to Model 2, with one exception. A moderate increase (0.378) in the WRMR

value was observed, indicating a slight decrement in fit between the two groups of interest when the factor variances were constrained to be equal. Also, a statistically significant WLSMV χ^2 difference test ($\Delta\chi^2$ [1] = 39.195, $p < .001$) suggested a difference between groups with respect to the parameter of interest. Again, however, because the $\Delta\chi^2$ fit index is sensitive to sample size, it is likely this value may have led to rejection of the model even though differences in the parameter of interest between groups were slight (Cheung & Rensvold, 2002; Kline, 2005). Taken together, these findings provide evidence of equivalence of factor variances across the two groups of interest for the Math section of the SAT Reasoning Test.

Invariance Levels of the Writing One-Factor Model Across Groups

Table 4.8 displays the fit indexes for the general factor multiple group model at three levels of invariance for the Writing section of the SAT Reasoning Test. Again, it is prudent to keep in mind the influence of sample size on fit statistics and to focus interpretations on the changes in goodness-of-fit indexes (e.g. CFI, TLI, RMSEA, WRMR) as constraints increase, rather than on the magnitude of the indexes (particularly the values of the χ^2 and $\Delta\chi^2$ fit indexes).

Assessment of the baseline model resulted in a good fit to the data. While the χ^2 was significant at 3,609.405 with 1,319 degrees of freedom, the TLI (0.974), RMSEA (0.026), and WRMR (1.969) all indicated good fit for Model 1. The CFI (0.941) indicated borderline fit. As a means of further evaluating the unconstrained model, the factor loadings for the Writing section were evaluated for each group separately (see Appendix E for standardized factor loadings). Examination of the single-factor model for the EWOD reveals that 37 of the 49 standardized pattern coefficients had values greater than 0.50, indicating that these items measure the unidimensional construct of writing reasonably well. In the EWD-T group, results were nearly identical; 37 (of 49) items had standardized pattern coefficient values greater than 0.50, again

suggesting that the majority of the items are good measures of the writing factor. Also of note, eleven of the items (Items ***3, 5, 8, 9, 10, 19, 21, 22, 27, 29, 32*** and 35) with factor loadings less than 0.50 in the EWD-T group were among the group of items with factor loadings less than 0.05 in the EWOD group (Items ***3, 5,*** 7, ***8, 9, 10, 19, 21, 22, 27, 29,*** and ***32***). Inspection of these parameter estimates showed that the two groups were virtually equivalent in terms of factor loading values.

Table 4.8: Goodness of Fit for Invariance Analyses Across Groups: Writing

Invariance Model	χ^{2a}	*df*[a]	CFI	TLI	RMSEA	WRMR	$\Delta\chi^2$	$\Delta\chi^2$ *df*
1. Factor loadings and thresholds freely estimated in both groups; residual variances fixed to one and factor means fixed to zero in both groups	3,609.405	1,319	0.941	0.974	0.026	1.969	-	-
2. Model 1 and all factor loadings and thresholds were constrained to be equal across groups	2,877.541	974	0.951	0.970	0.028	2.222	230.049**	55
3. Model 2 and factor variances invariant across groups	2,153.431	741	0.964	0.971	0.028	2.224	0.376	1

Note. All $\Delta\chi2$ values based on comparison to the previous model.
* $p <$ .001
[a] The $\chi2$ value and degrees of freedom (*df*) are calculated differently for the WLSMV estimator and require a special difference test (see Muthén & Muthén, 2006).

In Model 2, the estimates of the factor loadings and thresholds for the Writing section were held invariant across groups. As revealed in Table 4.8, when cross-group constraints on the factor loading and threshold estimates were imposed, there was no appreciable difference in the CFI, RMSEA, and TLI statistics from those observed for Model 1. However, evidence of a statistically significant $\Delta\chi^2$ test statistic ($\Delta\chi^2$ [55] = 230.049, p < .001), which may be the result of its sensitivity to large sample sizes, and a moderate change of +0.253 in the WRMR value called into question the tenability of strong metric invariance (e.g., the same pattern and equality

of factor loadings and the numerical equality of variable thresholds) across groups. When modification indexes for Model 2 were examined, 13 factor loadings (Writing Items 4, 7, 9, 16, 29, 30-35, 39, and 41) and four thresholds (Writing Items 6, 18, 26, and 44) were found to have large modification indexes (>10.00), suggesting that these parameters may not be invariant across groups.

Inspection of Table 4.8 (Model 3) indicated that constraining the factor variance to be equal (in addition to the constraints imposed in Model 2) resulted in virtually no change in goodness-of-fit indexes (e.g., CFI, TLI, and RMSEA) compared to Model 2. Furthermore, fixing the factor variance to be equal across groups resulted in no change in χ^2 ($\Delta\chi^2$ [1] = 0.376, $p > .05$). The final model for the Writing section, with values of all factor loadings and thresholds as well as the factor variances held invariant across groups, resulted in a reasonably good fit to the data (CFI = 0.964, TLI = 0.971, and RMSEA = 0.028). Such findings provided evidence of invariant factor variances across the two groups of interest.

Summary and Conclusions

In the present study, the researcher initially examined the factor structure of the Critical Reading, Math, and Writing sections of the SAT Reasoning Test across two groups (EWOD and EWD-T). A total of six competing models (two models per test section) concerning the constructs that are measured by the SAT Reasoning Test were evaluated separately for the two groups of interest. Findings indicated that the Critical Reading, Math, and Writing sections of the SAT Reasoning Test each measure a unidimensional construct for the two groups of interest. As a result, the one-factor model was considered adequate for the subsequent primary analyses.

In stage two of the study, three levels of model invariance for the final three models were tested across the two groups. The finding of invariance for the general factor model for the

Critical Reading, Math, and Writing sections across the EWOD and EWD-T groups provided a strong test of the hypothesis that the latent variables underlying the constructs of critical thinking (reading), reasoning (math), and writing are the same in these two populations. There was no evidence to reject the parsimonious hypothesis of strict equivalence of measurement of the latent variables for each test section in the two samples. Overall, the one-factor measurement model provided a good representation of both groups of examinees' responses to the ordered-categorical items comprising the Critical Reading, Math, and Writing sections of the SAT.

Results provided compelling evidence in support of the inference that the Critical Reading, Math, and Writing sections comprising the SAT reflect the same underlying constructs across both samples, suggesting that examinees in both groups respond to the items in a similar manner. If examinees in the EWD-T group were to respond differentially to items comprising each test section, then there would have been evidence of different item thresholds or factor loadings; there was only minimal evidence of such differential responding. In addition, constraining the factor variance to equality across the two groups indicated that this parameter is invariant across the two groups of interest.

In summary, despite notable differences in the mean scores across groups for each section of the SAT (see Table 3.1), and greater variability in the scores for examinees with learning disabilities or AD/HD than for the group of individuals without disabilities, results from all other statistical analyses suggested only negligible variability between groups. For instance, results of the internal consistency analyses of the Critical Reading, Math, and Writing sections of the SAT revealed no appreciable differences in reliability estimates across test sections and groups, and only slight SEM differences were found between the two groups of examinees (see Table 4.1).

Similarly, results of the preliminary single group analyses examining the factor structure of each test section were equivalent; findings indicated that the Critical Reading, Math, and Writing sections of the SAT Reasoning Test each measure a unidimensional construct for the two groups of interest. Finally, and most importantly, based on the results of the invariance analyses, the hypothesis that the items measuring the constructs of critical thinking, reasoning, and writing appear to function in the same way for the two groups of interest. Thus, there is no real evidence to suggest that the scores on the Critical Reading, Math, and Writing sections of the SAT Reasoning Test have different interpretations when examinees have an extended time administration as opposed to the standard time administration.

CHAPTER V

DISCUSSION

The main objective of this study was to examine whether the Scholastic Aptitude Reasoning Test (SAT®, 2005) measures the same construct across two groups of examinees. Specifically, the extent to which test scores of examinees without disabilities (EWOD) tested under standard (e.g., timed) conditions of the SAT are comparable to the scores of students with disabilities (EWD-T) tested with extended time was examined. The results from the preliminary single group confirmatory factor analyses confirmed that a general factor model for each of the three sections of the test fit the data from both groups of examinees. The overall results from the subsequent tests of measurement invariance provided support for the factorial invariance of the Critical Reading, Math, and Reading sections of the SAT Reasoning Test across the two groups of interest.

Accommodation policymaking and practice should be guided by empirical research and informed clinical judgment. Findings from the current study can assist in exploring the consequences of test use and provide information to test users about the validity of inferences that can be made from scores obtained from accommodated test administrations for students with disabilities. Following a presentation of the major contributions of the study from a measurement standpoint, the practical implications will be discussed from social, political, and legal perspectives.

Major Contributions

The assessment of measurement invariance verified the hypothesis that the Critical Reading, Math, and Writing sections of the SAT Reasoning Test are invariant across two groups

of examinees: students without disabilities who were administered a standard (e.g., timed) version of the test (EWOD) and examinees with learning disabilities (LD) or attention-deficit/hyperactivity disorder (AD/HD) who were administered the test with extended time (EWD-T). In the following section, the major contributions of the current study are further discussed. Initially, the factor structure of each of the three sections of the SAT, based on findings from the single group analyses, is described. Second, the value of conducting subsequent multi-group measurement invariance analyses is addressed. Finally, the importance of including a large sample consisting of a specific sub-group (e.g., examinees with LD or AD/HD) and isolating a single accommodation (e.g., extended time) is discussed.

The factor models used in the preliminary confirmatory factor analyses of the current study allowed tests of the hypothesis concerning the possible threat to validity from allowing examinees with disabilities extended time to take the SAT Reasoning Test. To assess this, the adequacy of fit of the two-factor model (Math) and three-factor model (Critical Reading and Writing) in contrast to a one-factor model had to be established for each of the two groups separately. According to the results of the preliminary confirmatory factor analyses, there was no evidence to reject the parsimonious hypothesis of unidimensionality for the Critical Reading, Math, and Writing sections of the SAT for each group.

This finding bears significance in that it provides evidence of configural invariance, upon which subsequent analyses were based. In order to compare parameter equivalence across groups, it is necessary to first establish that the basic factor structure is the same across groups in terms of the number of factors and the variables loading on each factor (i.e., configural invariance). If configural invariance is not supported, groups must be examined separately because what is being measured varies as a function of group membership. Differences in factor

structure represent differences in conceptualization and may represent a qualitative difference in the meaning of the underlying factor (Gregg, Bandalos, Coleman, Davis, Jiménez, Robinson & Blake, in press).

The findings from the preliminary confirmatory factor analyses also revealed important information about the dimensionality of the Critical Reading, Math, and Writing sections of the SAT that extends beyond the scope of this study. When the fit of the alternate two – (Critical Reading) and three- (Math and Writing) factor models in comparison to the general factor model was evaluated, the intercorrelations among the factors in the two-and three-factor models indicated a high degree of shared variance among the factors (Critical Reading: 96-98%; Math: 96-98%; Writing: 83-96%) for both groups. This suggests that there is not strong evidence to support using subscale scores. That is, because the two- and three dimensions of the mutli-factor models were highly correlated, it is not recommended that each be considered as a separate subscale. Subscore analyses can guide the diagnostic process only when the subscores have an interpretable dimensional structure.

A second major contribution of the current study emerged from the assessment of measurement invariance of the Critical Reading, Math, and Writing sections of the SAT across EWOD and EWD-T groups. For each step of the invariance analyses, in which increasing constraints were placed on the parameters, each model was evaluated for both groups simultaneously. This methodological approach is unprecedented in studies examining the factor structure of the SAT, and provides important information about the psychometric properties of the newly revised SAT above and beyond what has been previously reported.

For instance, previous studies involving the SAT have focused primarily on the predictive validity of the measure across groups (Cahalan, Mandinach, & Camara, 2002), the

interrelationship among the factors (Rock, Bennett, Kaplan, & Jirele, 1988), or the basic factor structure (e.g., configural invariance) of the SAT using exploratory (rather than confirmatory) approaches (Morgan & Huff, 2002). In the current study, however, multi-group assessment of measurement invariance allowed for comparison tests (e.g. chi-square difference test; goodness-of-fit statistics) of whether certain relationships are the same for the groups. That is, can the same factor structure and parameter values (e.g., factor loadings, thresholds, and factor variances) be used for the EWOD and EWD-T groups? If specific parameters were found to be non-invariant across groups, the type of methodology used would have allowed the researcher to determine the specific source of lack of invariance.

A third major contribution of the current study was the size of the sample (N = 2,476) that included a specific sub-group (e.g., examinees with LD or AD/HD) and isolated a single accommodation (e.g., extended time). A major limitation found among the majority of studies examining the influence of test accommodations for students with disabilities is inadequate sample size (e.g., N < 100). To circumvent methodological problems frequently associated with insufficient sample size (i.e., lack of generalizability of findings; limitation in the capability to detect a significant effect for the accommodation), many researchers have included groups comprised of examinees with multiple disabilities and/or several types of accommodations, both of which preclude the determination of the effects of each type of accommodation (separately) on test scores for examinees with specific disabilities. The present study, however, isolated a single, albeit varied, disability (learning) and accommodation (extended time) within a sufficiently large sample of examinees.

Practical Implications

Societal

A number of practical implications for the provision of extended time on high-stakes tests can be drawn from the current set of findings. Understanding the issues surrounding accommodations on high-stakes tests begins with recognition of the consequences for individuals with disabilities who are not provided equal opportunities to demonstrate knowledge. The ultimate purpose of tests, whether they are gatekeeper tests such as those for college admission, licensure, and certification, or assessments given as part of a promotional process in education or employment, may factor into the issue of how extensively accommodations should be used and how test scores should be subsequently interpreted (Pitoniak & Royer, 2001). Both legislation and educational initiatives have strengthened the need to afford individuals with disabilities the same testing opportunities, and thus access to the same life experiences in education and employment, as individuals without disabilities (Pitoniak & Royer). Lack of access to educational attainment has an unsettling effect on career development and adult income (Bowen & Bok, 1998; Vogel & Reder, 1999).

Another underlying societal concern is whether an accommodation such as extra time is truly fair— does it level the playing field or slant it for a select few who qualify as disabled? Unlike accommodations for physical or sensory disabilities (ramps for access, Braille tests, etc.), many argue that accommodations, such as extended time, for psychological processing disabilities (e.g., LD, AD/HD) could seemingly help any test taker. Fairness ultimately involves allowing any test taker the same accommodation (Ranseen & Parks, 2005). Thus, the principles of universal test design, which suggest building tests with greater content validity and more flexible administration conditions (e.g., ample time allowed for all examinees) should be

considered for future development of large-scale tests. At the same time, despite the intensity of some examinees' requests for extended time, some granted this accommodation may find it to make the exam too demanding and stressful.

Professional

The results from this study may be useful for professionals working in a variety of fields associated with large-scale assessments and students with disabilities. Testing agencies, college admissions officers, professional licensing boards, and policymakers can use these findings to make more informed decisions about the provision of extended time on standardized tests. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) emphasizes the importance of evaluating whether test accommodations alter the construct(s) measured by the test. According to the first standard in the chapter on testing individuals with disabilities (AERA, et al),

> In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement. (p. 106)

This standard provides justification for granting accommodations to obtain more valid measures of examinees' knowledge, but it also underscores the notion that if an accommodation alters the construct measured, scores from accommodated tests cannot have the same meaning as scores from standardized administrations. The questions addressed in the current study provide empirical evidence that the accommodation of extended time does not, in fact, appear to alter the construct(s) being measured by the SAT Reasoning Test for examinees with learning disabilities and/or AD/HD. College admissions officers should use these results as evidence that scores from

an extended time form of the SAT have the same meaning, and therefore can be interpreted in the same way, as scores from standardized administrations.

Requests for extended time are the most common but contentious accommodation requests made by students with learning disabilities and/or AD/HD taking licensing exams (Ranseen & Parks, 2005). Many licensing boards have grown wary of this steep increase in accommodation requests, prompting most to initiate detailed documentation review procedures (Ranseen, 2000). Many organizations deny some accommodation requests unless individuals can fully document diagnosis, provide evidence that impairments associated with the diagnosis are of a severity considered disabling, and offer a clear rationale for requested accommodations (Ranseen & Parks). Disagreements over the provision of test accommodations on licensing exams have been sufficiently contentious, and some individuals with learning disabilities and/or AD/HD who have been denied accommodations have filed suits under the Americans with Disabilities Act (ADA) against licensing boards (*Argen v. New York State Board of Law Examiners,* 1994; *Bartlett v. New York State Board of Law Examiners,* 1998; *D'Amico v. New York State Board of Law Examiners,* 1993; *Gonzales v. National Board of Medical Examiners,* 2000; *Pazer v. New York State Board of Law Examiners,* 1994; *Price, Singleton, & Morris v. National Board of Medical Examiners,* 1997). These cases form just some examples of the accumulating case law interpreting the complex legislation surrounding test accommodations and licensure exams.

It is hoped that findings from the present study will serve as an important basis upon which accommodation-related decisions can be made by licensing boards, and ultimately, prevent future cases from occurring. However, several important questions remain, including: Which accommodations are reasonable and for what disabilities? Does the accommodation of

extended time for examinees with LD and/or ADHD alleviate disabilities in a manner that meets the societal goal of being inclusive, or do they increase incompetent professional practice? Do accommodations lead to unfair advantage for those deemed disabled? None of these questions are particularly easy to answer. Thus, research examining the merits of different types of test accommodations for different disabling conditions should be conducted to help guide testing practices.

Legal/Political

The results from this study should be included in the collection of data used by those interpreting statutory language and writing regulations on testing accommodations. For instance, results are particularly relevant to the practice of "flagging" test scores (i.e., a notation on an examinee's score report to show that the test was administered in a nonstandard fashion, Mandinach, Cahalan, & Camara, 2002). Although many testing agencies (e.g., the College Board, Educational Testing Services [ETS], American College Testing [ACT]) no longer flag test scores when an examinee receives extended time due to a disability, the practice of flagging (for the accommodation of extended time only) continues to be used by other large testing agencies, such as the Law School Admission Council (LSAC) on the Law School Admission Test (LSAT) and the Medical College Admission Test (MCAT). When policymakers revise or update current legislation surrounding the issue of flagging, all current research on the topic should be taken into consideration. The findings from the current study are particularly important because they provide "… credible evidence of score comparability across regular and modified administrations" and therefore, "…no flag should be attached to a score" (*Standard* 10.11; AERA et al., 1999, p. 108).

In addition, such findings may shed light on the question of whether students with disabilities attending public school (K-12) should be included in large-scale assessments. And if so, whether test scores for students with disabilities (if the test was administered with accommodations) should be reported separately and/or included in building and district aggregate scores (Pomplun & Omar, 2000). While it would be inappropriate to generalize the findings of the present study to all students at the elementary and secondary levels, and/or to all types of large-scale tests, results should be included in the collection of data used by policymakers and educators at the local and state levels when making decisions about the meaning of test scores for students with disabilities. This study can also serve as a catalyst for changes in the way test scores from other college entrance exams (e.g., ACT, GRE) are interpreted and used.

Limitations and Future Directions

There are a number of study limitations that suggest the need for additional research. First, it was established in the present study that the provision of extended time to examinees with learning disabilities and/or AD/HD did not appear to alter the constructs being measured by the SAT Reasoning test. However, in other cases, a particular accommodation may alter the intended construct and/or provide an unfair advantage to students who receive the accommodation. Thus, accommodation decisions must take into account the construct measured by a test, the degree to which the accommodation is likely to alter the construct, and the specific needs of the examinee. Research to date has provided some information on what types of accommodations are likely to maintain fidelity to the construct and remove construct irrelevant variance (e.g. extended time). However, further research is needed to help guide college

admissions officers and testing/licensing agencies in the interpretation and use of test scores from accommodated tests.

The second limitation relates to the topic of partial measurement invariance. In the current study, chi-square difference tests were used to determine whether group differences existed within the set of parameters tested (e.g., factor loadings, thresholds, and factor variances); however, such tests do not indicate the specific parameters that resulted in a lack of invariance. Therefore, future research investigating partial invariance among the three sections of the SAT may be tenable (Byrne, Shavelson, & Muthén, 1989; Marsh & Hocevar, 1985). This would involve retaining indicators with non-invariant parameter values but allowing the values of these parameters to vary across groups. The identification of non-invariant parameter values can be identified by examining statistics such as the Modification Indexes (MIs) provided by the Mplus program.

A follow-up investigation of partial invariance may be particularly informative with regard to the Writing section, for which 13 factor loadings (Writing Items 4, 7, 9, 16, 29, 30-35, 39, and 41) and four thresholds (Writing Items 6, 18, 26, and 44) were found to have large MIs (>10.00), suggesting that these parameters may not be invariant across groups. Further examination of the factor structure of the Writing section may also be important considering the relatively weaker (in comparison to the Critical Reading and Math sections) reliability and standard error of measurement (SEM) estimates obtained for both groups (see Table 4.1). Differential item functioning (DIF) analyses, which can be conducted to evaluate test items designed to be used across groups, may also be useful in identifying potentially problematic Writing items (Gregg, Morgan, Hartwig, & Coleman, in press).

Finally, replication of the results using samples from other administrations is essential. In order to establish a strong program of construct validity for the SAT Reasoning Test, future studies of a similar nature are necessary. According to the 1999 *Standards* (AERA/APA/ NCME), validity is defined as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184). And according to Kline (2005), there is not a single, definitive test of construct validity, nor is it typically established in a single study. Therefore, results should be replicated on subsequent SAT Reasoning Test administrations to verify that the results are generalizable beyond the current data.

REFERENCES

American College Testing (2003, July). *ACT's decision to stop flagging ACT Assessment scores achieved with nonstandard time*. Retrieved November 29, 2005, from http://www.act.org/aap/disab/flag.html

American Council on Education (2000). Facts in brief: At-risk students succeed in high school and beyond. *Higher Education and National Affairs, 49*(14), 1-2.

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

Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 *et seq.*

Argen v. New York State Board of Law Examiners, 860 F. Supp. 84 (W.D.N.Y. 1994).

Bartlett v. New York State Board of Law Examiners, 156 F.3rd 321 (2d Cir. 1998).

Bennett, R. E., Rock, D. A., Kaplan, B. A., & Jirele, T. (1988). Psychometric characteristics. In W. Willingham, M. Ragosta, R. A. Bennett, H, Braun D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 83-97). Boston: Allyn & Bacon.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17(1)*, 10-22.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381-409.

Bowen, W. G. & Bok, D. (1998). The arrival of the Bowen-Bok study on racial preferences in college admissions. *Journal of Blacks in Higher Education, 20,* 120-122.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201–213.

Breland, H., Kubota, M., Nickerson, K., Trapani, C. & Walker, M. (2004). *New SAT®writing prompt study: Analyses of group impact reliability.* (College Board Research Report No. 2004-1). New York, NY: College Board.

Bridgeman, B. Curley, E. & Trapani, C. (Draft, 2001). *To what extent is the SAT I speeded?: Is the SAT I differentially speeded for ethnic/gender groups?* Princeton, NJ: Educational Testing Service.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT: I Reasoning Test*. (College Board Research Report No. 2000-1). New York, NY: College Board.

Bridgeman, B., Trapani, C., & Curley, E. (2003). *Impact of fewer questions per section on SAT I scores.* (College Board Research Report No. 2003-2). New York, NY: College Board.

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (Eds.) (1993). *GRE technical manual: Test development, score interpretation, and research for the graduate record examinations program*. Princeton, NJ: Educational Testing Service.

Byrne, B. M. (1998). *Structural equation modeling with Lisrel, Prelis and Simplis*. London: Lawrence Erlbaum Associates.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Cahalan, C., Mandinach, E., & Camara, W. (2002). *Predictive validity of SAT I: Reasoning Test for test takers with learning disabilities and extended time accommodations.* (College Board Research Report No. 2002-05). New York, NY: College Board.

Camara, W. J., Copeland, T., & Rothschild, B. (1998), *Effects of extended time on the SAT I: Reasoning test score growth for students with learning disabilities.* (College Board Report No. 98-7). New York: College Entrance Examination Board.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In S. Messick and H. Wainer (Eds.), *Principals of modern psychological measurement: A festschrigt for Frederic M. Lord*. Hillsdale, NJ: Erlbaum.

Cheung, G. & Rensvold, R. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Chiu, C. W. T., & Pearson, P. D. (1999). *Synthesizing the effects of test accommodations for special education and limited English proficient students*. Paper presented at the National Conference on Large Scale Assessment, June 13-16, Snowbird, UT.

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005). *2003 State policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Coffman, W. E. (1962). The Scholastic Aptitude Test 1926–1962. Paper presented to the Committee of Examiners on Aptitude Testing.

Cohen, A.S. & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal* of *Educational Measurement*, *42*(2), 133-148.

Cohen, A. S., Gregg, N, Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20* (4), 225-233.

College Board (2005a). Counselor's connection: About SAT. Retrieved November 21, 2005, from http://www.collegeboard.com/prof/counselors/tests/sat/about/ about_sat.html

College Board (2005b). Counselor's connection: Resources. Retrieved November 21, 2005, from http://www.collegeboard.com/prof/counselors/

College Board (2005c). Counselor's connection: Scores and reporting. Retrieved November 21, 2005, from http://www.collegeboard.com/prof/counselors/tests/ sat/scores/ scores_reporting.html

College Board (2005d). Higher ed: Recruitment and admission SAT program. Retrieved November 21, 2005, from http://www.collegeboard.com/highered/ra/sat/sat.html

College Board (2005e). Higher ed: Recruitment and admission test characteristics of the SAT. Retrieved December 30, 2005, from http://www.collegeboard.com/prod downloads/ about/ news_info/cbsenior/yr2005/10_test_characteristics_sat_0506.pdf

College Board (2005f). Our Organization. Retrieved December 30, 2005, from http://www. collegeboard.com/about/association/association.html

Commission on New Possibilities for the Admissions Testing Program (1990). *Beyond Prediction*. New York: College Entrance Examination Board.

Cook, L. L. (1994, April). *Recentering the SAT score scale: An overview and policy considerations*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Cronbach, L. J. (1992). Four *Psychological Bulletin* articles in perspective. *Psychological Bulletin, 112*, 389-392.

Cronbach, L. J. & Furby, L. (1970). How should we measure "change": Or should we? *Psychological Bulletin, 74*, 68-80.

D'Amico v. New York State Board of Law Examiners, 813 F. Supp. 217 (W.D.N.Y. 1993).

Davison, M. (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin, 97*(1), 94-105.

Disability Rights Advocates (2001). *Do no harm - High stakes testing and students with learning disabilities.* LD Access Foundation, Inc.: Oakland, CA

Donlon, T. F. (Ed.) (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Dorans, N. (2002). *The recentering of SAT® scales and its effects on score distributions and score interpretations*. (College Board Research Report No. 2002-11).). New York, NY: College Board.

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (RR-87-35*).* Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Lawrence, I. M. (1999). *The role of the unit of analysis in dimensionality assessment* (RR-99-14). Princeton, NJ: Educational Testing Service.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and proposed alternative. *Organizational Behavior and Human Decision Processes, 58*, 141-155.

Elliot, S. N., McKevitt, B.C., & Kettler, R.J. (2002). Testing accommodations research and decision making: The case of "Good" scores being highly valued but difficult to achieve for all students. *Measurement and Evaluation, 35*, 153-166.

Fuchs, L.S. & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator, 56*, 24-29.

GED Testing Service. (2002). *Who took the GED?* GED Statistical Report. Washington, DC: American Council on education.

Gonzales v. National Board of Medical Examiners, 225 F.3d 620 (6th Cir. 2000).

Gregg, N., Bandalos, D., Coleman, C., Davis, M., Jiménez, J., Robinson, K. & Blake, J. (in press). *Developmental Neuropsychology*.

Gregg, N., Morgan, D., Hartwig, J. & Coleman, C. (in press). Accommodations and the adult populations with learning disorders: State of the art research. In L. E. Wolf, H. Schreiber, & J. Wasserstein (Eds.), *Adult Learning Disorders: Contemporary Issues, Neuropsychology Handbook Series*. New York: Psychology Press.

Gregg, N., Mather, N., Shaywitz, S., & Sireci, S. (2002). *Flagging test scores of individuals with disabilities who are granted the accommodation of extended time.* An unpublished report of the majority opinion of the blue ribbon panel on flagging.

Haladyna, T. M. & Downing, S. M. (2004, Spring). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 17-27.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9* (2), 139-164.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research. 18,* 117–144.

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), 1-55.

Huff, K. L., & Sireci, S. G. (2001, October). *Appraising the dimensionality of a large-scale assessment across various demographic groups.* Paper presented at the Northeast Education Research Association, Ellenville, NY.

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

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16,* 215–224.

Kline, R. B. (2005). Principles and practice of structural equation modeling, second edition. New York: The Guilford Press.

Kobrin, J. L. & Schmidt, A. E., (2005). *The research behind the new SAT®*. (College Board Research Report RS-11). New York, NY: College Board.

Koenig, J. A. & Bachman, L. F. (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. National Research Council. Washington DC: The National Academics Press.

Lawrence, I., Rigol, G., VanEssen, T., & Jackson, C. (2003). *A historical perspective on the content of the SAT®*. (College Board Research Report No. 2003-3). New York, NY: College Board.

Liu, J., Feigenbaum, M., & Cook, L. (2004). *A simulation study to explore configuring the new SAT® critical reading section without analogy items*. (College Board Research Report No. 2004-2). New York, NY: College Board.

Liu, J., Feigenbaum, M., & Dorans, N. (2005). *Invariance of linkings of the revised 2005 SAT Reasoning Test™ to the SAT® I: Reasoning Test across gender groups.* (College Board Research Report No. 2005-6). New York, NY: College Board.

Lu, Y., & Sireci, S. G. (2003). *Validity issues in test speededness.* (Center for Educational Assessment Research Rep. No. 504). Amherst, MA: School of Education, University of Massachusetts.

Mandinach, E., Cahalan, C. & Camara, W. (2002). *The impact of flagging on the admission process: Policies, practices, and implications*. (College Board Research Report No. 2002-02). New York, NY: College Board.

Marsh, H. W. & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher-order factor models and their invariance across groups. *Psychological Bulletin*, 97, 562-582.

McDonald, R. P., & Ahlawat, K. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

Meade, A. W. & Lautenschlager, G. J. (2004). A Monte Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*(1), 60-72.

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

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Milewski, G. B., Johnsen, D., Glazer, N., & Kubota, M. (2005). A survey to evaluate the alignment of the new SAT® writing and critical reading sections to curricula and instructional practices. (College Board Research Report Summary No. 2005-1). New York, NY: College Board.

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

Millsap, R. E. & Kwok, O.M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Morgan, D. & Cahalan, C. (2003). Review of state policy for high stakes testing of students with disabilities on high school exit exams. Princeton, NJ: Educational Testing Service. Retrieved March 29, 2004 from the World Wide Web: http://www.ets.org/research.html

Morgan, D.L., & Huff, K. (2002). *Reliability and dimensionality of the SAT for examinees tested under standard timing conditions and examinees tested with extended time.* (Unpublished research report at the Educational Testing Service.) Princeton, NJ: Educational Testing Service.

Muthén, L. K. & Muthén, B. O. (1998-2006). *Mplus user's guide, fourth edition.* Los Angeles, CA: Muthén & Muthén.

Ofiesh, N., Mather, N., & Russell, A. (2005). Using speeded cognitive, reading, and academic measures to determine the need for extended test time among university students with learning disabilities. *Journal of Psychoeducational Assessment*.

Pazer v. New York State Board of Law Examiners, 849 F. Supp. 284 (S.D.N.Y. 1994).

Pitoniak, M., & Royer, J. (2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research. 71*(1), 53-104.

Pomplun, M. & Omar, M. H. (2000). Score comparability of a state mathematics assessment across students with and without reading accommodations. *Journal of Applied Psychology, 85*(1), 21-29.

Pomplun, M. & Omar, M. H. (2001). The factorial invariance of a test of reading comprehension across groups of limited English proficient students. *Applied Measurement in Education, 13*(3), 261-283.

Price, Singleton, & Morris v. National Board of Medical Examiners, 966 F. Supp. 419 (S.D.W.V. 1997).

Ranseen, J. D. (2000). Reviewing ADHD accommodation requests: An update. *Bar Examiner, 69,* 6–19.

Ranseen, J. D., & Parks, G. S. (2005). Test accommodations for postsecondary students: The quandary resulting from the ADA's disability definition. *Psychology, Public Policy, and Law*, *11*(1), 83-108.

Reilly, R., Bowden, S. C., Bardenhagen, F. J. & Cook, M. J. (in press). Equality of the psychological model underlying depressive symptoms in patients with temporal lobe epilepsy versus heterogeneous neurological disorders. *Journal of Clinical and Experimental Neuropsychology.*

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

Rock, D., Bennett, R.E., Kaplan, B.A., & Jirele, T. (1988). Construct validity. In Willingham, W.W., Ragosta, M., Bennett, R.E., Braun, H., Rock, D.A., & Powers, D.E. *Testing handicapped people* (pp. 99-107). Needham Heights, MA: Allyn and Bacon.

Scott, S., McGuire, J., & Shaw, S. (2003). Universal design for instruction. *Remedial and Special Education, 24*(6), 369-379.

Shaywitz, S. (2003). *Overcoming dyslexia.* New York: Alfred A. Knopf.

Sireci, S.(2005). Unlabeling the disabled: A psychomteric perspective on flagging scores from accommodated test administrations. *Educational Researcher, 24*(1), 3-12.

Sireci, S. & Gonzalez, E. J. (2003, April). *Evaluating the structural equivalence of tests used in international comparisons of educational achievement*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sireci, S., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature.* (Center for Educational Assessment Research Report No. 485). Amherst, MA: School of Education, University of Massachusetts.

Sireci. S., Zanetti, M., & Berger, J. (2003). Recent and anticipated changes in postsecondary admissions: A survey of New England colleges and universities. *Review of Higher Education, 26*, 323-342.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test validity* (147-169). Hillside, NJ: Erlbaum.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Available: http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm [January 2004]

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington: University of Kentucky, Mid-South Regional Resource Center Interdisciplinary Human Development Institute.

Vandenberg, R. J. (2002). Toward a fuller understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

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,* 4–69.

Vocational Rehabilitation Act of 1973, Section 504, Pub. L. No. 93-112, 29 U.S.C. § 794 (1977).

Vogel, S. & Reder, S. (Eds.). (1999). *Learning disabilities, literacy, and adult education*. Baltimore, MD: Paul H. Brookes Publishing Co.

Wagner, M., Newman, L., Cameto, R., Garza, N., & Levine, P. (2005). After high school: A first look at the post school experiences of youth with disabilities. *A Report from the National Longitudinal Transition Study-2.* Prepared for: Office of Special Education Programs, U.S. Department of Education (SRI Project P11182).

Wainer, H. & Keiley, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Ward, M. & Berry, H. (Summer 2005). Students with disabilities and postsecondary education: A tale of two data sets. *Heath Quarterly Newsletter.* The George Washington University HEATH Resource Center. Retrieved December 3, 2005 from http://www.heath.gwu.edu/newsletter/ Issue%2014/Issue%2014.htm

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wightman, L. (1993). *Test takers with disabilities: A summary of data from special administrations of the LSAT.* (LSAC Research Report No. 93-03). Newton, PA: Law School Admissions Council.

Willingham, W.W. (1976). *Validity and the Graduate Record Examinations program*. (Educational Testing Service Research Report No. GREB-76-01SR). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Needham Heights, MA: Allyn and Bacon.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Doctoral dissertation, University of California, Los Angeles.

Zumbo, B. D., Sireci, S. G. & Hambleton, R. K. (2003, April). *Revisiting exploratory methods for construct comparability and measurement invariance: Is there something to be gained from the ways of old?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Zurcher, R. & Bryant, D. P. (2001). The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities, 34,* 462-471.

Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13 (1),* 99-117.

APPENDIX A

NEW SAT® FOR THE PRESS: FACT SHEET

Introduction Schedule

Fall 2004	PSAT/NMSQT®*
March 2005	First administration of the new SAT

* Reflected changes to the SAT, excluding the student-written essay.

Changes to the SAT

Critical Reading

- The critical reading section, previously known as the verbal section, includes short and long reading passages.
- Analogies have been eliminated, but sentence-completion questions remain.

	Previous SAT	**New SAT**
Critical Reading		
Time	75 minutes Two 30-minute sections; one 15-minute section	70 minutes Two 25-minute sections; one 20-minute section.
Content	Sentence Completions, Passage-Based Reading, Analogies Measuring: Extended Reasoning, Literal Comprehension, Vocabulary in Context	Sentence Completions, Passage-Based Reading Measuring: Extended Reasoning, Literal Comprehension, Vocabulary in Context
Score	V 200-800	CR 200-800

Math

- The new math section includes topics from third-year college-preparatory math, such as exponential growth, absolute value, functional notation, and negative and fractional exponents.
- Quantitative comparisons have been eliminated.

	Previous SAT	**New SAT**
Math		
Time	75 minutes Two 30-minute sections; one 15-minute section.	70 minutes Two 25-minute sections; one 20-minute section.

Content	Multiple-Choice Items, Student-Produced Responses, Quantitative Comparisons Measuring: Number and Operations, Algebra I and Functions Geometry; Statistics, Probability, and Data Analysis.	Multiple-Choice Items, Student-Produced Responses Measuring: Number and Operations, Algebra I, II, and Functions Geometry; Statistics, Probability, and Data Analysis
Score	M 200-800	M 200-800

Writing

- A new writing section has been added to the test. Students are asked to write an essay that requires them to take a position on an issue and use reasoning and examples to support their position.
- The essay is similar to the type of writing required on in-class college essay exams.
- Multiple-choice questions measure a student's ability to identify sentence errors, improve sentences, and improve paragraphs.

	Previous SAT	**New SAT**
Writing		
Time	No test	60 minutes 35-minute multiple choice; 25-minute essay
Content	No test	Multiple-Choice: Identifying Errors, Improving Sentences and Paragraphs Student-Written Essay: Effectively Communicate a Viewpoint, Defining and Supporting a Position
Score		W 200-800 Multiple-Choice Subscore: 20-80 Essay Subscore: 2-12

Total Testing Time: 3 hours and 45 minutes, including an unscored 25-minute variable section (which helps in the development of future test questions).

APPENDIX B

CollegeBoard
connect to college success™

Instructions for Completing the 2005–2006 *Student Eligibility Form*

for Accommodations on College Board Tests Based on Disability
(SAT Reasoning Test™, SAT Subject Tests™, Advanced Placement Program® Exams, PSAT/NMSQT®)

Eligibility

A student with a disability, a condition that substantially limits his or her learning, may be eligible for accommodations on College Board tests. A *Student Eligibility Form* needs to be submitted for each student requesting accommodations. To be eligible, the student must:

1. have a disability that necessitates testing accommodations,
2. have documentation on file at school that supports the need for requested accommodations and meets the *Guidelines for Documentation,* and
3. receive and use the requested accommodations, due to the disability, for school-based tests.

If all of these requirements are not met, a student may still be eligible for accommodations on College Board tests. The student may send his or her disability documentation with the *Student Eligibility Form* to the address provided on page 8 and the College Board will review the documentation and make a determination (documentation must adhere to *Guidelines for Documentation* provided on this page).

NOTE:

- All students seeking accommodations on the basis of disability on the SAT Reasoning Test, SAT Subject Tests, Advanced Placement Program Exams, and PSAT/NMSQT must complete a *Student Eligibility Form*.
- Only one *Student Eligibility Form* needs to be completed for each student. It will cover all noted College Board testing programs for as long as the student's school verifies annually that the eligibility requirements and *Guidelines for Documentation* continue to be met.
- Step-by-step directions, definitions of terms, and reference information are found on pages 2-8 of these *Instructions for Completing the 2005-2006 Student Eligibility Form for Accommodations on College Board Tests Based on Disability (Instructions).*
- Test scores will not be provided if the accommodations are not approved by the College Board prior to the test administration.

For additional information, visit the College Board Web site at http://www.collegeboard.com/ssd/ or contact Services for Students with Disabilities (SSD) at 609 771-7137 (voice), 609 882-4118 (TTY) or ssd@info.collegeboard.org.

Guidelines for Documentation

The following *Guidelines for Documentation* list the information the College Board considers fundamental in determining that a student is eligible, based on disability, for accommodations on its tests, and what accommodations appropriately meet a student's individual needs for College Board tests.

When a student's school-generated IEP, 504 Plan, or other formal written educational plan/program aligns with the College Board's *Guidelines,* and officials at the student's school verify this to be accurate, the College Board accepts what the school verifies.

Some IEPs, 504 Plans, and other formal written educational plans/programs developed at schools to meet local needs, however, do not align with the following *Guidelines.* In those instances, a student may 1) elect to work with school officials to ensure that the disability documentation includes the information below before forwarding his or her *Student Eligibility Form* to the College Board; or 2) have the College Board review his or her disability documentation to determine the appropriate accommodations.

1. **State the specific disability,** as diagnosed;
2. **be current** (in most cases, the evaluation and testing should be completed within five years of the request for accommodations). For psychiatric disabilities, an annual evaluation update must be within 12 months of the request for accommodations;
3. **provide relevant educational, developmental, and medical history;**
4. **describe the comprehensive testing and techniques** used to arrive at the diagnosis (including evaluation date[s] and test results with subtest scores from measures of cognitive ability, academic achievement, and information processing). (For additional information such as a list of tests and their uses, please visit our Web site at http://www.collegeboard.com/ssd/prof/limitation.html/);
5. **describe the functional limitations** (how the disability impacts learning) See additional information at http://www.collegeboard.com/ssd/prof/limitation.html/;
6. **describe the specific accommodations** requested, including the amount of extended time required if applicable. State why the disability qualifies the student for such accommodations on standardized tests; *and*
7. **establish the professional credentials of the evaluator,** including information about license or certification and area of specialization.

APPENDIX C

STANDARDIZED FACTOR LOADINGS:

CRITICAL READING

	EWOD		*EWD-T*	
Items	***Loading***	***SE***	***Loading***	***SE***
Item 1	0.653	0.000	0.741	0.000
Item 2	0.740	0.089	0.835	0.082
Item 3	1.037	0.124	1.035	0.102
Item 4	0.533	0.069	0.501	0.054
Item 5	0.776	0.096	0.796	0.080
Item 6	0.458	0.069	0.464	0.058
Item 7	0.832	0.107	0.854	0.091
Item 8	0.295	0.057	0.361	0.053
Item 9	0.572	0.074	0.721	0.070
Item 10	0.640	0.081	0.754	0.075
Item 11	0.808	0.094	0.903	0.084
Item 12	0.864	0.104	0.927	0.089
Item 13	0.580	0.077	0.563	0.062
Item 14	0.555	0.071	0.577	0.060
Item 15	0.682	0.082	0.731	0.072
Item 16	0.652	0.086	0.805	0.077
Item 17	0.736	0.094	0.794	0.075
Item 18	0.691	0.088	0.88	0.083
Item 19	0.418	0.059	0.469	0.052
Item 20	0.577	0.083	0.525	0.066
Item 21	0.590	0.074	0.532	0.057
Item 22	0.698	0.084	0.784	0.075
Item 23	0.990	0.113	1.03	0.095
Item 24	0.556	0.073	0.45	0.052
Item 25	0.842	0.099	0.755	0.073
Item 26	0.670	0.087	0.61	0.063
Item 27	0.640	0.080	0.65	0.068
Item 28	0.559	0.070	0.566	0.057
Item 29	0.529	0.071	0.548	0.062
Item 30	0.517	0.077	0.458	0.060
Item 31	0.421	0.074	0.551	0.066
Item 32	0.692	0.082	0.732	0.070
Item 33	0.545	0.075	0.517	0.061
Item 34	0.539	0.084	0.579	0.073
Item 35	0.498	0.064	0.469	0.053
Item 36	0.556	0.073	0.644	0.065
Item 37	0.808	0.097	0.918	0.083
Item 38	0.428	0.061	0.553	0.057
Item 39	0.558	0.073	0.612	0.063
Item 40	0.649	0.079	0.746	0.072
Item 41	0.734	0.089	0.74	0.072
Item 42	0.489	0.067	0.493	0.054
Item 43	0.792	0.100	0.802	0.080
Item 44	0.749	0.091	0.762	0.073
Item 45	0.817	0.096	0.814	0.075
Item 46	1.012	0.117	1.009	0.094
Item 47	0.505	0.070	0.506	0.058
Item 48	0.669	0.084	0.607	0.062
Item 49	0.614	0.088	0.695	0.075
Item 50	0.571	0.075	0.587	0.094
Item 51	0.4	0.058	0.379	0.058
Item 52	0.673	0.084	0.655	0.069
Item 53	0.746	0.090	0.719	0.072
Item 54	0.625	0.083	0.648	0.073
Item 55	0.768	0.092	0.808	0.076
Item 56	0.58	0.080	0.641	0.067

	EWOD		***EWD-T***	
Items	***Loading***	***SE***	***Loading***	***SE***
Item 57	0.737	0.089	0.8	0.075
Item 58	0.284	0.051	0.32	0.045
Item 59	0.581	0.073	0.637	0.063
Item 60	0.530	0.072	0.528	0.059
Item 61	0.661	0.083	0.714	0.070
Item 62	0.441	0.071	0.507	0.062
Item 63	0.536	0.069	0.686	0.069
Item 64	0.535	0.068	0.505	0.056
Item 65	0.708	0.086	0.704	0.070
Item 66	0.923	0.109	0.911	0.085
Item 67	0.610	0.076	0.718	0.072

APPENDIX D

STANDARDIZED FACTOR LOADINGS:

MATH

	EWOD		***EWD-T***			***EWOD***		***EWD-T***	
Items	***Loading***	***SE***	***Loading***	***SE***	***Items***	***Loading***	***SE***	***Loading***	***SE***
Item 1	0.442	0.000	0.567	0.000	Item 28	1.233	0.355	1.261	0.269
Item 2	0.702	0.202	1.051	0.217	Item 29	0.652	0.192	0.820	0.169
Item 3	0.513	0.152	0.671	0.142	Item 30	0.669	0.187	0.776	0.158
Item 4	0.429	0.134	0.677	0.142	Item 31	0.796	0.223	0.969	0.197
Item 5	0.742	0.215	0.930	0.188	Item 32	0.838	0.232	1.088	0.220
Item 6	0.823	0.231	1.131	0.230	Item 33	1.019	0.285	1.299	0.270
Item 7	0.754	0.209	0.999	0.203	Item 34	1.055	0.292	1.139	0.239
Item 8	0.546	0.161	0.580	0.123	Item 35	0.822	0.231	0.847	0.176
Item 9	0.613	0.176	0.734	0.152	Item 36	1.127	0.323	1.334	0.282
Item 10	0.838	0.237	1.048	0.215	Item 37	1.007	0.289	1.026	0.219
Item 11	0.588	0.171	0.734	0.153	Item 38	1.039	0.379	1.246	0.310
Item 12	0.784	0.214	0.845	0.174	Item 39	0.397	0.144	0.585	0.136
Item 13	0.833	0.235	1.001	0.204	Item 40	0.695	0.211	0.928	0.194
Item 14	1.069	0.295	1.214	0.255	Item 41	0.478	0.150	0.628	0.136
Item 15	1.030	0.288	1.182	0.244	Item 42	0.662	0.190	0.708	0.148
Item 16	0.829	0.238	0.753	0.157	Item 43	0.617	0.182	0.724	0.151
Item 17	0.680	0.195	0.688	0.151	Item 44	1.060	0.293	1.209	0.247
Item 18	0.652	0.194	0.694	0.152	Item 45	0.711	0.200	0.805	0.166
Item 19	1.269	0.362	1.284	0.271	Item 46	0.584	0.166	0.715	0.148
Item 20	0.278	0.131	0.436	0.117	Item 47	0.497	0.146	0.483	0.106
Item 21	0.775	0.232	0.924	0.196	Item 48	0.574	0.165	0.616	0.132
Item 22	0.797	0.230	0.917	0.191	Item 49	0.486	0.145	0.553	0.122
Item 23	0.557	0.164	0.655	0.141	Item 50	0.804	0.232	0.859	0.181
Item 24	0.774	0.217	0.879	0.184	Item 51	0.831	0.240	0.795	0.177
Item 25	0.627	0.179	0.749	0.157	Item 52	0.509	0.156	0.524	0.118
Item 26	0.710	0.211	0.721	0.160	Item 53	0.490	0.158	0.495	0.123
Item 27	0.662	0.190	0.766	0.160	Item 54	0.550	0.169	0.579	0.134

APPENDIX E

STANDARDIZED FACTOR LOADINGS:

WRITING

	EWOD		***EWD-T***	
Items	***Loading***	***SE***	***Loading***	***SE***
Item 1	0.609	0.000	0.646	0.000
Item 2	0.514	0.077	0.528	0.069
Item 3	0.436	0.070	0.475	0.063
Item 4	0.758	0.098	0.912	0.101
Item 5	0.464	0.068	0.456	0.059
Item 6	0.644	0.082	0.663	0.079
Item 7	0.498	0.068	0.632	0.075
Item 8	0.391	0.061	0.445	0.057
Item 9	0.447	0.067	0.366	0.057
Item 10	0.494	0.078	0.479	0.072
Item 11	0.521	0.079	0.508	0.077
Item 12	0.627	0.097	0.618	0.080
Item 13	0.530	0.073	0.628	0.076
Item 14	0.719	0.102	0.761	0.088
Item 15	0.639	0.091	0.711	0.087
Item 16	0.507	0.072	0.759	0.090
Item 17	0.695	0.089	0.618	0.071
Item 18	0.764	0.096	0.851	0.096
Item 19	0.435	0.072	0.458	0.062
Item 20	0.570	0.091	0.588	0.089
Item 21	0.442	0.067	0.412	0.057
Item 22	0.463	0.068	0.473	0.063
Item 23	0.566	0.077	0.667	0.081
Item 24	0.647	0.091	0.664	0.084
Item 25	0.667	0.106	0.689	0.111
Item 26	0.719	0.115	0.714	0.099
Item 27	0.431	0.087	0.461	0.080
Item 28	1.048	0.130	1.104	0.123
Item 29	0.483	0.069	0.378	0.055
Item 30	0.926	0.110	0.774	0.088
Item 31	0.758	0.096	0.646	0.077
Item 32	0.454	0.065	0.324	0.051
Item 33	0.754	0.095	0.570	0.068
Item 34	0.957	0.116	0.652	0.074
Item 35	0.766	0.097	0.490	0.066
Item 36	0.584	0.097	0.618	0.090
Item 37	0.682	0.097	0.770	0.087
Item 38	0.706	0.102	0.809	0.094
Item 39	0.610	0.087	0.829	0.096
Item 40	0.545	0.075	0.606	0.072
Item 41	0.655	0.095	0.806	0.090
Item 42	0.725	0.096	0.715	0.081
Item 43	0.683	0.089	0.741	0.083
Item 44	0.507	0.072	0.563	0.070
Item 45	0.591	0.081	0.661	0.079
Item 46	0.718	0.094	0.716	0.085
Item 47	0.596	0.078	0.559	0.070
Item 48	0.581	0.081	0.631	0.080
Item 49	0.615	0.093	0.535	0.078