ED 469 245 TM 034 467

AUTHOR De Champlain, Andre F.

TITLE Assessing the Effect of Multidimensionality on LSAT Equating for Subgroups of Test Takers. Statistical Report. LSAC Research Report Series.

INSTITUTION Law School Admission Council, Newtown, PA.

REPORT NO LSAC-R-95-01

PUB DATE 1995-05-00

NOTE 32p.

PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS Blacks; *College Entrance Examinations; *Equated Scores; *Ethnic Groups; Higher Education; Hispanic Americans; Item Response Theory; Law Schools; Test Results; Whites

IDENTIFIERS *Dimensionality (Tests); *Law School Admission Test

ABSTRACT

This study investigated whether differences in the dimensional structure of the form of the Law School Admission Test (LSAT) across selected ethnic subgroups of test takers had any impact on equating results using an item response theory (IRT) true-score procedure. Whether there were any differences in the underlying latent trait composite across ethnic subpopulations that might yield noticeably different conversions for certain groups of test takers was studied. Data were from Caucasian (n=34,726), African American (n=3,548), and Hispanic (n=1,351) test takers. Results obtained with respect to the dimensionality of the LSAT with the three ethnic groups show that a two-dimensional model, specifying Analytical Reasoning and Logical Reasoning plus Reading Comprehension as the two factors, adequately accounted for the item responses of Caucasian and African American test takers, while a more complex model was required for the Hispanic subgroup. Equating results indicate that the differences between the conversion lines for the three ethnic groups and the total test-taker population were negligible. Results suggest that the African American and Hispanic conversion lines are statistically equivalent to the equating function for the majority Caucasian group and to one derived for the total test taker population. (Contains 5 tables, 15 figures, and 75 references.) (SLD)

TM

Assessing the Effect of Multidimensionality on LSAT Equating for Subgroups of Test Takers

Andre F. De Champlain

Law School Admission Council
Statistical Report 95-01
May 1995

LAW Services

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 193 law schools in the United States and Canada.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up to date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School nission Council.

Table of Contents

Tables

Figures

Executive Summary

Testing organizations typically disclose test forms after they have been administered to large test-taker populations. Therefore, several test forms must be developed annually to be as similar as possible to one another in terms of statistical and content attributes. Although a great deal of effort is placed on assembling comparable tests, forms will tend to vary somewhat in terms of their statistical characteristics. Hence, scores must be transformed in order to enable direct comparisons across forms. The process by which scores are adjusted so as to make them comparable to each other is referred to as equating. The Law School Admission Council (LSAC) employs item response theory (IRT) true-score equating to equate the LSAT.

One of the main assumptions underlying IRT true-score equating methods is unidimensionality of the construct underlying the items of the forms to be equated. This assumption of the model must be met in order to benefit from the many advantages of IRT-based procedures, namely population invariance. Simply stated, IRT equating functions should theoretically be independent of the groups from which they were derived, assuming the postulates of the model hold.

Several studies had previously attempted to assess how multidimensionality might affect the quality of IRT true-score equating results. Most studies examining this issue concluded that (unidimensional) IRT true-score equating procedures were quite robust to departures from the assumption of unidimensionality. The effects of multidimensionality on the quality of IRT true-score equating results were found to be negligible (Bogan & Yen, 1983; Camilli, Wang, & Fesq, 1995; Cook & Douglass, 1982; Cook, Dorans, Eignor, & Petersen, 1985; Dorans & Kingston, 1985; Kolen & Whitney, 1982; Modu, 1982; Snieckus & Camilli, 1993; Stocking & Eignor, 1986; Wang, 1985; Yen, 1984). However, these studies generally focused on dimensionality at the content level only, that is, they did not specifically examine the interaction of both content and test-taker population as suggested by Lord and Novick (1968) and Bejar (1983). Other studies, which centered on assessing how the interaction of both multidimensional test content and heterogeneous populations might impact on IRT true-score equating, did show that conversions could differ substantially across diverse groups of test takers, most notably when the content was also heterogeneous (Angoff & Cowell, 1985; Cook, Eignor, & Taft, 1988; Eignor & Cook, 1991). Although informative, none of these studies systematically investigated how differences in the latent trait composite of subgroups might affect IRT true-score equating results. The purpose of this study was therefore to assess the dimensionality of one form of the Law School Admission Test (LSAT) with respect to three ethnic groups of test takers and to investigate whether differences in their latent trait composite have any noticeable impact on IRT true-score equating results for these subgroups. More precisely, the equating functions estimated for African American and Hispanic test takers were compared to those derived for the majority Caucasian group as well as the total test-taker population to see if there existed any noteworthy differences.

Results obtained with respect to the dimensionality of the LSAT with the three ethnic groups showed that a two-dimensional model, specifying Analytical Reasoning and Logical Reasoning + Reading Comprehension as two factors, adequately accounted for the item responses of both Caucasian and African American test takers whereas a more complex model was required for the Hispanic subgroup.

Equating results indicated that the differences between the conversion lines obtained for the three ethnic groups and the total test-taker population were negligible. The largest residuals obtained when comparing the minority-group conversion lines to either the Caucasian or total population equating functions were well within one conditional standard error of measurement for score differences which again would indicate that the variations are of no practical significance.

Also, the effect of matching Caucasian test takers on the basis of the African American raw score frequency distribution did tend to increase the disparities between the equating functions at the extremes, hence contributing to a slightly larger mean absolute residual value. However, the discrepancies between the two conversion lines in the middle of the scale were smaller. These findings support those of Cook, Eignor, and Schmitt (1990) as well as Kolen (1990) who stated that matching generally did not contribute to a more accurate equating.

Regardless, the results obtained in this study suggest that African American and Hispanic conversion lines are statistically equivalent to the equating function of the majority Caucasian group as well as to the one derived from the total test-taker population. In other words, the current practice of applying a conversion function obtained from the total population to all test takers, irrespective of ethnicity, does not penalize minority test takers.

Introduction

As the Law School Admission Test (LSAT) is usually disclosed after each national administration, the Law School Admission Council (LSAC) prefers not to readminister the same test form repeatedly to different groups of test takers. Therefore, several test forms of the LSAT are constructed to be as similar as possible with regard to the content of the items, the skills being targeted by the test as well as the statistical characteristics of both the items and the test. This is accomplished by developing test items and forms that conform to a well-defined set of content as well as statistical specifications. However, in practice, LSAT test forms, like test forms developed for all large-scale admission or licensure testing programs, do vary slightly with respect to their statistical attributes. This makes it impossible to compare scores across different forms without adjusting for these differences. This adjustment is done through a process called score equating. Lord (1977) states that two test forms are equated when it is a matter of indifference to each test taker, or to anyone using the results, which test form he or she takes. More precisely, Lord (1980) states that scores on two tests can be equated if they show evidence of meeting the following four conditions:

- *Unidimensionality*, that is, the same underlying construct should be present in both test forms;
- *Equity*, that is, for each group of test takers of identical ability, the conditional frequency distribution of scores on one form (e.g., *Y*), after transformation, should be the same as the conditional frequency distribution of scores on the other test (e.g., *X*);
- *Population invariance*, that is, the equating function should be independent of the group from which it was derived;
- *Symmetry*, that is, the equating is transposable. The function that transforms scores from Form *X* to Form *Y* is the same as the one that maps scores from Form *Y* to Form *X*.

The sufficient, though perhaps not necessary, conditions of unidimensionality and population invariance set out by Lord (1980) are especially relevant for item response theory (IRT)-based equating procedures. One of the theoretical advantages of IRT, within an equating application, is that the function derived to transform scores from one test form to another should be independent of the population on which it was based (Cook & Petersen, 1987). However, this property does not hold if the assumptions of the models, one of which in most instances is unidimensionality of the latent ability space, are violated. Common IRT models (Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983) assume that item response probabilities are a function of a single latent trait. One such model is the three-parameter logistic function, given by,

$$P_i\,(u_i = 1 \mid \theta_j) = c_i + (1{-}c_i)\,\frac{e^{Da_i\,(\theta_j - b_i)}}{1 + e^{Da_i\,(\theta_j - b_i)}}. \tag{1}$$

This function specifies the probability that a randomly selected test taker of latent trait level θ_j will correctly answer a given dichotomous item *i*. The item difficulty parameter or b_i, corresponds to the latent trait value at the point of inflexion of the item characteristic curve (ICC) whereas the discrimination parameter or a_i, is the value of the slope of the ICC at its point of inflexion. The lower asymptote parameter or c_i, corresponds to the minimum $P(\theta_j)$ value sometimes called the *pseudo-guessing* parameter. The value *D* is a constant used to approximate the normal ogive model (≈1.7). Although most IRT models assume that item response probabilities can be estimated in a unidimensional space, this condition is rarely met in practice (Traub, 1983). A mathematics test, for example, might entail not only mathematical proficiency but also the capability to read and understand the problems being presented. Hence, the advantages of using IRT models to equate scores on two test forms, namely the population invariance property, might not generalize to conditions where the assumption of unidimensionality has been compromised. This led Petersen, Kolen, and Hoover (1989) to state that the unidimensionality and population invariance conditions set out by Lord (1980) are intertwined. Equating functions derived from tests that measure different latent traits will probably vary for different groups of test takers (Petersen, Kolen, & Hoover, 1989). Indeed, Lord and Novick (1968) as well as Bejar (1983) have stressed that dimensionality must be viewed as an interaction of a given sample of test takers with a specific set of items rather than solely as a characteristic of the content of the test.

Several researchers have attempted to assess the effect of multidimensionality on IRT true-score equating functions (Bogan & Yen, 1983; Camilli, Wang, & Fesq, 1995; Cook & Douglass, 1982; Cook, Dorans, Eignor, & Petersen, 1985; Dorans & Kingston, 1985; Kolen & Whitney, 1982; Modu, 1982; Snieckus & Camilli, 1993; Stocking & Eignor, 1986; Wang, 1985; Yen, 1984). A comprehensive review of these papers can be found in Harris (1993). The majority of these studies concluded that although multidimensionality of the latent ability space did affect the quality of IRT true-score equating, the impact often appeared to be minimal and of no practical significance. These findings confirm an earlier statement made by Divgi (1981) to the effect that IRT applications that deal with entire tests such as equating, are more likely to be robust to departures from model assumptions, e.g., unidimensionality of the latent trait space. Kolen and Whitney (1982) found that differences in the dimensionality of various tests of General Educational Development (GED) did not affect IRT true-score equating results in a noticeable fashion. Bogan and Yen (1983) and Yen (1984), in vertical equating studies, similarly concluded that unsystematic errors of equating are to be expected when equating two multidimensional tests. However, substantial systematic errors might be expected solely when attempting to equate two tests that measure very discrepant latent trait composites. Wang (1985) and Goldstein and Wood (1989) also stated that the impact of multidimensionality on the quality of IRT equatings is likely to be negligible as long as the same linear composite of latent traits underlies the item responses on both tests.

Dorans and Kingston (1985), compared conversion tables based on calibrations of homogeneous Graduate Record Examinations (GRE) verbal scale items versus those based on calibrations of heterogeneous items. They concluded that the differences in estimates were quite small, especially when the latent traits were moderately to highly correlated. This finding supported Divgi's (1981) view that (unidimensional) IRT equating methods appear to be sufficiently robust to departures from the assumption of unidimensionality. Cook, Dorans, Eignor, and Petersen (1985) noted that the violation of the assumption of unidimensionality observed with Scholastic Assessment Test (SAT) verbal and mathematics item responses did not seriously affect the quality of IRT true-score equating as measured by scale drift. These researchers concluded that IRT true-score equating results were acceptable with these data sets because of the presence of a dominant latent trait underlying the item responses to each test form, a point previously alluded to by Drasgow and Parsons (1983).

More recently, Camilli, Wang, and Fesq (1995) demonstrated that the effect of multidimensionality on the IRT true-score equating of some LSAT forms was negligible. The authors did suggest, however, that the impact of multidimensionality should be investigated for those subgroups of test takers whose latent trait composite differs from that of the total test-taker population. Snieckus and Camilli (1993), in a simulation study, showed that the effect of a two-dimensional test structure on IRT true-score equating was insignificant as measured by scale drift, except when the means on the secondary dimension were very discrepant across simulated groups of test takers. Even in that instance, the authors questioned whether the differences were of any practical significance.

Although these studies have provided useful information with respect to the quality of unidimensional IRT true-score equating functions in the presence of multidimensionality, they have generally tended to focus on the impact of multidimensional test content rather than investigating the interaction of both the characteristics of the items and the test-taker population responding to them, as had been stressed by Lord and Novick (1968) as well as Bejar (1983).

A number of studies have attempted to assess the impact of both heterogeneous subpopulations and multidimensional test content on the quality of IRT true-score and classical equating results (Angoff & Cowell, 1985; Cook, Eignor, & Taft, 1988; Eignor & Cook, 1991; Kingston, Leary, & Wightman, 1988; Stocking & Eignor, 1986). Angoff and Cowell (1985) noted that linear and equipercentile equating functions derived with the GRE quantitative scale tended to be relatively invariant across subgroups of test takers when the forms were homogeneous with respect to content. However, the conversions were quite discrepant with forms that showed more evidence of content heterogeneity.

Kingston, Leary, and Wightman (1988) assessed the degree of invariance of IRT true-score equating functions, derived from the Graduate Management Admissions Test (GMAT), across several subpopulations (e.g., males and females and various age groups). They concluded that equating functions did not vary substantially across the subpopulations that were examined in their study. The authors did stress, however, that the results were obtained with subgroups that were very similar with respect to latent trait distribution, and test forms that were quite homogeneous with regard to content. Hence, their findings should not be generalized beyond these conditions.

Cook, Eignor, and Taft (1988) reported that different equating functions were obtained using groups of students who took the same test at different administration dates. The researchers concluded that "curricular progress" probably accounted for these differences. The implications of these findings are important for organizations that routinely administer test forms throughout the year (Eignor & Cook, 1991). Skaggs and Lissitz (1986) felt that calibrations based on samples from different parts of the United States were probably not comparable, which could also potentially have an impact on IRT true-score equating. Stocking and Eignor (1986) suggested that a difference in mean latent trait value from pretest to operational form as well as differences in the dimensionality of both forms might account for some of the disparities found in the conversions derived at the various stages of a test. Skaggs and Lissitz (1988) noted that IRT true-score methods were relatively robust to differences in test taker latent trait levels and suggest that multidimensional test content probably accounts for the lack of invariance reported in previous studies.

Skaggs (1990) stated that the multidimensional nature of both the test and population examined is a complex issue that should be addressed more extensively. The somewhat misunderstood relationship between multidimensionality and equating in general prompted Braun and Holland (1982) to state that it might be useful, whenever referring to an equating function, to add a qualifying phrase describing the population for which the conversion table is likely to hold. Cook and Petersen (1987) have also suggested that it might be necessary, in certain instances, to provide a description of the group for which (equated) scores can be considered to have the same meaning. It would therefore seem imperative to investigate how possible differences in the dimensional structure of a test for various subgroups of test takers might affect score equating using an IRT true-score procedure. Goldstein and Wood (1989) emphasize the importance of conducting these types of studies when they stated:

> For various reasons to do with, say, curriculum or culture, equating relationships may vary over subpopulations, so that an overall relationship may not reflect at all accurately the relationships to be found within subgroups or subpopulations. This raises the potentially serious issue of bias and discrimination against certain subgroups. Unhappily, there appears to be little formal recognition of this in the equating literature and a lack of serious empirical study of the issue (p. 157).

Purpose

The purpose of this study was to investigate whether differences in the dimensional structure of a form of the LSAT across selected ethnic subgroups of test takers had any impact on equating results using an IRT true-score procedure. Specifically, are there any differences in the underlying latent trait composite across ethnic subpopulations that might yield noticeably different conversions for certain groups of test takers as compared to the majority group and total population equating functions?

Methods

The LSAT Equating Design

As was previously mentioned, equating enables the comparison of scores from different test forms. This procedure requires forms that are linked together through a common strand or *equating chain*. The general form of the equating chain used with the LSAT is presented in Table 1.

TABLE 1
Hypothetical LSAT equating chain

Operational forms	Pre-operational forms	Pretests
Base Form (1)	2	a, b, c, d
2	3	e, f, g, h
3	4	i, j, k, l
4	5	m, n, o, p
5	6	q, r, s, t

The equating design that is used with the LSAT is referred to as a *section pre-equating* design. As is shown in Table 1, any given LSAT form can be comprised of up to three sets of items: operational items, pre-operational items, and pretest items. The reported LSAT score that a test taker receives is based solely on the operational items. Hence, every test taker is exposed to every operational item of the LSAT. The pre-operational sections are administered to test takers in order to gather statistical information that will enable us to scale the form operationally in the future. These pre-operational forms are *spiralled*. That is, each section of pre-operational items is administered to a different group of test takers. Finally, the statistical characteristics of new items are assessed through the use of pretests. Pretest items have never been administered in a previous LSAT form. As was the case with pre-operational items, pretest sections are also spiralled to various groups of test takers. Either pre-operational or pretest items are included in a *variable* section that does not contribute to the test taker's final reported score. It is referred to as a variable section because different groups of test takers are exposed to different pre-operational or pretest sections. A summary of the LSAT equating design is presented in Table 2.

TABLE 2
LSAT section pre-equating design

	Operational Sections				Variable Sections				
Sample	AR	LA[1]	LB	RC	V_1	V_2	V_3	.	V_z
S_1	✓	✓	✓	✓	✓				
S_2	✓	✓	✓	✓		✓			
S_3	✓	✓	✓	✓			✓		
.	.	.	.	.				.	
S_z	✓	✓	✓	✓					✓

[1] LR = LA & LB

The three item types on the LSAT are Analytical Reasoning (AR), Logical Reasoning (LR), and Reading Comprehension (RC). Briefly, AR items measure a test taker's deductive reasoning. For example, some current AR items require a test taker to determine the proper ordering of people or objects. LR items measure a test taker's inductive reasoning. For example, some LR items require the test taker to identify flaws in a text. Finally, RC items measure the test taker's capability to read and interpret material. For some RC passages, test takers must identify the part of the stimulus that supports an inference. The LSAT form that was examined in this study contained 24 AR items, 51 LR items, and 27 RC items. Hence, the raw number-right score on this form could range from zero to 102.

Generally, when deriving a conversion table for a form of the LSAT, the first step consists of obtaining IRT parameter estimates using the marginal maximum likelihood estimation procedure implemented in the computer program BILOG (Mislevy & Bock, 1990).

Second, the item parameters obtained are scaled to the LSAT equating chain. The initial item parameters obtained for a given group of test takers cannot be compared to past populations because the metric defined by each calibration is distinct. The "arbitrariness" or indeterminancy associated with the estimation of item and ability parameters for the groups can be eliminated however by a linear transformation of item and ability parameters. The characteristic curve (CC) method developed by Stocking and Lord (1983) enables parameters on one form of a test to be scaled to another form, such that,

$$\theta^* = \lambda\theta + \kappa$$
$$b_i^* = \lambda b_i + \kappa$$
$$a_i^* = a_i / \lambda ,$$

where λ and κ are scaling parameters that are selected to minimize the difference between two test characteristic curves (TCCs) obtained from two different administrations of the same form. The process of obtaining scaled parameter values that can be directly compared across groups is referred to as *item pre-equating*.

Finally, once the item and latent trait parameters have been placed on the same scale through item pre-equating, the latent trait score (θ_j estimate) for a given individual will be the same (within measurement error) irrespective of the group from which it was estimated. Therefore, if latent trait values could be reported as final test-taker scores, the equating process would be completed. However, two more steps are undertaken to equate LSAT scores from a given form to the base form. First, the $P(\theta_j)$ values are summed across all 102 LSAT items for all test takers using the equation outlined in (1). This sum of $P(\theta_j)$ values is commonly referred to as a test taker's *expected true-score*. That is, the expected true-score (denoted by τ) for a test taker with latent trait value θ_j is given by,

$$\tau = \sum_{i=1}^{n} P_i(\theta_j). \tag{2}$$

However, we do not know what a test taker's true score is in reality. We only know the test taker's observed score (x). In order to overcome this situation, we treat the test taker's observed score (x) as his or her true score (τ). This approach is not without problems, however. For example, the lowest estimated true score that a test taker can achieve is equal to Σc_i (i.e., the sum of the lower asymptote item parameters) whereas the lowest observed score is in actuality, zero. Hence, the equating procedure does not function well for test takers with $x < \Sigma c_i$. Fortunately, this affects a very small proportion of test takers (in the order of 0.25% for this form of the LSAT) and, therefore, is not a major impediment to the equating process. Once this process has been completed, we can equate the scores obtained on a given form to those of a *base form*. A base form is a test which serves as a comparison point in order to assess how test takers would have performed had they been exposed to this form. Finally, equating allows us to place scores onto the LSAT score scale, which ranges from 120 to 180.

Test Takers

This study focused on one form of the LSAT that was administered to 45,918 test takers. Test takers who required an accommodated testing situation were excluded from the analyses. A breakdown of the test-taker population is given in Figure 1.

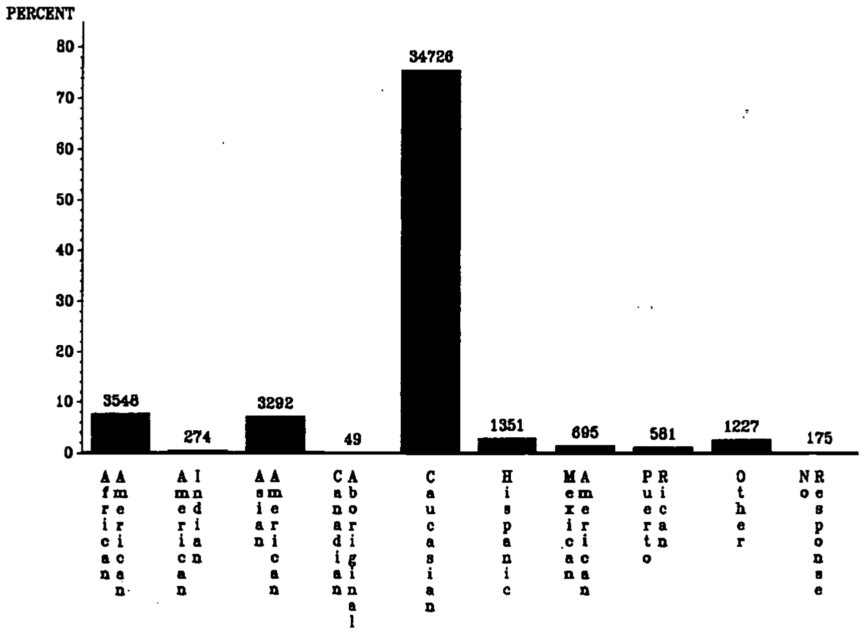


FIGURE 1. *LSAT ethnicity frequency distributions*

As can be seen, the majority of the population was composed of Caucasian test takers (75.6%). African American, Asian American and Hispanic test takers formed the three largest minority groups (respectively 7.7%, 7.2%, and 2.9%). It is important to underscore that these values were self-reported. Also, the Hispanic group of test takers is restricted to test takers who identified themselves as such, that is, it does not include test takers who described themselves as being Puerto Rican or Mexican American. The analyses in this study were focused upon the majority Caucasian group as well as two minority groups, i.e., African American and Hispanic test takers.

Dimensionality Assessment Procedures

The first set of analyses was centered on assessing the dimensionality of the LSAT form with the Caucasian, African American, and Hispanic subgroups. Dimensionality was assessed using two currently promising procedures: Stout's essential dimensionality procedure and *T* statistic (Stout, 1987, 1990), as well as an approximate χ^2 statistic based on McDonald's nonlinear factor analysis (NLFA) model (De Champlain, 1992; Gessaroli & De Champlain, 1994; McDonald, 1967, 1982).

Stout's Essential Dimensionality Procedure

Stout proposed a nonparametric procedure that is based on his notions of *essential independence* and *essential dimensionality* (Nandakumar, 1991, 1993; Stout, 1987, 1990). Essential dimensionality corresponds to the number of latent traits that are required to satisfy the assumption of essential independence, that is, a mean absolute residual covariance value that tends towards zero at fixed latent trait values,

$$\frac{1}{N(N-1)} \sum_{1\le i\le j\le N} |COV(U_i,U_j|\theta)| \approx 0. \qquad (3)$$

A test ($U_1, U_2, \ldots, U_N$) is said to be essentially unidimensional if for all subsets ($U_1, U_2, \ldots, U_M$) of length M (<N) and all values of Y_p,

$$\frac{1}{M(M-1)} \sum_{1\le i\le j\le N} |COV(U_i,U_j|Y_p)| \approx 0. \qquad (4)$$

where Y_p is the proportion correct score on the longer subtest and ($U_1, U_2, \ldots, U_M$) are shorter subtests with length n = N-M. Stout (1987, 1990) proposed the *T* statistic to test the assumption of essential unidimensionality. The steps involved in the calculation of the *T* statistic are outlined in Stout (1987) and Nandakumar (1991). In a series of studies carried out by Stout and his colleagues, the *T* statistic was found to be quite accurate in correctly determining essential unidimensionality and departure from the assumption with multidimensional data sets (Junker & Stout, 1994; Nandakumar, 1994) except when the test contained few items (less than 25) and the sample sizes were small (less than 750 test takers; De Champlain, 1992; Nandakumar, 1987).

Nonlinear Factor Analysis

Another approach that is gaining popularity for the assessment of dimensionality is the one that treats IRT as a special case of NLFA. Several researchers have shown that common IRT models are a special case of a more general NLFA model and that the functions are mathematically equivalent (Bartholomew, 1983; Goldstein & Wood, 1989; McDonald, 1967, 1994; Takane & De Leeuw, 1987). Based on this IRT-NLFA relationship, some researchers have suggested that the most suitable method of assessing dimensionality should be based on the analysis of the residual covariance matrix obtained after fitting an *m*-factor NLFA model (Gessaroli, 1994; Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1981, 1994). An approximate χ^2 statistic, based on McDonald's NLFA model, was investigated by De Champlain (1992) and Gessaroli and De Champlain (1994) as a potentially useful procedure for assessing dimensionality. The approximate χ^2 was originally proposed by Bartlett (1950) and outlined by Steiger (1980a, 1980b). The approximate χ^2 statistic tests the null hypothesis that the off-diagonal elements of a residual correlation matrix are equal to zero after fitting an *m*-factor NLFA model. The statistics are based on the estimation of parameters for an *m*-factor model using the NLFA approach outlined by McDonald (1967) and implemented by Fraser and McDonald (1988) in the computer program NOHARM. The approximate χ^2 statistic can be defined as,

$$\chi^2 = (N-3)\sum_{i=1}^{k}\sum_{j=1}^{i-1} z_{ij}^{2(r)},$$

where Zij2(r) is the square of the Fisher Z corresponding to the residual correlation between items *i* and *j*, (*i,j* = *1* . . . *k*) and *N* is the number of test takers in the sample. This statistic is distributed approximately as a central χ^2 with df = .5*k* (*k* - 1) - *t* where *k* is equal to the number of items and *t* is the total number of independent parameters estimated. The performance of the approximate χ^2 statistic was assessed with simulated unidimensional and multidimensional data sets that varied according to test length, sample size, item parameter values as well as the number of items defining each latent trait (De Champlain, 1992; Gessaroli & De Champlain, 1994). With unidimensional data sets, the empirical Type I error rates tended to be lower than the nominal α for the shorter test lengths examined but increased to values close to expected α probabilities for the longer tests. Also, the empirical Type I error rates obtained for the approximate χ^2 statistic were not affected by the nonnormality of the latent trait distributions (De Champlain & Tang, 1993). With multidimensional data sets, rejection rates based on the approximate χ^2 statistic were generally high, even in some instances with data sets containing as few as 15 items and 500 test takers, which was not the case with Stout's *T* statistic (De Champlain, 1992; Gessaroli & De Champlain, 1994).

Dimensionality Assessment Analyses

Initially, the fit of a unidimensional model was assessed by computing Stout's *T* statistic for the Caucasian, African American, and Hispanic LSAT data sets using the computer program DIMTEST (Stout, Junker, Nandakumar, Chang, & Steidinger, 1991). It was not possible to analyze all items of the LSAT form due to program restrictions. Hence, the last two (RC) items were dropped, (i.e., the first 100 items were subjected to DIMTEST). Also, the approximate χ^2 statistic proposed by Gessaroli and De Champlain (1994) was calculated for the same three data sets with the computer program CHIDIM (De Champlain & Tang, 1994), after fitting a one-factor model using NOHARM (Fraser & McDonald, 1988).

The fit of more complex models (e.g., two- and three-factor models) was also assessed with NOHARM and the approximate χ^2 statistic, as computed by CHIDIM. Past research has shown that there appear to be two distinct, though correlated latent traits underlying the item responses to recent forms of the LSAT (Ackerman, 1994; Camilli, Wang, & Fesq, 1995; Roussos & Stout, 1994), including the form that was the focus of this study (De Champlain, 1994a). More precisely, the first factor corresponds to the latent trait required to answer AR items whereas a second factor is needed to correctly answer LR and RC items. The appropriateness of this model was ascertained for the three ethnic groups by fitting a confirmatory two-factor model to the matrices of item responses and by calculating the approximate χ^2 statistic to see if the sum of the squared residuals differed significantly from zero. Finally, the fit of a three-factor model, specifying AR, LR, and RC as separate dimensions, was examined for these same three groups of test takers, again with the approximate χ^2 statistic.

Given that χ^2 -distributed statistics often suffer from an inflated Type I error rate with large sample sizes (Marsh, Balla, & McDonald, 1988), random samples of Caucasian and African American test takers were selected for all dimensionality analyses. Specifically, a random sample of 1351 Caucasian test takers and African American test takers was selected for these analyses in order to match the sample size of the Hispanic subgroup. Both procedures were shown to be generally accurate with respect to correctly identifying the unidimensional or multidimensional nature of item response matrices with these sample sizes (De Champlain, 1992; Gessaroli & De Champlain, 1994; Nandakumar, 1994).

Equating Analyses

The second set of analyses entailed deriving and comparing separate equating functions for the Caucasian, African American, and Hispanic subgroups to see if any noticeable discrepancies, possibly attributable to differences in the dimensional structure of the ethnic group item response matrices, might exist. Specifically, the equating functions for the three groups were compared to the total population conversion line, that is, the one actually used in the past for score reporting. In addition, the African American and Hispanic equating functions were plotted against the majority group (i.e., Caucasian) conversion line to see if any noticeable discrepancies occurred. In order to obtain these equating functions, three steps were followed.

First, separate IRT parameter estimates were obtained for each group of test takers using the computer program BILOG (Mislevy & Bock, 1990). Default BILOG program values were used for all analyses.

Second, the item parameters obtained were scaled to the LSAT equating chain using the CC method. For this study, the IRT parameters obtained for each ethnic group were scaled to the LR pre-operational form parameters (51 items) obtained from the total population using the SCALE program (McKinley, 1993a). Past research has shown that LR items on the LSAT exhibited the most evidence of unidimensionality across forms and random samples (De Champlain, 1994b; Roussos & Stout, 1994). Also, past studies have demonstrated that 40 common items generally result in an adequate scaling when using a characteristic curve procedure (Wingersky, Cook, & Eignor, 1986). The IRT parameters obtained for the total test-taker population were scaled to pre-operational parameter values. That is, all 102 items were included in the scaling analysis. It is also important to point out that the item parameters for the LSAT form that was the focus of this study were scaled to pre-operational parameters obtained from a similar administration in terms of overall latent trait distribution. Therefore, it is unlikely that any differences between the groups can be ascribed to the fact that the items were calibrated in two highly divergent administrations.

Once the item and latent trait parameters were placed on the same scale through item pre-equating, the $P(\theta_j)$ values were summed across all 102 LSAT items for all test takers in order to obtain their estimated true scores. As was mentioned previously, the true-score conversion table obtained for each group was then applied to their respective raw scores given that true-scores are unknown in reality (Lord, 1980). Once this process was completed, scores for all test takers were equated to those of a base form and placed on the LSAT score scale. This was accomplished by using the EQUATE program (McKinley, 1993b).

Results

Descriptive Statistics

LSAT raw number-right score descriptive statistics obtained with the three groups of test takers and the total population are presented in Table 3. Also, their respective distributions are plotted in Figure 2.

TABLE 3
LSAT raw number-right score descriptive statistics by ethnicity

Ethnic Group	Mean	Standard deviation	Skewness	Kurtosis
Total (N=45,918)	59.291	15.729	-0.087	-0.343
Caucasian (N=34,726)	61.384	14.817	-0.090	-0.228
African Americans (N=3,548)	45.198	13.994	0.441	0.020
Hispanics (N=1,351)	52.882	15.713	0.196	-0.489

The differences that were obtained between the groups with respect to both mean raw number-right score and frequency distribution are very similar to those reported with other national testing programs, most notably the GRE General test (Briel, O'Neill, & Scheuneman, 1993).

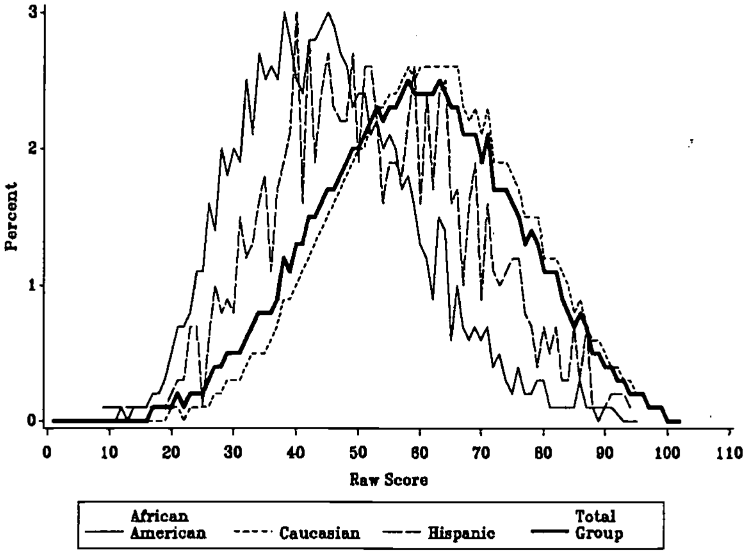


FIGURE 2. *LSAT raw score percentage distributions for African American, Caucasian, Hispanic, and total test-taker groups*

Dimensionality Analyses

Fit of a Unidimensional Model

Initially, the fit of a unidimensional model to the item response matrices of the Caucasian, African American, and Hispanic subgroups was ascertained using Stout's T statistic and the approximate χ^2 statistic obtained after fitting a one-factor NLFA model. These results are summarized in Table 4.

TABLE 4
Assessing the fit of a unidimensional model

Ethnic Group	Stout's T statistic	Approximate χ^2 and p values 1-F NOHARM model
Caucasian	2.90, p<.001	χ^2 = 9,239.50; p<.001
African Americans	2.71, p<.003	χ^2 = 8,288.29; p<.001
Hispanics	3.30, p<.001	χ^2 = 13,625.44; p<.001

Irrespective of the procedure employed, there is ample evidence to confirm the multidimensional nature of this form of the LSAT with all three subgroups of test takers. Stout's T statistic was statistically significant for all data sets. In addition, an inspection of the approximate χ^2 statistic values reveals that the sum of the squared residual correlations differed significantly from zero after fitting a one-factor NOHARM model, again clearly confirming the multidimensional nature of the data set for the three ethnic groups.

Fit of the Two- and Three-factor Models

The fit of a two- and three-factor model to the item response matrices of the Caucasian, African American, and Hispanic subgroups was also investigated using NLFA and the approximate χ^2 statistic. The approximate χ^2 statistic values are shown in Table 5.

TABLE 5
Assessing the fit of two- and three-factor models

Ethnic group	Approximate χ^2 and p values 2-factor NOHARM model	Approximate χ^2 and p values 3-factor NOHARM model
Caucasian	χ^2 = 4,890.76; p<.074	χ^2 = 4,655.36; p<.163
African Americans	χ^2 = 4,542.291; p<.985	χ^2 = 4,427.44; p<.990
Hispanics	χ^2 = 5,875.80; p<.001	χ^2 = 5,663.28; p<.001

Results indicate that the two-factor confirmatory NLFA model fit the item response matrices of both the Caucasian and African American test takers. In other words, the residual correlations did not differ significantly from zero after fitting the two-dimensional model. These results support those of Ackerman (1994), Camilli, Wang, and Fesq (1995), De Champlain (1994a) as well as Roussos and Stout (1994) who had suggested that two dominant latent traits were needed to correctly answer the items on the LSAT, that is, one for AR items and a second one for LR and RC items. However, this two-factor model did not adequately account for the item responses of the Hispanic subgroup of test takers, χ^2 (4946, N = 1351) = 5875.80, p<.001. In addition, the fit of a three-factor model, specifying AR, LR, and RC as separate dimensions, was still inadequate in accounting for the item responses of Hispanic test takers, χ^2 (4842, N = 1351) = 5663.28, p<.001.

Equating Analyses

Total Population and African American Equating Comparisons

The raw to unrounded scaled score conversion functions derived for the African American subgroup and the total test-taker population are plotted in Figure 3.

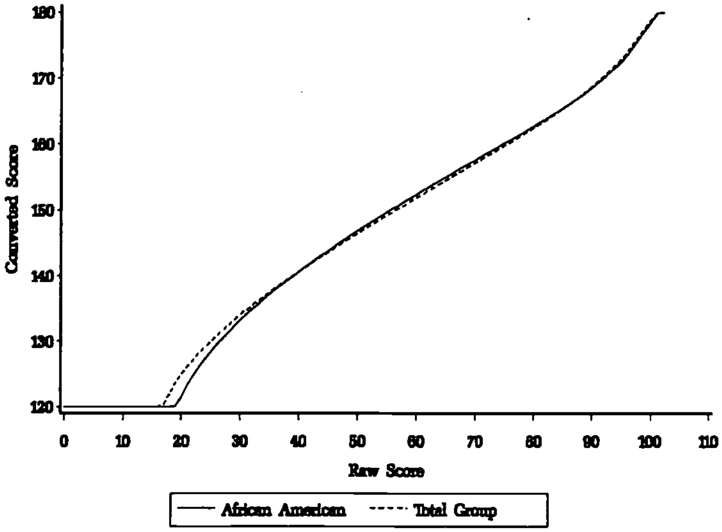


FIGURE 3. *LSAT total group and African American raw to scaled score conversion lines*

The differences obtained between the two conversions were very small. The mean absolute difference between the two score conversions, weighted by the frequency of African American test takers at each corresponding raw score point, was equal to 0.50 with a standard deviation of 3.08. A plot of the differences in the equated scores for the two groups across the raw score scale is provided in Figure 4.

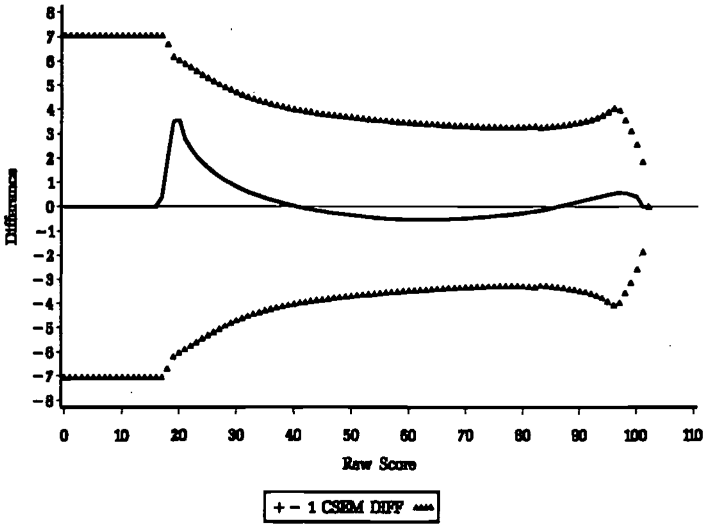


FIGURE 4. *Total group - African American equating residual plot*

Again, this plot clearly shows that the differences between the two conversions are of no practical significance and are well within one conditional standard error of measurement of score differences (CSEM DIFF; Dorans, 1984) across the entire raw score scale of the LSAT. In other words, if we were to compute the differences between scores for all pairs of total test-taker population and African American test takers of the same true latent trait value, 68% of these differences would fall within one pair of CSEM values at each score point. None of the differences plotted in Figure 4 were beyond these ranges. Not surprisingly, the largest discrepancies occurred at the lower end of the scale where the paucity of scores contributes to a poorer fit of the model and consequently a larger amount of measurement error.

Caucasian and African American Equating Comparisons

The raw to unrounded scaled score conversion functions derived for the African American and Caucasian subgroups are plotted in Figure 5.

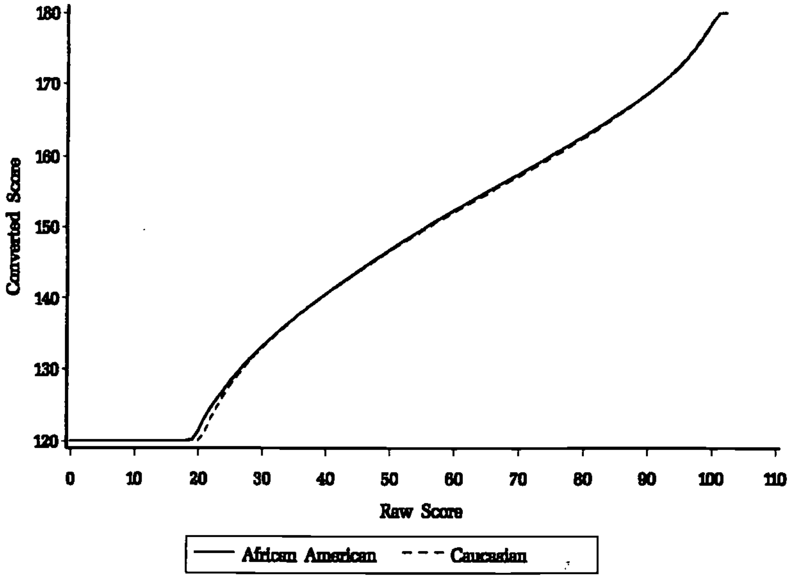


FIGURE 5. *LSAT African American and Caucasian raw to scaled score conversion lines*

The differences obtained between the two conversions were also very small. The mean absolute difference between the two conversions, weighted by the frequency of African American test takers at each corresponding raw-score point, was equal to 0.29 with a standard deviation of 1.47. A plot of the differences in the equated scores for the two groups across the raw score scale is provided in Figure 6.

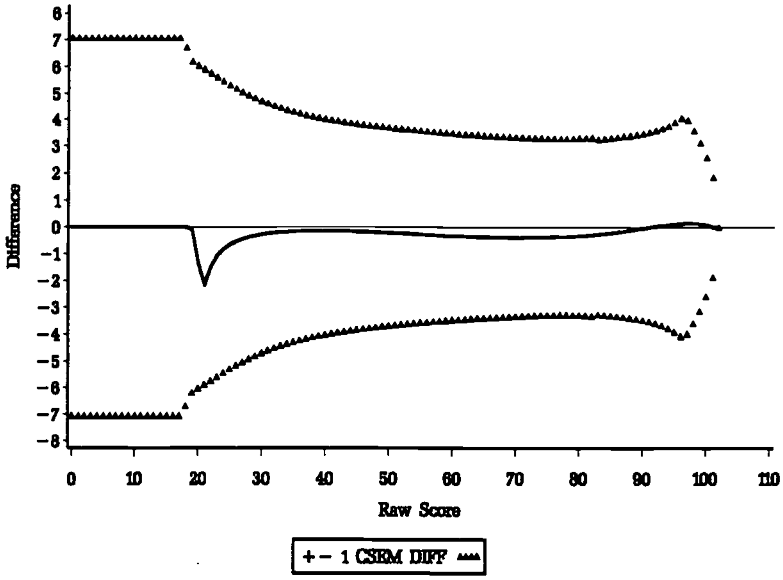
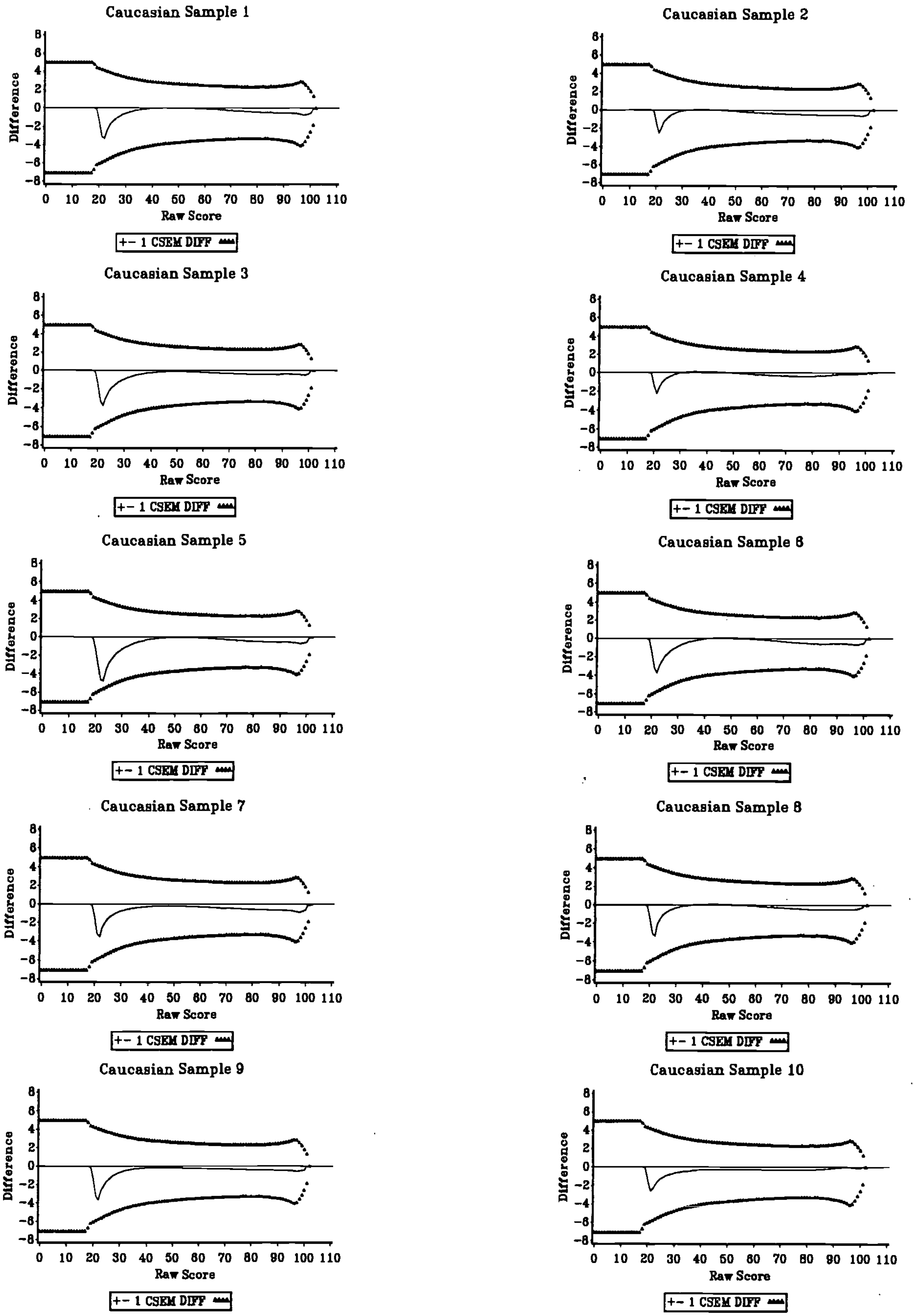


FIGURE 6. *Caucasian - African American equating residual plot*

Again, this plot clearly shows that the differences between the two conversions are of no practical significance and are well within one CSEM DIFF value (Dorans, 1984) across the entire raw score scale of the LSAT. None of the residuals plotted in Figure 6 were beyond their respective CSEM DIFF values. Once more, the largest discrepancies occurred at the lower end of the scale where the scarcity of scores yielded a poorer fit of the model and consequently a larger amount of measurement error. The true-score conversions obtained from 10 randomly selected samples of Caucasian test takers (N=3,000) were also plotted against the African American equating function in order to determine if there was any noticeable pattern in the residuals. These plots are shown in Figure 7.



Caucasian Sample 1

Caucasian Sample 2

Caucasian Sample 3

Caucasian Sample 4

Caucasian Sample 5

Caucasian Sample 6

Caucasian Sample 7

Caucasian Sample 8

Caucasian Sample 9

Caucasian Sample 10

FIGURE 7. *Random sample Caucasian - African American equating residual plots*

The results were quite similar to those obtained with the total Caucasian population in that the differences across most of the scaled score range were negligible for the 10 comparisons. Also, the largest residuals occurred at the lower end of the scale which again can be attributed to the larger amount of error concentrated in the segment of the scale that contains very few scores.

Given the differences in raw score distributions previously noted between the two groups, it was also of interest to examine whether matching Caucasian test takers on the basis of the African American raw score frequency distribution might lead to even smaller discrepancies between the two functions. However, it was not possible to obtain a perfect match at the lower end of the raw score scale. The raw score frequency distributions for the matched Caucasian subgroup and the African American population are shown in Figure 8.

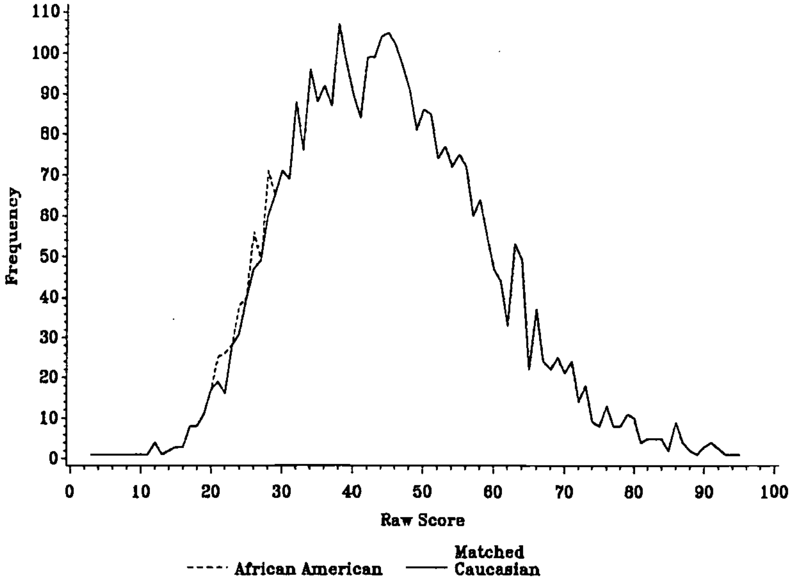


FIGURE 8. *LSAT raw number-right score frequency distributions for African American and matched Caucasian sample test takers*

As shown in Figure 8, with the exception of the very low end of the raw score scale, the two distributions were identical. A plot of both unrounded score conversions as well as the residuals between the two equating functions are presented in Figures 9 and 10.

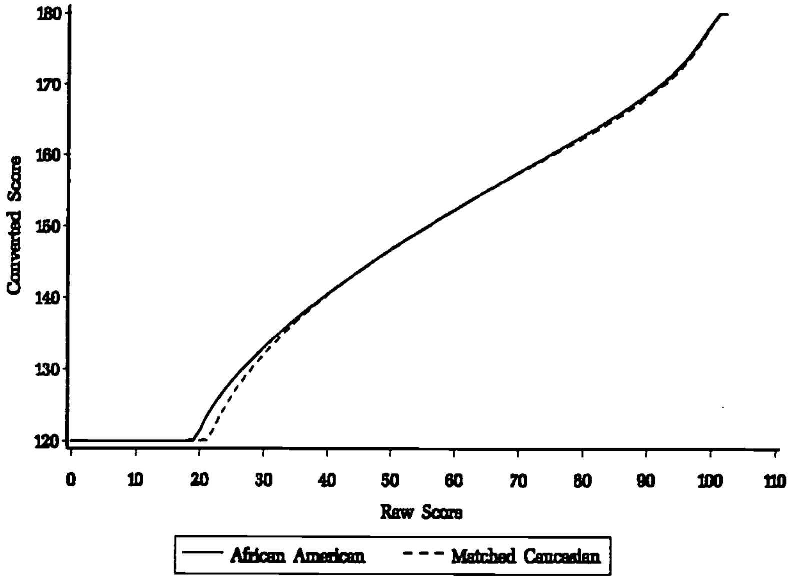


FIGURE 9. *LSAT African American and matched Caucasian sample raw to scaled score conversion lines*

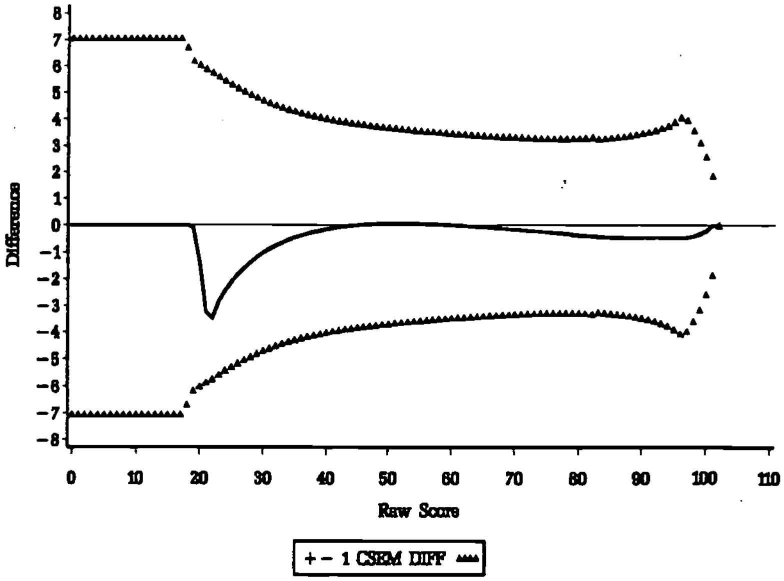


FIGURE 10. *Matched Caucasian sample - African American equating residual plot*

The differences obtained between the two conversions were slightly larger when matching on number-right score, but still inconsequential. The mean absolute difference between the two conversions, weighted by the frequency of African American test takers at each corresponding raw score point, was equal to 0.40 with a standard deviation of 3.88. Matching on the basis of raw score had a slight impact on the differences between the two conversion lines. Matching appears to have increased the differences between the two conversions at the very low end of the raw score scale. It's important to stress, however, that these differences were still within one CSEM DIFF across the entire raw score scale. Conversely, matching did reduce the discrepancies between both conversions in the middle-range of the distributions, where most scores are concentrated. However, the differences at the upper end of the raw score scale were slightly higher than previously observed with the (unmatched) Caucasian group. It is important to note that the highest raw score for the African American group (and consequently, the matched Caucasian group) was 95. Therefore, the differences noted at the extreme end of the scale can perhaps partially be imputed to the interpolation process employed by the EQUATE program in the absence of information (i.e., scores), rather than to any meaningful disparities.

Total Population and Hispanic Equating Comparisons

The raw to unrounded scaled score conversion functions derived for the Hispanic subgroup and the total test-taker population are plotted in Figure 11.

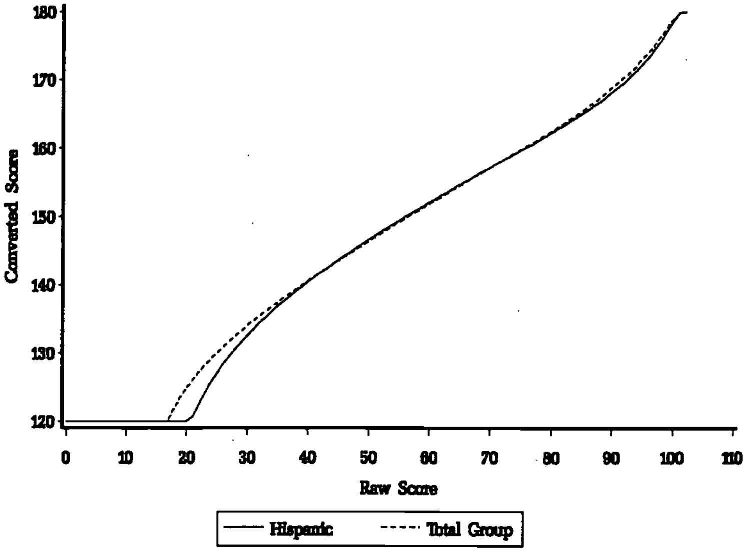


FIGURE 11. *LSAT Hispanic and total group raw to scaled score conversion lines*

The differences obtained between the two conversions were very small. The mean absolute difference between the two conversions, weighted by the number of Hispanic test takers at each corresponding raw score point, was equal to 0.41 with a standard deviation of 2.63. A plot of the differences in the equated scores for the two groups across the raw score scale is provided in Figure 12.

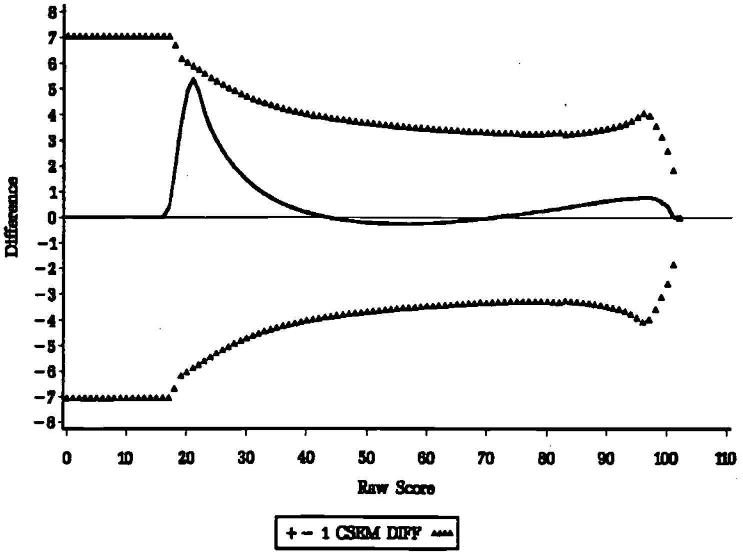


FIGURE 12. *Total Group - Hispanic equating residual plot*

Again, this plot clearly shows that the differences between the two conversions are of no practical significance and are within one CSEM DIFF value (Dorans, 1984) across the entire raw score scale of the LSAT. As was the case for previous comparisons, none of the differences plotted in Figure 12 were beyond these ranges. The largest discrepancies also occurred at the lower end of the scale due to the paucity of scores and consequently the larger amount of measurement error in that segment of the scale.

Caucasian and Hispanic Equating Comparisons

The raw to unrounded scaled score conversion functions derived for the Hispanic and Caucasian subgroups are plotted in Figure 13.

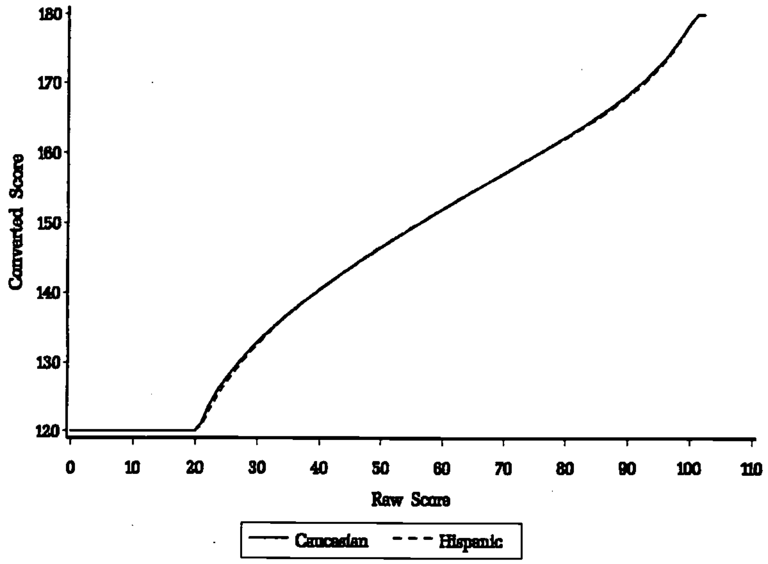


FIGURE 13. *Caucasian and Hispanic LSAT raw to scaled score conversion lines*

Once more, the differences obtained between the two conversions were virtually nil. The mean absolute difference, weighted by the frequency of Hispanic test takers at each raw-score point, was equal to 0.11 with a standard deviation of 0.55. A plot of the differences in the equated scores for the two groups across the raw score scale is provided in Figure 14.

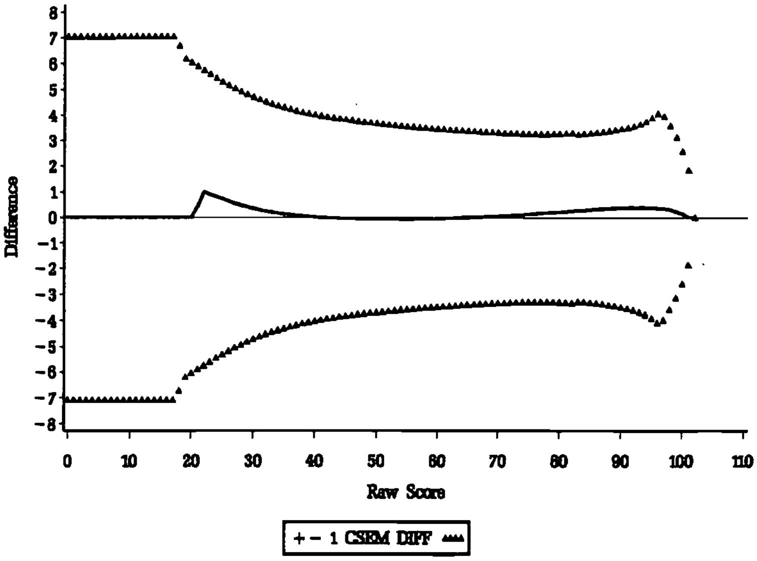


FIGURE 14. *Caucasian - Hispanic equating residual plot*

It is clear from this plot that the differences between the two conversions are of no practical significance and are again well within one CSEM DIFF value. As was the case with the previous comparisons, none of the differences plotted in Figure 14 were beyond this range. Also, the largest differences occurred at the two extremes of the raw score scale where few test-taker scores are located. Residual plots comparing the equating functions derived from the 10 random samples of Caucasian test takers (N=3,000) against the Hispanic population conversion are shown in Figure 15.

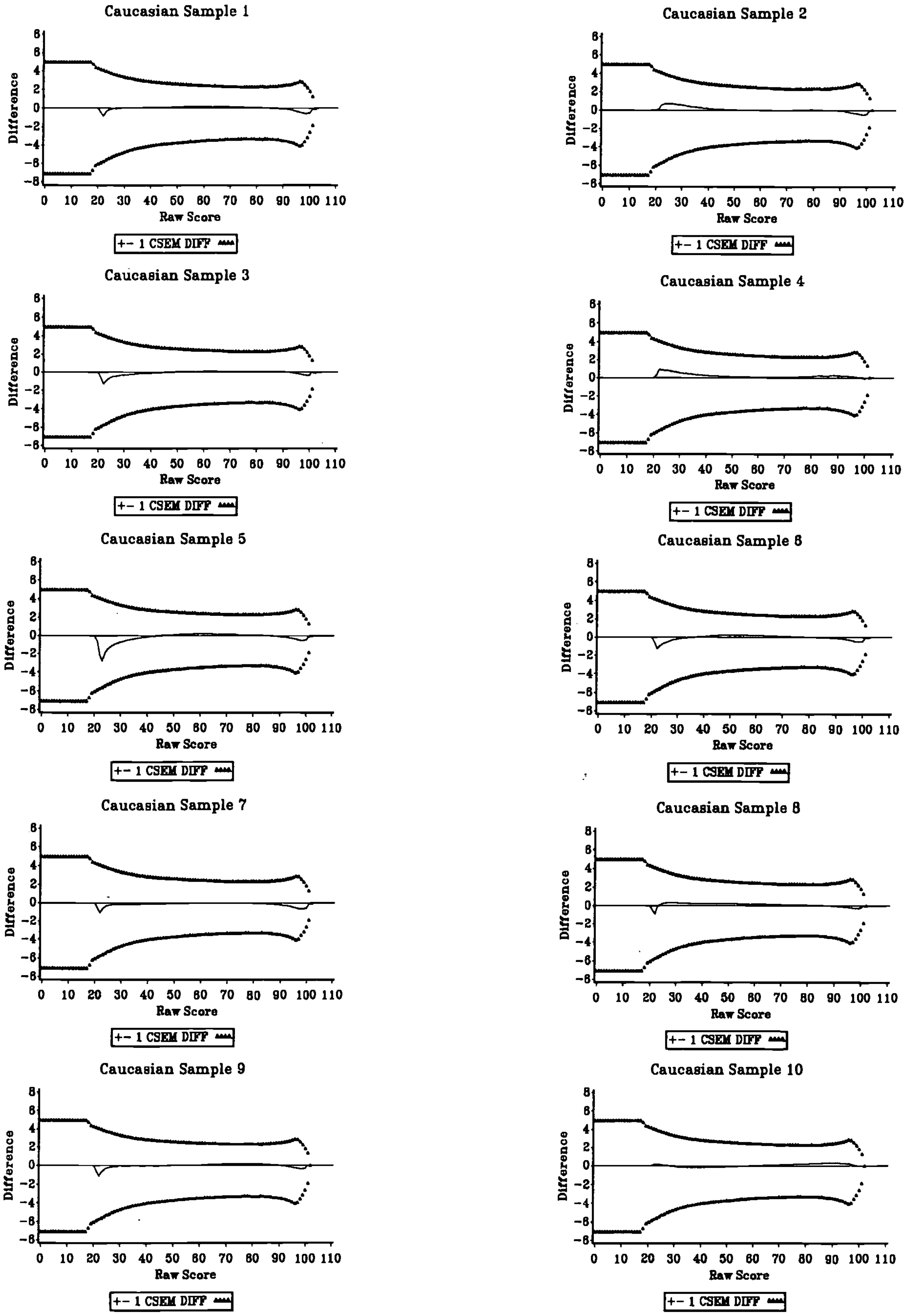


FIGURE 15. *Random sample Caucasian - Hispanic equating residual plots*

Again, the differences between the pairs of conversions were negligible and generally tended to be concentrated at the lower end of the raw score scale.

Discussion

Common IRT true-score equating methods assume that the construct underlying the items of the forms to be equated is unidimensional in nature. This assumption of the models must be met in order to benefit from the many advantages of IRT-based procedures, including population invariance. Simply stated, IRT equating functions should theoretically be independent of the groups from which they were derived, assuming the postulates of the models hold.

Previous studies that had assessed the relationship between multidimensionality and IRT true-score equating results generally concluded that this procedure was quite robust to departures from the assumption of unidimensionality (Bogan & Yen, 1983; Camilli, Wang, & Fesq, 1995; Cook & Douglass, 1982; Cook, Dorans, Eignor, & Petersen, 1985; Dorans & Kingston, 1985; Kolen & Whitney, 1982; Modu, 1982; Snieckus & Camilli, 1993; Stocking & Eignor, 1986; Wang, 1985; Yen, 1984). However, these studies focused on multidimensionality as a property solely of the content, rather than as an interaction of both test-taker population and the characteristics of a set of items. Studies that did examine the interaction of both multidimensional test content and heterogeneous population concluded that this combination of factors might impact on unidimensional IRT true-score equating results (Angoff & Cowell, 1985; Cook, Eignor, & Taft, 1988; Eignor & Cook, 1991; Kingston, Leary, & Wightman, 1988; Stocking & Eignor, 1986). The purpose of this study was therefore to assess the dimensionality of one form of the LSAT with three ethnic groups of test takers and to investigate whether differences in the latent trait composite have any noticeable impact on IRT true-score equating results for these subgroups. Specifically, the conversion lines estimated for African American and Hispanic test takers were compared to the equating functions derived from the majority Caucasian group as well as the total test-taker population to see if notable differences existed between the functions.

Results obtained with respect to the dimensionality of the LSAT with the three ethnic groups showed that a two-dimensional model, previously reported with the total population, adequately accounted for the item responses of both Caucasian and African American test takers. Specifically, it appears that one latent trait is required to answer AR items while a second one is needed to answer LR and RC items. However, the degree of misfit of this model was quite large for Hispanic test takers. In fact, a three-factor model, specifying AR, LR, and RC as separate dimensions, was still inadequate with regard to explaining Hispanic test-taker item responses. Note that the makeup of the Hispanic test-taker group might be quite varied and contain distinct subgroups with respect to their LSAT latent trait composite. For example, it is possible that the Hispanic population is comprised of one group of students who are quite fluent in English and a second group for whom English is a second language. This would have to be investigated more thoroughly before making any definite conclusions regarding the factor structure of the LSAT for these test takers.

Equating results indicated that the differences between the conversion lines obtained for the three ethnic groups and the total test-taker population were negligible, especially throughout the segment of the scale that contained most of the scores. In this sense, these findings confirm those reported in previous studies that examined the degree of invariance of IRT true-score equating functions across subgroups of test takers (Angoff & Cowell, 1985; Kingston, Leary, & Wightman, 1988).

However, it is important to note that the results obtained in this study also suggest that differences in the underlying latent trait composite between groups yielded conversions that did not differ substantially. The differences that were noted occurred primarily at the lower end of the raw score scale, which is to be expected given the small number of scores concentrated in that segment and hence, the larger amount of measurement error. Dorans and Kingston (1985) as well as Petersen, Cook, and Stocking (1983) similarly attributed larger residuals in the tails of the scale to the paucity of observations contributing to a poorer fit of the model for these scores. Also, it is known that the IRT true-score equating procedure does not function particularly well for test takers whose observed score is lower than their expected true score. Therefore, the discrepancies noted at the lower end of the scale are probably attributable to the shortcomings of the model rather than latent trait composite divergences between groups.

Also, the largest residuals obtained when comparing the minority group conversion lines to either the Caucasian or total population equating functions were well within one CSEM DIFF value, which again would indicate that the variations are of no practical significance. It is particularly surprising to see that the differences obtained between the Hispanic conversion line and the total population as well as the Caucasian equating functions were so insignificant given their distinctly different underlying latent trait composite. Perhaps, the two dimensions that account for the Caucasian and African American item response matrices also predominantly underlie Hispanic test taker item responses, in addition to other minor ones. Again, a more thorough investigation would seem necessary before drawing any definite conclusions as to why the equating functions were invariant across ethnic subgroups even in the presence of different latent trait composites.

In addition, it is important to point out that the operational item parameter estimates for all groups were scaled to pre-operational values from a similar administration in terms of latent trait distribution. It might be interesting in the future to examine whether scaling parameters to administrations that differ slightly with respect to latent trait distribution has any impact on the resulting conversion table.

Matching Caucasian test takers on the basis of the African American raw number-right score frequency distribution tended to increase the disparities between the equating functions at the extremes, hence contributing to a slightly larger mean absolute residual value. However, the discrepancies between the two conversion lines in the middle of the scale were smaller. These findings support those of Cook, Eignor, and Schmitt (1990) as well as Kolen (1990) who stated that matching generally did not contribute to a more accurate equating. Perhaps, as Skaggs (1990) pointed out, the complex multidimensional relationship that exists between test form and population might account for the questionable benefits that are sometimes reported in the literature with respect to the effect of matching on equating.

Regardless, the results obtained in this study suggest that African American and Hispanic conversion lines are statistically equivalent to the equating function of the majority Caucasian group as well as to the one derived for the total test-taker population. In other words, the current practice of applying a conversion function obtained from the total population to all test takers, irrespective of ethnicity, does not penalize minority test takers, as evidenced by the residual plots produced in this investigation. Also, the equating residual plots between minority- and majority-group test takers did not seem to conform to any apparent pattern, as displayed in the two sets of figures comparing the random Caucasian samples and minority group equating functions.

Investigating the relationship between multidimensionality, population invariance, and equating is especially relevant within a computer adaptive testing (CAT) framework. Currently, at each administration, all test takers are exposed to a single form of the LSAT. In that sense, the current paper-and-pencil LSAT administration represents a highly constrained environment. However, within a CAT perspective, a large number of test forms are tailored to groups of test takers in order to estimate their latent trait level with the highest possible degree of accuracy. The number of factors that must be taken into consideration with a CAT far outweighs the amount currently contended with within a paper-and-pencil framework. For example, the distributed nature of CAT forms, attributable to "on-demand" scheduling, potentially creates a more complex environment than its paper-and-pencil counterpart. Previous evidence has shown that equating functions could diverge noticeably across different administrations (Cook & Eignor, 1991). How might this affect the equating of CAT forms given the much larger number of administrations? Although the results obtained in this study are encouraging, the analyses should be replicated over several subsets of items in order to better understand how the equating of CAT forms might be affected by multidimensionality and heterogeneous subpopulations.

Hopefully, the results obtained in these and other studies will provide valuable guidelines to LSAC regarding the degree of robustness of IRT-based equating procedures to violations of unidimensionality and the degree of invariance to be expected with heterogeneous subgroups of test takers.

References

Ackerman, T. (1994, April). *Graphical representation of multidimensional IRT analysis.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Angoff, W. H., & Cowell, W.R. (1985). *An examination of the assumption that the equating of parallel forms is population independent* (Report No. 85-22). Princeton, NJ: Educational Testing Service.

Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics, 22,* 229-243.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology,* 3, 77-85.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Bock, D. R., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 4,* 443-459.

Bogan, E. D., & Yen, W.M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Monterey, CA: CTB/McGraw-Hill. (ERIC Document Reproduction Service No. ED229450).

Braun, H. I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp.9-48). New York: Academic Press, Inc.

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

Camilli, G., Wang, M.M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 1,* 79-96.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating.* (ETS Research Report NO. RR-85-30). Princeton, NJ: Educational Testing Service.

Cook, L. L., Douglass, J. B. (1982). *Analysis of fit and vertical equating with the three-parameter model.* Paper presented at the meeting of the American Educational Research Association.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., Eignor, D. R., & Schmitt, A.P. (1990). *Equating achievement tests using samples matched on ability* (College Board Research Report No. 90-2). Princeton, NJ: Educational Testing Service.

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

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 3,* 225-244.

De Champlain, A. (1992). *Assessing test dimensionality using two approximate chi square statistics.* Unpublished doctoral dissertation, University of Ottawa, Ottawa.

De Champlain, A. (1994a). [Assessing the dimensionality of two forms of the LSAT]. Unpublished raw data.

De Champlain, A. (1994b, February). *Assessing the dimensionality of the LSAT at the section level.* Paper presented at the University of Illinois, Department of Statistics, Champaign, IL.

De Champlain, A. F., & Tang, K. L. (1994). *CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's non-linear factor analytic model*. Manuscript submitted for publication.

De Champlain, A., & Tang, K. L. (1993, April). *The effect of nonnormal ability distributions on the assessment of dimensionality*. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

Divgi, D. R. (1981). *Potential pitfalls in applications of item response theory*. Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles.

Dorans, N. J. (1984). *Approximate IRT formula score and scaled score standard errors of measurement at different ability levels* (ETS Research Report no. SR-84-118). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 4*, 249-262.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Eignor, D. R., & Cook, L. L. (1991, April). *The effects of sample and test variation on achievement test equatings*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Gessaroli, M. E. (1994). The assessment of dimensionality via local and essential independence: A comparison in theory and practice. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 93-104). Ottawa, ON: Edumetrics Research Group, University of Ottawa.

Gessaroli, M. E., & De Champlain, A. (1994). *Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items*. Manuscript submitted for publication.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*, 234-246.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nyjhoff.

Harris, D. J. (1993, April). *Practical issues in equating*. Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

Hulin, C. L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Junker, B. W. & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds), *Modern theories in measurement: Problems and issues* (pp. 31-61). Ottawa, ON: Edumetrics Research Group, University of Ottawa.

Kingston, N., Leary, L., & Wightman, L. (1988). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (GMAC Occasional paper). Graduate Management Admission Council: Los Angeles.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*, 97-104.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement, 19,* 279-293.

Lord, F. M. (1977). Practical applications of characteristic curve theory. *Journal of Educational Measurement, 14,* 117-138.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometrika Monograph No. 15, 32*(4,Pt. 2).

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117.

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

McDonald, R. P. (1989). Future directions for item response theory. *International Journal of Educational Research, 13,* 205-220.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 63-86). Ottawa, ON: Edumetrics Research Group, University of Ottawa.

McKinley, R. (1993a). *SCALE*. Newtown, PA: Law School Admission Council.

McKinley, R. (1993b). *EQUATE*. Newtown, PA: Law School Admission Council.

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

Mislevy, R. J., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software, Inc.

Modu, C. C. (1982, April). *The robustness of latent trait models for achievement test score equating*. Paper presented at the meeting of the American Educational Research Association, New York.

Nandakumar, R. (1987). *Refinement of Stout's procedure for assessing latent trait dimensionality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

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

Nandakumar, R. & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18,* 41-68.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses - Comparison of different approaches. *Journal of Educational Measurement, 31,* 17-35.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics,* 2, 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), *Educational measurement* (pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

Roussos, L., & Stout, W. (1994, April). *Analysis and assessment of test structure from the multidimensional perspective.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Skaggs, G. (1990). To match or not match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3,* 105-113.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 4,* 495-529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of test taker ability on test equating invariance. *Applied Psychological Measurement, 1,* 69-82.

Snieckus, A. H., & Camilli, G. (1993, April). *Equated score scale stability in the presence of a two-dimensional test structure*. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

Steiger, J. H. (1980a). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Steiger, J. H. (1980b). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research,* 15, 335-352.

Stocking, M. L., Eignor, D. R. (1986). *The impact of different ability distributions on IRT preequating* (Research Report No. 86-49). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

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

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Stout, W. F., Junker, B. W., Nandakumar, R., Chang, H. H., & Steidinger, D. (1991). *DIMTEST and TESTSIM.* Urbana, IL: University of Illinois.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Wang, M. M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software, Inc.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986, April). *Specifying the characteristics of linking items used for item response theory item calibration.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement,* 2, 125-145.