CollegeBoard
connect to college success™

Research Report
No. 2005-6

Invariance of Linkings of the Revised 2005 SAT Reasoning Test™ to the SAT® I: Reasoning Test Across Gender Groups

Jinghua Liu, Miriam Feigenbaum, and Neil J. Dorans

www.collegeboard.com

College Board Research Report No. 2005-6
ETS RR-05-17

Invariance of Linkings of the Revised 2005 SAT Reasoning Test™ to the SAT® I: Reasoning Test Across Gender Groups

Jinghua Liu, Miriam Feigenbaum, and Neil J. Dorans

College Board, New York, 2005

Jinghua Liu is a lead measurement statistician at Educational Testing Service.

Miriam Feigenbaum is a principal statistical associate level II at Educational Testing Service.

Neil J. Dorans is a distinguished presidential appointee at Educational Testing Service.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,700 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three and a half million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #050481413) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Contents

Figures

Abstract

Score equity assessment was used to evaluate linkings of new SAT® to the current SAT Reasoning Test™[1]. Population invariance across gender groups was studied on the linkage of a new SAT critical reading prototype to a current SAT verbal section, and on the linkage of a new SAT math prototype to a current SAT math section. The results indicated that the conversion lines obtained through subgroup-only linkings were very similar to those obtained using the total group linking for both critical reading and math prototypes. Even though there were slight degrees of divergence from the total group conversion, the differences were smaller than the differences associated with rounding rules for the SAT. Hence, on the basis of the field trial data, it appears that population invariance was achieved with respect to gender groups.

Introduction

The goal of equating is to ensure that scores from one version of a test can be used interchangeably with scores from another version of the same test. When different versions of a test are constructed to the same explicit content, and statistical specifications are administered under the same conditions, the process through which minor differences in test form difficulty are adjusted and scores are placed on a common scale is called *equating* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

A testing program needs to make revisions from time to time in order to strengthen the alignment with school reform, curriculum changes, and test-taker changes. When a test undergoes changes in specifications and/or administrative conditions, it could introduce changes to equating practices. How do we judge whether or not the newer version of the test is equatable to the older version of the test? There are five equating requirements that are often regarded as basic to test equating (Dorans and Holland, 2000), among which population invariance is the most critical. That is, tests are equatable to the extent that the same equating function is obtained across significant subpopulations. Dorans (2004a) further modified this framework and proposed the concept of score equity assessment by using population invariance as a criterion to evaluate whether or not scores that are supposed to be used interchangeably are in fact interchangeable.

The framework of score equity assessment has important implications for the SAT Reasoning Test (referred to as the SAT in this report). The SAT is an objective, standardized test that measures verbal and mathematical reasoning abilities that students develop over time, both in and out of school. In order to strengthen the alignment of the test to current curriculum and institutional practices in high schools and colleges, changes are being made to the SAT.

The major content changes in the verbal section (SAT-V) include the elimination of the analogy items and the introduction of a new item type—paragraph reading items. The total length of the section will be reduced to 67 items from the current 78 items, a 14 percent reduction. The new section represents increasingly heavier reliance on a reading construct. The prototype for the new section consists of approximately 72 percent reading comprehension items, as compared to 51 percent in the current verbal section. The name of the section will be changed from verbal to critical reading to emphasize the change in focus.

The major content changes for the math section (SAT-M) are the elimination of the quantitative comparison items, and the expansion to include more advanced content, such as Algebra II. The prototype of the new version contains 54 items in total, as compared to the current 60 items. The number of student-produced response (SPR) items remains at 10, but the number of 5-choice items increases to 44 from the current 35. Correspondingly, the proportion of 5-choice items in the test increases to 81 percent from the current 58 percent.

Another significant change is the addition to the test battery of a new writing section, containing multiple-choice questions and a student-written essay. In addition, testing timing has been changed. The critical reading and math prototypes contain three sections each (two 25-minute sections and one 20-minute section) versus the current three sections (two 30-minute sections and one 15-minute section). The variable section is changed to 25 minutes from the current 30 minutes. The writing section consists of two multiple-choice sections (one 25-minute and one 10-minute) and one 25-minute essay section. Therefore, the total testing time increases to 3 hours and 45 minutes from the current 3 hours.

The purpose of this study is to evaluate score equity of the new SAT to the current SAT from the perspective of population invariance; that is, to examine whether the term *equating* could be defended for this application under this specific criterion. The proposed changes in the content, structure, context, and administration conditions will have implications for test equating if the current College Board 200-to-800 scale is maintained with the new critical reading and math measures. On the one hand, the prototypes of the new SAT are not developed according to the same content specifications, nor are they administered under the same conditions as the current test. From a psychometric perspective, the new blueprint for the development of the test may alter the meaning of the

[1] This research study was conducted before the introduction of the new SAT in March 2005. Therefore, references to the "current SAT Reasoning Test" apply to the SAT I: Reasoning Test, which was administered prior to March 2005.

scores in a nontrivial way. On the other hand, though the new test and the current test differ in content specifications and administration conditions, the study explores whether or not new versions of the SAT can be mapped onto the current SAT scales so that the characteristics of the new SAT scores can match those of the current test scores very closely. This is done by examining the invariance of linkings across subpopulations. Above all, this report tries to answer the following question: Is the new SAT equatable to the current SAT, even though the new SAT is different from the current SAT?

Section I of this report reviews previous research on score equity assessment. Section II describes the equating designs and methods for the current SAT and for the new SAT prototypes administered in the spring 2003 field trial. Section III presents and compares the results of score equity assessment between the current SAT and new SAT. Finally, Sections IV and V summarize the research implications and limitations.

I. Score Equity Assessment

Score equity assessment (Dorans, 2004a) focuses on whether or not scores that are supposed to be used interchangeably are in fact interchangeable. The key questions are: Does the test measure what it measures in the same way, across different subpopulations, as it does in the full population? Does the relationship between the two tests depend on whether examinees are male or female, or are white or African American or Hispanic or Asian? Score equity assessment uses population invariance/variance of linking functions across important subgroups, such as gender groups or ethnic groups, to assess the degree of interchangeability.

Subpopulation Invariance of Equating Function

Dorans and Holland (2000) summarized five requirements that are often regarded as basic to test score equating: the same construct requirement, the equal reliability requirement, the symmetry requirement, the equity requirement, and the subpopulation invariance requirement. Of the five requirements, subpopulation invariance is the most critical for score equatability. That is, the score equating function should be the same across subpopulations as it is in the total population.

One direct way to assess equatability is to check whether the linking functions between a pair of tests are the same across important subpopulations (e.g., gender and/or ethnicity). When the linking functions used to link pairs of score distributions are not invariant across different subpopulations, equating has not been attained. Note that no equating function can be perfectly subpopulation invariant. Rather, when the dependence of the linking functions on the subpopulations is small enough to be ignored, results of score linkings can be treated as if they were equatings.

An early work on score equity assessment is the study conducted by Dorans and Feigenbaum in 1994, when salient changes occurred to the older SAT version. That study examined the equating variability across gender groups and ethnicity groups by using two stringent indices: the percentage of raw scores for which the total and subgroup conversions differed by more than five points and the percentage of examinees for whom these conversions created scores that differed by more than five points. This study provided an initial framework in studying population invariance.

Dorans and Holland (2000) further developed the concept of population invariance and introduced general measures, namely, the standardized Root Mean Square Difference (RMSD) and the standardized Root Expected Mean Square Difference (REMSD), to quantify the degree to which linkings are subpopulation invariant. They examined several examples. In some, the expectation of equitability was very high, such as SAT-M to SAT-M and SAT-V to SAT-V, since alternative forms of the test are designed to be parallel in both content and statistical specifications, and equated scores are routinely reported on the 200-to-800 scales as if they were exchangeable. In other cases, such as the link from SAT-V to SAT-M, the expectation of equatability was not strong. The RMSD and REMSD suggested by Dorans and Holland appear to be appropriate measures of the degree of invariance.

More research has been conducted after Dorans and Holland's milestone work. While the initial research by Dorans and Holland was restricted by data collection design (random groups and single-group designs) and linking method (parallel-linear), von Davier, Holland, and Thayer (2003) extended the measures to the nonequivalent group anchor test design and examined its application to nonlinear linking methods. Yang, Dorans, and Tateneni (2003) investigated whether the multiple-choice to composite linking functions that determine Advanced Placement Program® (AP®) grades remain invariant over subgroups defined by region.

The concept of score equity assessment was introduced later, in a study to examine differential mean score differences on the free-response and multiple-choice sections for two AP Exams across gender groups (Dorans, 2004a). Do these differential mean score differences affect the equatability of AP scores and the invariance of AP grade assignments across males and females? Dorans used the population sensitivity of linking functions to assess score equity and placed score equity assessment within a fairness

framework that encompasses differential prediction and differential item functioning as well as population sensitivity of equating functions. Dorans compared score equity assessment to differential item functioning (DIF). DIF analysis evaluates whether the function relating item score to total score is invariant across subpopulations, whereas the score equity assessment evaluates whether the function linking one test to another test at the reported score level is invariant across subpopulations. When population invariance does not hold, it tells us that the differential difficulty of the two tests to be equated is not consistent across different subgroups. Instead, there is an interaction between the relative difficulty of the two tests and group membership, or there is an interaction among score level, difficulty, and group. For example, we equate test *X* to test *Y*. Relative to their performance on test *Y*, females find test *X* easier than test *Y* while males find test *X* harder than test *Y*. Therefore, the same test *X* score converts to a lower conversion for females than it does for males. If this difference is nontrivial, then test *X* scores and test *Y* scores are not exchangeable.

Other studies of score equity assessment provide more examples of how population invariance can be used to assess whether test scores are equatable or not. Von Davier and Wilson (2004) examined the population invariance of IRT equating for an AP Exam. Liu and Holland (2004) examined the population invariance of parallel-linear linkings across different subpopulations of the Law School Admission Test. Yang and Gao (2004) looked at invariance of linking computer-administered College-Level Examination Program® data across gender groups. Dorans, Liu, and Hammond (2004) examined the role of the anchor test in achieving population invariance across subpopulations and test administration of the SAT. These studies demonstrate across a variety of settings that population invariance is a valuable tool for evaluating test equatability.

Equatability Indices by Using Subpopulation Linking

Dorans and Holland (2000) and Dorans et al. (2003) suggested using the standardized Root Mean Square Difference (RMSD) to describe the differences between the subgroup linking functions and the total group linking functions at a given score value, and using the Root Expected Mean Square Difference (REMSD) to summarize overall differences between the linking functions. The following descriptions of RMSD and REMSD are adapted from Dorans and Holland (2000).

Root Mean Square Difference (RMSD)

The two tests to be linked are denoted by *X* (new test) and *Y* (old test), and the observed scores from these two tests are denoted by *x* and *y*, respectively. *P* represents the total population of examinees, and the subpopulations of *P* are denoted by subscripts, such as P_j. For our purposes, the set of subpopulations, $\{P_j: j = 1, 2,\ldots\}$, will always *partition P* into a set of *mutually exclusive* and *exhaustive* subpopulations.

We denote *e* as an equivalent of score *y* to score *x*, so $y = e(x)$ denotes a linking function based on any method (e.g., linear, equipercentile). Then $e_P(x)$ represents transformed scores of form *X* to the scale of form *Y* for the total group, and $e_{Pj}(x)$ represents transformed scores of form *X* to the scale of form *Y* for subgroup P_j.

Dorans and Holland's first measure of subpopulation dependence is defined at each *X* score level, *x*. It is the standardized RMSD of the subpopulation linking functions from the total group linking function for a given *x* value,

$$RMSD_{(x)} = \frac{\sqrt{\sum_j w_j \left[e_{Pj}(x) - e_P(x)\right]^2}}{\sigma_{YP}}, \quad (1)$$

where $w_j = \frac{N_j}{N}$ denotes the relative proportion of examinees from *P* that are in P_j so that $\sum_j w_j = 1$.

This measure is computed at each *x* level to quantify the difference between all subgroup linking functions and the total group linking function. The contribution of each subgroup is weighted by its proportional representation in the total group. The square root is used to bring the measure back to the scale of "unsquared" *Y* score points. The divisor, σ_{YP}, is used to make the units of this measure the proportion of the standard deviation of *Y* scores in *P*. For example, a value of 0.1 for RMSD (*x*) is interpreted as a root mean square difference of 10 percent of the standard deviation of *Y* scores in *P* in the linking functions at score *x* of *X*. Thus, RMSD (*x*) is a type of *effect size* for each *x* value (Dorans and Holland, 2000).

Note that in Formula (1), *x* denotes each *raw score* level. The divisor, σ_{YP}, is used to quantify the sum of differences between total group and subgroup linked raw scores in standard deviation units. In the present study, the linkings converted the raw scores into scaled scores on the familiar College Board 200-to-800 scale. Because most readers can understand and readily interpret values on this scale, a modified version of Formula (1) was used, which expressed the differences in scaled score units rather than in standard deviation units:

$$RMSD_{(SS)} = RMSD_{(x)} \cdot \sigma_{YP} = \sqrt{\sum_j w_j \left[e_{Pj}(X) - e_P(X)\right]^2}, \quad (2)$$

where $e_P(X)$ represents scaled scores of form *X* to the scale of form *Y* for the total group, and $e_{Pj}(X)$ represents scaled scores of form *X* to the scale of form *Y* for subgroup P_j.

$RMSD_{(SS)}$ represents RMSD at each scaled score level.

Root Expected Mean Square Difference (REMSD)

To obtain a single number summarizing the values of RMSD (x), Dorans and Holland (2000) introduced a summary measure by averaging over the distribution of X in P. This is the Root Expected Mean Square Difference (REMSD):

$$REMSD = \frac{\sqrt{E_P\left\{\sum_j w_j\left[e_{Pj}(x) - e_P(x)\right]^2\right\}}}{\sigma_{YP}} = \frac{\sqrt{\sum_j w_j E_P\left\{\left[e_{Pj}(x) - e_P(x)\right]^2\right\}}}{\sigma_{YP}}. \quad (3)$$

REMSD is a weighted average of differences between subpopulation linking functions and the total group linking function. It is a double weighted average. First, at each score level x, the difference between each subpopulation linking function is squared. These squared differences are then averaged over subpopulations using the relative size of each subpopulation as the weight for each subpopulation. Then these weighted sums of squared differences are averaged across score levels weighted by the relative number of candidates in the total population at each score level. Finally, taking the square root of that weighted average and dividing the result by the standard deviation of the old test scores in the total group produces a measure of overall equatability in a metric in which the standard deviation of the composite score is unity.

Similarly, we modified Formula (3) and put this summary measure on the 200-to-800 scale:

$$REMSD_{(SS)} = \sqrt{E_P\left\{\sum_j w_j\left[e_{Pj}(X) - e_P(X)\right]^2\right\}} = \sqrt{\sum_j w_j E_P\left\{\left[e_{Pj}(X) - e_P(X)\right]^2\right\}}, \quad (4)$$

where X denotes a random X scaled score sampled from the total population P, and $E_P\{\cdot\}$ denotes averaging over this distribution. We weighted the expected values of the squared differences using the relative frequencies of the data for X at each scaled score point.

These two measures have different uses. The REMSD can be used to summarize the overall differences between the linking functions, whereas the RMSD can give detailed information as to which X score points are the most affected by the subpopulation differences. In the present study, we first examined RMSD at each score point and then averaged the RMSD over the distribution of X in P to obtain the single number, REMSD.

Difference That Matters (DTM): How Big Is a Big Difference?

To evaluate the relative magnitude of RMSD and REMSD, Dorans and Feigenbaum (1994) used the notion of scaled score differences that matter (DTM) in the context of linking the new SAT to the old SAT. On the SAT scales, scores are reported in 10-point units (200, 210, 220…780, 790, 800). For a given raw score, if the unrounded scaled scores resulting from two separate linkings differ by fewer than 5 points, then ideally the scores should be rounded to the same reported score. For example, at a raw score of 50, the corresponded scaled scores are 710.1589 for total group and 712.3467 for females. The rounded reported scores for both groups are 710. Consequently, these two conversions at this raw score point (50) were treated as being equivalent. Dorans et al. (2003) adapted the above indices used in SAT practice to other tests and considered DTM to be half of a score unit for unrounded scores. In the present study, the DTM was therefore defined as half of the SAT score unit—5.

II. Equating Designs and Methods

In this study, score equity assessment on two current SAT forms was first conducted as a baseline. The equatability assessment on the prototypes was then carried out and compared to the baseline. This section describes data collection designs and statistical procedures used in current SAT equating practices, as well as those used in the field trial.

Equating Designs Employed in the Current SAT®

There are two types of data collection designs for equatings employed in the current SAT program: the nonequivalent groups anchor test design (*NEAT*) and random/equivalent groups (*EG*) design.

NEAT Design

At each administration of a new form, the new form is equated to four old SAT forms through an external anchor test design. One of the old forms was administered at the same time of the year as the new form. This old form is called the short leg. Each of the other three old forms was administered at one of three core administrations of the SAT that contribute large numbers of test-takers to the SAT cohort. These three old forms are called the long legs. This design has produced stable equatings because it directly acknowledges the important role that the old form linking plays in placing a new form on scale. Typical SAT equatings employ a variety of equating models, linear as well as nonlinear, observed scores as well as true score models, and equatings that

employ poststratification. In this report the focus is on nonlinear methods only.

EG Design

At each SAT administration with two new forms, the first new form is equated using the *NEAT* design, while the second new form is equated to the first one through an *EG* design. The spiraling procedure used in the administration and the large numbers of examinees who take each form usually ensure equivalent groups in the same administration. The equating results from the linear method and the equipercentile method, with and without smoothing, are then evaluated. In the baseline analyses for this study, we equated the second new form to the first new form through the *EG* design.

Equating Designs for the Prototypes in the Field Trial

There were two data collection designs in the field trial, where the new critical reading and math prototypes were administered, containing different sets of booklets for different purposes (Liu, Feigenbaum, and Walker, 2004). Design 1 was set up as an *EG* design. Books in Design 1 focused on the entire test battery: students either took a complete current SAT test (current SAT-V and SAT-M) or a complete new SAT test (critical reading, new math, and writing). Design 2 was set up as a counterbalanced single group design. Each book in Design 2 contained two versions of one component only. For example, a current SAT-V and a new critical reading prototype were administered to the same group of test-takers. For one group of the test-takers, the current test preceded the prototype; for another group, the prototype preceded the current test.

The linking results reported in this study were conducted through the *EG* design. The linear method and the equipercentile method were conducted in each of the following groups: Total, Male, and Female test-takers.[2] The smoothed equipercentile conversion was deemed most appropriate for the Total group equating, the Male-only equating, and the Female-only equating. Therefore, only the results from the smoothed equipercentile equatings are reported in the following sections.

Statistical Methods of Linking—Smoothed Equipercentile Linking

Equipercentile Method

The equipercentile linking function is set so that the cumulative distribution function of scores on form *X* converted to the form *Y* scale is equal to the cumulative distribution function of scores on form *Y* (Braun and Holland, 1982). This nonlinear transformation for the total population *P* can be expressed as:

$$e_Y(x) = G^{-1}[F(x)], \tag{5}$$

where *F* represents the cumulative distribution function of *X*, *G* is the cumulative distribution function of *Y*, and G^{-1} is the inverse of the cumulative distribution function of *Y*. The effect is that the transformed scores on *X* have the same distribution function as the scores on *Y*.

Similarly, for a subpopulation P_j, the transformation equation is:

$$e_{Yj}(x) = G_j^{-1}[F_j(x)], \tag{6}$$

where F_j is the cumulative distribution function for *X* obtained from subgroup *j*, and G_j^{-1} is the inverse of the cumulative distribution function for *Y* based on subgroup P_j.

Smoothing

Equipercentile linkings were performed on original data and on data obtained by smoothing the relevant frequency distributions. Smoothing was performed to remove irregularities in the data due to sampling variation and to remove the teeth associated with formula-scoring the tests. For both *X* and *Y* forms, smoothing was performed using the loglinear univariate model, preserving six marginal moments. The resulting plots of original and smoothed distributions were evaluated and compared to each other. It was determined that the smoothed distributions faithfully represented the trends in the data while removing the undesirable irregularities.

III. Results of Score Equity Assessment

In this section, the results are illustrated in the following order. First, the score equity analyses are presented on an old SAT-V equating as a baseline. Second, the results of linking the new SAT critical reading prototype to the old SAT-V from the field trial are discussed. Third, the score equity results of the critical reading prototype from the field trial are compared to the baseline. Similarly, the analyses on an old SAT-M equating and on the new SAT math prototype linking from the field trial are illustrated, respectively, and the results from the prototype versus the baseline are compared.

[2] Prototypes were also equated in each of the four ethnic groups: white, African American, Asian American, and Hispanic. However, due to the small sample sizes, equipercentile equatings used for these subgroups may not be acceptable. Therefore, only gender groups are discussed in this report.

Population Invariance in a Current SAT-V Equating

As a baseline check, the equatings of two editions of the SAT-V (*X* and *Y*) that were given in the same SAT administration were examined for population invariance. As discussed on page 5, the second new form *X* was equated to the first new form *Y* through an *EG* design. In other words, we can view form *X* as the new form and form *Y* as the old form. The smoothed equipercentile linkings were conducted for Total, Male, and Female examinees. We examined the differences between each gender subgroup linking and the total group linking, the percentage of formula scores for which the total group and subgroup conversions differ by at least 5 points, the percentage of the examinees for whom these conversions provide scores that differ by at least 5 points, the RMSD values at each score level, the REMSD value, and the DTM value.

Descriptive Statistics

Table 1 provides descriptive statistics for Total, Male, and Female examinees on *X* and *Y* raw scores. Data in Table 1 show that *X* and *Y* scores on the total population and the two subpopulations were fairly symmetric, as indicated by the near zero skewness values. All the distributions were slightly platykurtic as compared with the normal distribution. Each group obtained a higher mean on the new form *X*, with slightly higher variation in scores on *X*.

Scaled Score Differences

Figure 1 presents the differences between the actual conversions for Male-Total and Female-Total. The differences between each subgroup conversion and total group conversion all fell within 5 scaled score points (a DTM). Therefore, the linkings for this test were considered invariant across the Male and Female subgroups and hence were considered to be equatings.

Another pertinent piece of evidence about population invariance is captured in Figure 2, where the RMSD curve

Table 1

Formula Score Descriptive Statistics in an Old SAT-V Equating

		N	*Mean*	*S.D.*	*Min.*	*Max.*	*Skewness*	*Kurtosis*
Y	Total	115,089	35.86	16.87	-11	78	.07	2.29
	Male	52,509	36.00	17.05	-11	78	.07	2.25
	Female	62,579	35.74	16.72	-10	78	.08	2.32
X	Total	112,367	37.39	17.70	-11	78	-.04	2.20
	Male	50,974	37.93	17.85	-11	78	-.10	2.18
	Female	61,393	36.94	17.56	-11	78	.01	2.22

Note: The summation of Male and Female sample sizes may not be equal to the sample size of total group due to nonresponse.

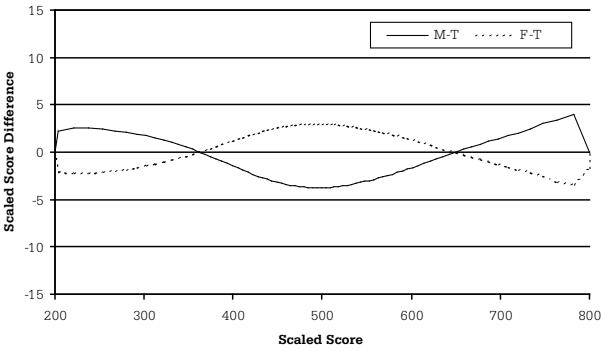

Figure 1. Form *X* Verbal scaled score differences by gender.

and the overall REMSD line are compared to the DTM line. The solid line at an ordinate value of 5 denotes the DTM. Another line around 2 denotes the average REMSD, which is well below the DTM line. The curve denotes the RMSD at each score level. For all the score levels, the RMSD fell well below the DTM of 5.

While the figures present one view of subpopulation invariance, Table 2 provides the exact values that highlight the differences between each pair of linkings (M-T and F-T) and RMSD values at each score level. The first set of three columns present linked scores of *X* for the total group and for each subgroup at each formula score level. Because the scores are reported on the 200-to-800 scale, scores below 200 were truncated to 200, and scores above 800 were truncated to 800. The second set of two columns show the differences between the pairs of linkings for Male versus Total and Female versus Total, respectively, with differences less than 5 displayed as "0." The next two columns present RMSD and REMSD values, respectively. As can be seen, the results were virtually identical across groups, with no conversion differences of 5 or larger for either Males-Total or Females-Total. The RMSD values were smaller than DTM across the entire scale range. The REMSD of 2.16 was smaller than the DTM of 5.

Table 3 summarizes the differences in equating results between each pair. For each subgroup, means and standard deviations are listed for that subgroup when the total group

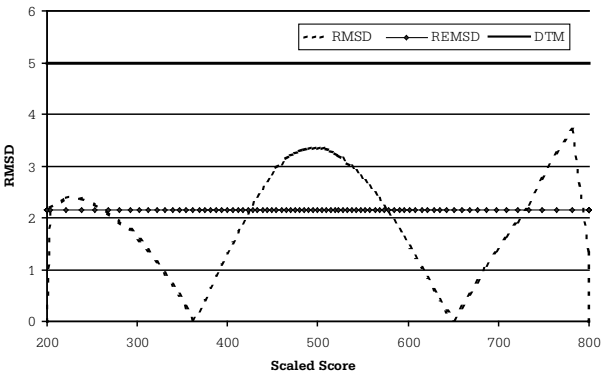

Figure 2. Form *X* Verbal RMSD by gender.

Table 2

Unrounded and Truncated Linkings for the *X* Form in an Old SAT-V Equating

	Scaled Score			*Differences*			
FS	*Total*	*Male*	*Female*	*M-T*	*F-T*	*RMSD*	*REMSD*
-19	200	200	200	0	0	0	2.16
.	.	.	.	.	.	.	.
-3	204.0746	206.3486	201.9659	0	0	2.19	.
-2	221.2555	223.7829	219.0046	0	0	2.38	.
.	.	.	.	.	.	.	.
73	764.0509	767.4845	760.9194	0	0	3.27	.
74	781.9661	785.9866	778.5631	0	0	3.70	.
.	.	.	.	.	.	.	.
78	800	800	800	0	0	0	2.16

Note: Scores below 200 were truncated to 200, and scores above 800 were truncated to 800.

conversion was used and when the subgroup conversion was used. Also listed are the difference in means, the percentage of formula scores with an absolute unrounded scaled score difference equal to or larger than 5, and the percentage of examinees whose conversions resulted in scores that differ by at least 5 points. Examination of the data reveals that the mean difference between each subgroup conversion and the total group conversion was quite small, with standardized mean difference near zero for each subgroup. The proportion of formula scores for which scaled scores between the total group conversion and the subgroup conversion differed by more than 5 points was zero. The percentage of examinees whose conversions resulted in scores that differed by more than 5 points was also zero. Even though the results exhibited slight degrees of departure from invariance across gender groups, it was not large enough to warrant any concern.

Table 3

Summary Statistics of Scaled Scores Based on Total Group Equating and Subgroup Equating in an Old SAT-V

	Total	*Male*	*Female*
Sample size and % of each subgroup in total	112,367	50,974	61,393
		45.36%	54.64%
Mean & S.D. based on total group conversion	506.7	509.5	504.4
	101.6	102.6	100.7
Mean & S.D. based on subgroup conversion		507.5	506.0
		102.7	100.6
Total conv. mean - subgroup conv. mean		2.0	-1.6
Standardized mean difference		.02	-.02
% FS with |unrounded scaled score diff.| ≥ 5		0	0
% Examinees with |unrounded scaled score diff.| ≥ 5		0	0

In summary, no evidence was found to question the score equity with respect to gender on this particular SAT-V equating. This finding was consistent with the results on the study by Dorans and Holland (2000).

Population Invariance in the New SAT Critical Reading Prototype Linking

The two designs in the field trial were *EG* design and single group counterbalanced design. For the purpose of comparison to the baseline where *EG* design was employed, it is preferable to use the results from the same design as the baseline. Plus, the preliminary results indicated the existence of a large order effect between test-takers' performance on both the current test and the prototype when they were given in different orders. Because of these factors, it was decided that the linkings and score equity assessment would be performed using the *EG* design.

Descriptive Statistics

Table 4 provides descriptive statistics for Total, Male, and Female examinees for *X* = critical reading (*CR*) scores and *Y* = current verbal (*OV*) scores. Both *CR* and *OV* scores were positively skewed for the Total, Male, and Female groups. For each group, the means and standard deviations on the *OV* and *CR* are not directly comparable due to the different test lengths. To compare how relatively test-takers performed on *OV* and *CR*, we use the percentages of means and standard deviations out of the possible range of scores (e.g., -19 to 78 for *OV* and -17 to 67 for *CR*). All three groups performed worse on *CR* with relatively larger variation in scores, and Females performed worse to a larger extent than Males.

The correlation between *OV* and *CR* was calculated

Table 4

Formula Score Descriptive Statistics in the SAT-V (*OV*) and Critical Reading (*CR*) from the Field Trial

		N	*Mean*	*Mean %*	*S.D.*	*S.D. %*	*Min.*	*Max.*	*Skew*	*Kurtosis*
OV	Total	5,344	33.01	53.62	17.25	17.78	-8	77	.17	2.26
	Male	2,283	32.31	52.90	17.55	18.09	-8	76	.22	2.25
	Female	3,055	33.56	54.19	17.00	17.53	-7	77	.14	2.28
CR	Total	9,194	27.00	52.38	15.52	18.48	-8	67	.23	2.30
	Male	3,801	26.85	52.20	15.89	18.92	-8	66	.22	2.27
	Female	5,374	27.13	52.54	15.24	18.14	-8	67	.23	2.32
		N	*Corr.*							
Corr. btw. *OV* & *CR*	Total	3,126	.912							
	Male	1,314	.900							
	Female	1,807	.921							

Note: The summation of Male and Female sample sizes may not be equal to the sample size of total group due to nonresponses.

by using data obtained from the single group design. The observed-score *CR-OV* correlations for the Total group, Males, and Females were .912, .900, and .921, respectively. These correlations are larger than .866, which is the minimum correlation needed for equating two tests, suggested by Dorans (2004b). The reliability estimates for both tests were .93. Hence, the estimated true-score correlations were about .981, .968, and .990 for Total, Males, and Females, respectively. The magnitudes of these numbers indicate that the two tests measure the same construct in nearly the same way within the three groups.

Scaled Score Differences

Figure 3 graphically displays the differences in linking results for each pair (M-T and F-T). As illustrated in Figure 3, Females would have had slightly higher scores in the middle portion of the score scale and lower scores at the ends of the score scale if the Female-only conversion were used. Conversely, Males would have obtained slightly lower scores in the middle range and higher scores at the ends of the score range if the Male-only conversion were used. The plot indicates that even though the subgroup conversions diverge slightly more from the total conversion than was seen in the baseline (see Figure 1), the differences were less than 5 scaled score points across the entire score range.

Figure 4 depicts RMSD values compared to the DTM line at each score level. As seen in Figure 4, the RMSD fell below the DTM line virtually across the entire score range. The REMSD value of approximately 3 was below the DTM of 5. Therefore, the lack of invariance of the linkings for the prototype across the gender groups was not enough to cause any concern.

Table 5 shows the differences between subgroup and total group linkings, where the differences of at least 5 points are highlighted in bold. An examination of the data indicates that the smoothed equipercentile equivalents were virtually identical across groups, except for the conversion differences of 5 or more at 3 formula scores (-2, -1, and 65) for Males, and at 1 formula score (-2) for Females. Compared to the DTM of 5, the REMSD of 2.98 was well below the DTM. The RMSD values were smaller than the DTM across the entire scale range (except at formula score –2), which indicates that the linking functions for each subgroup could be considered to be

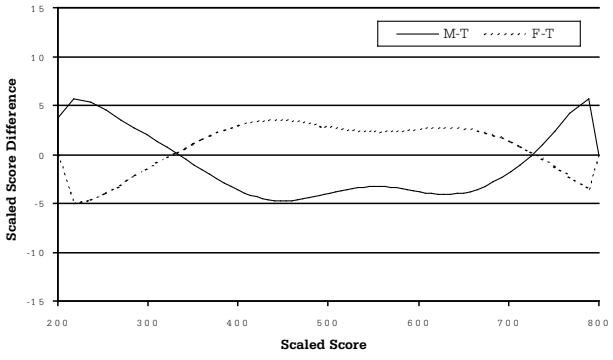

Figure 3. Critical Reading scaled score differences by gender.

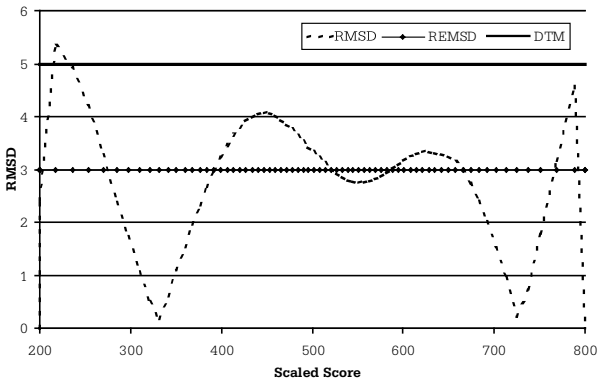

Figure 4. Critical Reading RMSD by gender.

Table 5

Unrounded and Truncated Linking for *CR* to *OV* from the Field Trial

	Scaled Score			*Differences*			
FS	*Total*	*Male*	*Female*	*M-T*	*F-T*	*RMSD*	*REMSD*
-17	200	200	200	0	0	0	2.98
.	.	.	.	.	.	.	.
-2	217.3360	223.0539	212.2967	**6**	**-5**	**5.33**	.
-1	236.1519	241.5320	231.5228	**5**	0	4.95	.
.	.	.	.	.	.	.	.
65	788.6496	794.3890	785.0992	**6**	0	4.58	.
.	.	.	.	.	.	.	.
67	800	800	800	0	0	0	2.98

Note: Scores below 200 were truncated to 200, and scores above 800 were truncated to 800.

Table 6

Summary Statistics of Scaled Score Based on Total Group Linking and Subgroup Linking in the Critical Reading Prototype

	Total	*Male*	*Female*
Sample size and % of each subgroup in total	9,194	3,801	5,374
		41.34%	58.45%
Mean & S.D. based on total group conversion	479.4	477.9	480.4
	107.8	111.0	105.3
Mean & S.D. based on subgroup conversion		474.9	482.8
		110.0	105.8
Total conv. mean - subgroup conv. mean		3.0	-2.3
Standardized mean difference		.03	-.02
% FS with |unrounded scaled score diff.| ≥ 5		3.5	1.2
% Examinees with |unrounded scaled score diff.| ≥ 5		0.7	0.4

Table 7

Comparison of Population Invariance in the Baseline Versus in the Prototype

	Baseline		*Prototype*	
	M	*F*	*M*	*F*
Total conv. mean - subgroup conv. mean	2.0	-1.6	3.0	-2.3
Standardized mean difference	.02	-.02	.03	-.02
% FS with |unrounded scaled score diff.| ≥ 5	0	0	3.5	1.2
% Examinees with |unrounded scaled score diff.| ≥ 5	0	0	0.7	0.4
REMSD	2.16		2.98	

Note: Data presented in this table are also shown in Table 3 and Table 6, for baseline and prototype, respectively. They are combined in this table for the purpose of comparison.

invariant from that of the total group and hence could be considered to be equating.

Table 6 summarizes differences between each subgroup conversion and the total group conversion. In general, Table 6 shows that Males would have had a lower mean with a Male-only conversion than they obtained with the total group conversion, while Females would have had a higher mean with a Female-only conversion than they obtained with the total group conversion. Therefore, the total group conversion seems to advantage Males but disadvantage Females. The percentage of examinees for whom the conversions differed by 5 or more was 0.7 percent and 0.4 percent, respectively, for Males and Females.

Comparison of SAT-V Baseline Versus Critical Reading Prototype

Table 7 presents the differences between subgroup linkings from the total group linking in the baseline and in the prototype, respectively. For the Male group, the difference between the Total-conversion mean and the Male-conversion mean was slightly larger in the prototype than in the baseline, but the standardized mean differences were very close to each other. The percentage of formula scores with unrounded scaled score differences larger than 5, and the percentage of examinees whose scaled score differed by more than 5 were also slightly larger in the prototype than in the baseline. For the Female group, the difference between the Total-conversion mean and the Female-conversion mean was also slightly larger in the prototype, but the standardized mean differences were the same. Both of the percentage indices were slightly higher in the prototype than in the baseline. The comparison of REMSD between the baseline and the prototype showed that the departures of the gender group conversions from the total conversion in the prototype were slightly larger than those of the current test, but not beyond the DTM range.

Overall, the comparison of the prototype and the baseline analyses suggested more divergence of the subgroup conversions from the total conversion in the critical reading prototype, but not enough to cause any concerns. Using the stringent standards for this study, population invariance for linking the new SAT critical reading prototype to the current SAT-V was achieved across gender groups.

Population Invariance in a Current SAT-M Equating

A similar set of analyses was performed on the current SAT–M. The new form *X* and the old form *Y* were administered to the large equivalent groups in the same

Table 8

Formula Score Descriptive Statistics in an Old SAT-M Equating

		N	*Mean*	*S.D.*	*Min.*	*Max.*	*Skew*	*Kurtosis*
Y	Total	115,089	29.90	13.06	-10	60	-.03	2.33
	Male	52,509	32.37	13.25	-8	60	-.19	2.37
	Female	62,579	27.84	12.54	-10	60	.05	2.37
X	Total	112,367	29.69	13.84	-7	60	-.08	2.21
	Male	50,974	31.97	14.02	-7	60	-.23	2.25
	Female	61,393	27.80	13.40	-7	60	.02	2.25

Note: The summation of Male and Female sample sizes may not be equal to the sample size of total group due to nonresponses.

SAT administration, and the new form *X* was linked to the old form *Y* through an *EG* design on three different groups: Total, Males, and Females. In each group, the linear method and the smoothed and unsmoothed equipercentile methods were performed. Again, we evaluated the linking results for all the methods, and the smoothed equipercentile conversions were deemed most appropriate for the Total group, the Male-only, and the Female-only linkings. Therefore, in the following discussion, we focus only on the results based on smoothed equipercentile (*SE*) linkings.

Descriptive Statistics

Table 8 provides descriptive statistics for Total, Males, and Females on *X* and *Y* scores. The scores of the Total group and the Female group on both *X* and *Y* were fairly symmetric, as indicated by the near zero skewness values. For the Male group, both *X* and *Y* scores were clearly negatively skewed. All the distributions were slightly platykurtic when compared with normal distribution. Each group obtained a slightly lower mean on the new form *X*, with slightly more variation in scores.

Scaled Score Differences

Figure 5 provides the differences between each of the subgroup conversions and the total group conversion. It appears that Males would have obtained higher scaled scores if the Male-only conversion were used, while Females would have received lower scaled scores if the Female-only conversion were used. The differences between each subgroup and the total group fell within 5 scaled score points along almost the entire score range.

Figure 6 presents the RMSD curve and the REMSD line, compared to the DTM line. As shown in Figure 6, the RMSD fell below the DTM line for all of the score levels. The REMSD line was well below the DTM.

Table 9 highlights the conversion points where the difference between each of the subgroup conversions and the total group conversion was larger than or equal to 5. As shown in Table 9, most of these differences occurred between scaled scores of 700 to 800. Note that the tabular portion of the display for the Female group has three conversion points where differences are greater than 5 at the high end, while the Male group has none. All in all, the linking results for the subgroups were virtually identical to the total group linking results. The REMSD value (2.37) was smaller than the DTM.

Table 10 presents the summary statistics based on the linking results for each group. The absolute mean differences in subgroup linking and the total group linking were -2.4 and 2.1, and the standardized mean differences were near zero. The percentage of formula scores for which the scaled score differences were 5 or more was 0

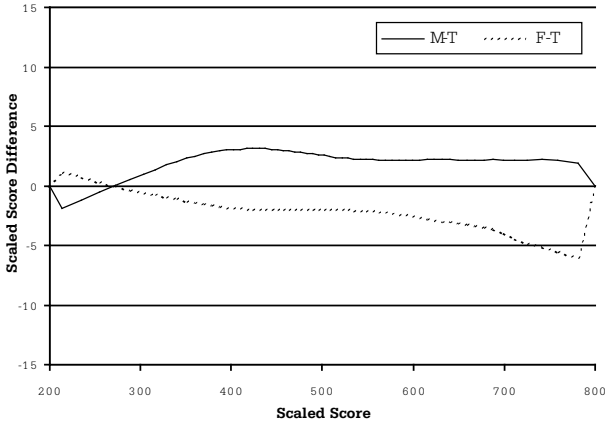

Figure 5. Form *X* Math scaled score differences by gender.

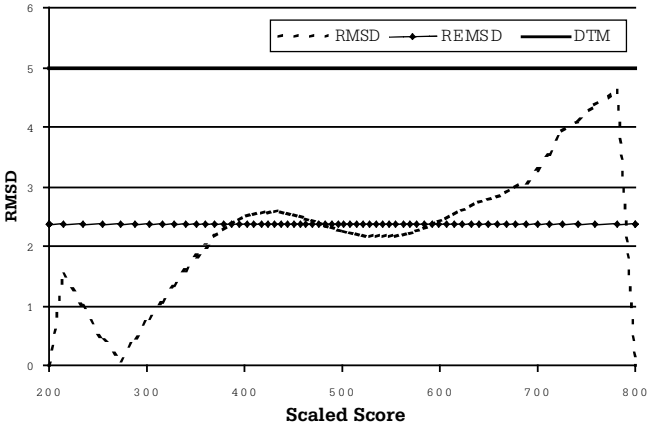

Figure 6. Form *X* Math RMSD by gender.

Table 9

Unrounded and Truncated Linkings for the *X* Form in a Current SAT-M Equating

	Scaled score			*Differences*			
FS	*Total*	*Male*	*Female*	*M-T*	*F-T*	*RMSD*	*REMSD*
-14	200	200	200	0	0	0	2.37
.	.	.	.	.	.	.	.
56	740.9575	743.1866	735.8140	0	**5**	4.09	.
57	759.1151	761.2243	753.5423	0	**6**	4.36	.
58	780.7958	782.6984	774.7824	0	**6**	4.63	.
.	.	.	.	.	.		.
60	800	800	800	0	0	0	2.37

Note: Scores below 200 were truncated to 200, and scores above 800 were truncated to 800.

Table 10

Summary Statistics of Scaled Score Based on Total Group Equating and Subgroup Equating in an Old SAT-M

	Total	*Male*	*Female*
Sample size and % of each sub-group in total	112,367	50,974	61,393
		45.36%	54.64%
Mean & S.D. based on total group conversion	515.8	532.9	501.7
	103.4	105.9	99.1
Mean & S.D. based on subgroup conversion		535.2	499.5
		105.7	98.5
Total conv. mean - subgroup conv. mean		-2.4	2.1
Standardized mean difference		-.02	.02
% FS with |unrounded scaled score diff.| ≥ 5		0.0	4.0
% Examinees with |unrounded scaled score diff.| ≥ 5		0.0	1.2

percent and 4 percent for Male and Female, respectively. For the Male group, no examinees showed an unrounded scaled score difference larger than or equal to 5 between the Male-only conversion and the Total conversion; and only 1.2 percent of the Female examinees had a difference of 5 or more. The effects can be considered negligible.

Overall, the above analyses reveal that in this standard SAT linking, the two Math forms were equatable from the perspective of population invariance, and hence it could be considered as equating.

Population Invariance in the New SAT Math Prototype Linking

The results described below were performed by using data from *EG* design in order to compare to the baseline results.

Descriptive Statistics

Table 11 provides descriptive statistics for the current Math (*OM*) scores and the new Math (*NM*) prototype scores from the field trial. For the Total and the Female groups, both the *OM* and the *NM* scores were positively skewed. For the Male group, both the *OM* and the *NM* scores were fairly symmetric. The means and standard deviations of the two tests were not directly comparable due to the different test lengths. Therefore, we compared the percentages of the means and standard deviations out of the possible range of score. Total and Female groups performed slightly worse on the *NM*, whereas Males performed similarly on the *NM* and the *OM*. All three groups had slightly larger variation in *NM* scores.

The observed-score *OM*–*NM* correlations were obtained from single group design in Design 2. For the Total group, Males, and Females, the correlations were .922, .923, and .918, respectively. These values meet the minimum criterion of .866, suggesting that the two tests

Table 11

Formula Score Descriptive Statistics in the Old Math and New Math from the Field Trial

		N	*Mean*	*Mean %*	*S.D.*	*S.D. %*	*Min.*	*Max.*	*Skew*	*Kurtosis*
OM	Total	5,344	25.29	53.09	13.68	18.49	-7	60	.21	2.31
	Male	2,283	27.12	55.57	14.20	19.19	-6	60	.09	2.21
	Female	3,055	23.94	51.27	13.12	17.73	-7	60	.27	2.41
NM	Total	9,194	23.16	52.55	12.58	19.35	-7	54	.14	2.31
	Male	3,801	25.09	55.52	13.05	20.08	-7	54	.01	2.23
	Female	5,374	21.79	50.45	12.05	18.54	-7	53	.21	2.41
		N	*Corr.*							
Corr. btw. *OM* & *NM*	Total	3,019	.922							
	Male	1,299	.923							
	Female	1,717	.918							

Note: The summation of Male and Female sample sizes may not be equal to the sample size of total group due to nonresponses.

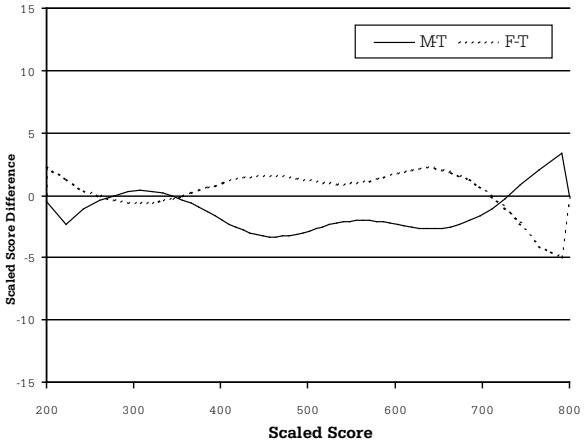

Figure 7. New Math scaled score differences by gender.

measure the same construct. The reliability estimates for the two tests were .92 and .93. The estimated true-score correlations were about .997, .998, and .992 for Total, Males, and Females, respectively.

Scaled Score Differences

The comparisons between each subgroup conversion and the total group conversion are provided in Figure 7. The data indicate that Males would have obtained lower scores if the Male-only conversion were used, and Females would have received slightly higher scores if the Female-only conversion were used. This was the case at most score levels, with the exception of the high end of the score scale, where the subgroup conversions diverged from the Total group conversion in the opposite direction. Overall, when compared with the baseline, the degrees of divergences of the subgroup conversions from the total group conversion were similar and fell within the range of 5 scaled score points.

Figure 8 displays how RMSD varies across score levels, and the REMSD line compared with the DTM line. For all the score levels, the RMSD was below the DTM. This

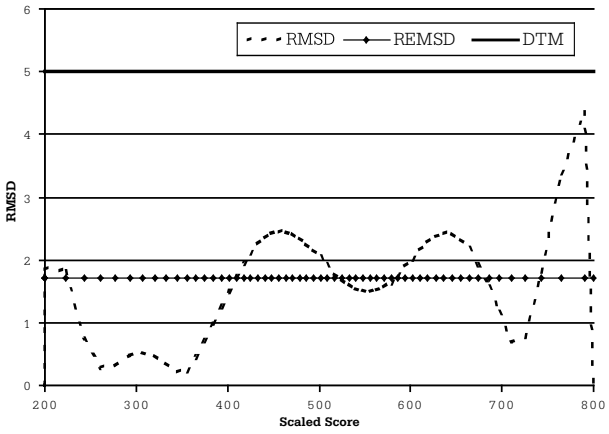

Figure 8. New Math RMSD by gender.

Table 12

Unrounded and Truncated Linking for the New Math to the Old Math from the Field Trial

FS	Scaled Score			Differences		RMSD	REMSD
	Total	Male	Female	M-T	F-T		
-14	200	200	200	0	0	0	1.71
.	.	.	.	.	.	.	.
-2	222.9097	220.5220	224.223	0	0	1.83	.
-1	242.8932	241.8416	243.2545	0	0	0.73	.
.	.	.	.	.	.	.	.
52	764.5600	766.6080	760.5160	0	0	3.36	.
53	790.9161	794.2690	785.9645	0	-5	4.36	.
54	800	800	800	0	0	0	1.71

Note: Scores below 200 were truncated to 200, and scores above 800 were truncated to 800.

Table 13

Summary Statistics of Total Group Linking and Subgroup Linking in the Math Prototype

	Total	Male	Female
Sample size and % of each subgroup in total	9,194	3,801	5,374
		41.34%	58.45%
Mean & S.D. based on total group conversion	485.9	501.8	474.6
	105.9	110.1	101.3
Mean & S.D. based on subgroup conversion		499.7	475.6
		109.9	101.6
Total conv. mean - subgroup conv. mean		2.1	-1.0
Standardized mean difference		0.02	-0.01
% FS with |unrounded scaled score diff.| ≥ 5		0.0	0.0
% Examinees with |unrounded scaled score diff.| ≥ 5		0.0	0.0

Table 14

Comparison of Population Invariance in the Math Baseline and in the Prototype

	Baseline		Prototype	
	M	F	M	F
Total conv. mean - subgroup conv. mean	-2.4	2.1	2.1	-1.0
Standardized mean difference	-.02	.02	.02	-.01
% FS with |unrounded scaled score diff.| ≥ 5	0.0	4.0	0.0	0.0
% Examinees with |unrounded scaled score diff.| ≥ 5	0.0	1.2	0.0	0.0
REMSD	2.37		1.71	

Note: Data presented in this table are also shown in Table 10 and Table 13, for baseline and prototype, respectively. They are combined in this table for the purpose of comparison.

indicates that the linking functions for each subgroup were very close to that of the total group.

Table 12 shows the portions of separate linkings where the subgroup conversion difference from the total group conversion was larger than or equal to 5. As can be seen, there were no such 5-point or larger differences for Males, and there was one 5-point difference for Females. The REMSD value of 1.71 was much smaller than the DTM of 5, which indicates that the linkings for this test were invariant across the Male and Female subgroups.

Table 13 presents the summary statistics based on the total group and subgroup linkings. The two percentage indices are reported as well. Males would have had a lower mean with the Male-only conversion, and Females would have received a slightly higher mean with the Female-only conversion. The standardized mean difference was near zero for each subgroup. The percentage statistics were negligible (zero). This further confirms that the linkings for the *NM* prototype to the *OM* were invariant across the Male and Female subgroups. Therefore, it could be considered as equating.

Comparison of SAT-M Baseline Analyses Versus New SAT Math Prototype Analyses

Table 14 presents the comparison between the baseline analyses and the prototype analyses. As can be seen, the raw differences between the total conversion means and the subgroup conversion means were slightly smaller in the prototype for both Males and Females. The REMSD value in the prototype was much smaller than the DTM value and deviated to a lesser extent than in the baseline. In summary, no evidence was found to question population invariance across gender groups when linking the new Math prototype to the current SAT-M.

An interesting finding here is that the divergences of the subgroup conversions from the total group conversion exhibited an opposite direction for the new Math prototype than for the current Math. For the current Math, Males would have obtained a higher mean under the Male-only conversion, while Females would have had a lower mean under the Female-only conversion. For the prototype, we observed results on the opposite direction: Males would have had a lower mean under the Male-only conversion, whereas Females would have obtained a higher mean under the Female-only conversion. What this tells us is that Males found the current Math harder while Females found it easier; and Males found the new Math prototype easier while Females found it harder. This might be caused by the addition of more advanced items. This needs further exploration. Nonetheless, the differences are not substantial.

IV. Implications for Equating Practices with New SAT

In the present study, score equity assessment was performed on the linkings of the new SAT critical reading prototype to the current SAT-V, and on the new SAT math prototype to the current SAT-M. The analyses focused on gender groups. The subgroup-only conversions were compared to the total group conversion. Both RMSD and REMSD measures were examined and compared to the Difference That Matters (DTM) criterion. A major reason for conducting this study was to examine whether linking functions for the new SAT and the current SAT were essentially the same across different major subpopulations. If the relationship between the two tests depends on whether examinees are male or female, then the tests are probably not measuring the same thing with comparable degrees of reliability, and treating the score from the tests as if they were interchangeable would be questionable practice.

The results of the equatability analysis suggest that population invariance for the new critical reading section was achieved from the perspective of gender groups. The critical reading section was equatable to the current SAT-V according to the stringent criterion of equatability employed in this study. Even though the prototype exhibited more divergence than the SAT-V baseline, the degree of divergence is not enough to cause concern.

For math, the results suggest that the linkages of the new Math prototype to the current SAT-M were invariant with respect to gender groups. Males performed slightly better on the new Math than they did on the SAT-M, whereas females performed slightly worse on average on the new Math than they did on SAT-M. However, the differences were negligible for the most part. The prototype even showed smaller divergence than the SAT-M baseline. Based on the results from this study, the new Math is equatable to the current Math.

V. Summary

The field trial data suggest that the new SAT critical reading prototype measures the same construct as the current SAT-V in the Male and Female groups relative to the Total group. For the new Math prototype, the data suggest that the same construct is being measured in the Male group and Female group relative to the Total group. However, there are certain possible criticisms of these analyses that need to be pointed out.

First, the field trial population was not representative of the regular SAT test-taking population. While nonrepresentativeness may affect the generalizability of effect results, it does not invalidate the differences that exist across the current and new material in this study. As indicated above, the data used in these analyses were based on the random assignment of forms to examinees. Groups were randomly assigned to either new or old test material. While the samples were nonrepresentative of the general population, comparisons of the groups on different types of test material are still valid.

Second, the sample contained unmotivated examinees. In order for lack of motivation to have effects on the comparative analyses presented here, the effects have to have been differential across the old and new test. In other words, those who took the new material would have to have been more or less motivated than those who took the old material. There is no reason to expect this differential effect. If motivation was an issue, it was likely an issue across all the material, and while it may have diminished overall performance and added noise to the data, it did so independently of material taken.

Third, we need to keep in mind that scores on one test can be considered successfully equated to scores on another test only *with respect to some population or populations*. One can invariably define other populations for which the scores on the two tests in question are not exchangeable. To the extent that gender groups constitute populations of primary interest, the results of the current study provide evidence that the new SAT critical reading and math prototypes can be equated to the current SAT. However, as pointed out earlier, the ethnic group sample sizes were too small to support sound linkings at the ethnic subgroup level that would yield valid inferences about population invariance. Examination of score equity across ethnic groups requires further data collection and study.

References

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

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus post-stratification methods for equating and test linking. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program Examinations* (ETS RR-03-27, pp. 19–35). Princeton, NJ: Educational Testing Service.

von Davier, A. A., & Wilson, C. (2004). Population invariance of IRT equating for Advanced Placement (AP) Program Exams. Paper presented in the symposium, *Population Invariance of Test Equating and Linking: Theory Extension and Applications Across Exams*, at the annual meeting of the National Council on Measurement in Education, San Diego.

Dorans, N. J. (2004a). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41* (*1*), 43–68.

Dorans, N. J. (2004b). Equating, concordance and expectation. *Applied Psychological Measurement*, *28* (*4*) 227–246.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT*. (ETS Research Memorandum 94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37* (*4*), 281–306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program Examinations. In N. J. Dorans, (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program Examinations*. (ETS RR-03-27, pp. 79–118). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Liu, J., & Hammond, S. (2004). The role of the anchor test in achieving population invariance across subpopulations and test administrations. Paper presented in the symposium, *Population Invariance of Test Equating and Linking: Theory Extension and Applications Across Exams*, at the annual meeting of the National Council on Measurement in Education, San Diego.

Liu, J., Feigenbaum, M., & Walker, M. E. (2004). New SAT and PSAT/NMSQT® Spring 2003 field trial design. Paper presented in the symposium, *Analysis of spring 2003 new SAT and new PSAT/NMSQT field trial*, at the annual meeting of the National Council on Measurement in Education, San Diego.

Liu, M., & Holland, P. W. (2004). Exploring the population sensitivity of linking functions using LSAT subpopulations. Paper presented in the symposium, *Population Invariance of Test Equating and Linking: Theory Extension and Applications Across Exams*, at the annual meeting of the National Council on Measurement in Education, San Diego.

Yang, W.-L., Dorans, N. J., & Tateneni, K. (2003). Effect of sample selection on AP multiple-choice score to composite score linking. In N. J. Dorans, (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program Examinations*. (ETS RR-03-27, pp. 57–78). Princeton, NJ: Educational Testing Service.

Yang, W.-L., & Gao, R. (2004). Invariance of Score Linking Across Gender Groups for Testlet-Based CLEP Examinations. Paper presented in the symposium, *Population Invariance of Test Equating and Linking: Theory Extension and Applications Across Exams*, at the annual meeting of the National Council on Measurement in Education, San Diego.