Research Report

Score Equity Assessment: Development of a Prototype Analysis Using SAT® Mathematics Test Data Across Several Administrations

Neil J. Dorans

Jinghua Liu

March 2009

ETS RR-09-08

Score Equity Assessment: Development of a Prototype Analysis Using SAT$^{®}$ Mathematics Test Data Across Several Administrations

Neil J. Dorans and Jinghua Liu
ETS, Princeton, NJ

March 2009

Abstract

The equating process links scores from different editions of the same test. For testing programs that build nearly parallel forms to the same explicit content and statistical specifications and administer forms under the same conditions, the linkings between the forms are expected to be equatings. Score equity assessment (SEA) provides a useful tool to check form equatability. We suggest use of SEA as a quality control tool to evaluate how well a test assembly process works over several administrations. The examination of multiple forms should provide a proper assessment of the fairness of the test assembly process. We illustrate how to include SEA into statistical and psychometric practice with data from several administrations of the SAT®-Math.

Key words: SAT, test assembly, score equity assessment, equating, linking

Table of Contents

Page

List of Tables

List of Figures

Page

1. Overview

Dorans (2004b) introduced and placed score equity assessment (SEA) within a fairness context that included differential prediction analysis (DPA) and differential item functioning (DIF). The notion of subpopulation invariance is central to all three analyses. Fairness in the SEA, DIF, and DPA contexts is defined in terms of the invariance of a relationship across important subpopulations. DPA examines whether a test (alone or in conjunction with other information) predicts an external criterion in much the same way across different subpopulations. DIF examines how performance on the item varies across different subpopulations; lack of DIF means the relationship is the same across subpopulations. SEA examines whether the linking relationship in the total population holds up across different subpopulations. In all three cases, assessment of fairness involves checking for invariant relationships across subpopulations.

DIF, SEA, and DPA can be fit into a 2-by-2 framework for fairness procedures by crossing item/test with internal/external validity. DIF examines the internal validity of a test at the item level, SEA examines internal validity at the test level, and DPA evaluates the external validity of a test at the test level. All three aspects of fairness should be addressed. Some testing programs routinely address DIF, but not SEA nor DPA. SEA, like differential prediction, addresses issues that are most germane to the major product of the assessment process, test scores. Unlike differential prediction, SEA does not require the collection of additional data on external criteria. This paper examines how SEA might be incorporated into operational testing programs, and be used as a stand-alone fairness procedure or as a complement to existing DIF procedures.

In section 2 of this paper, we summarize how fairness has been assessed over the past several decades beginning with differential prediction and moving onto DIF and then finish the review with SEA. We describe the SEA indexes in section 3. In section 4, we describe the nature of SAT equating. In section 5, we conduct the subgroup equatings, apply the indexes to operational SAT data, and report our results. Finally we address questions such as required sample sizes, why certain groups are or are not studied under different data collection designs, and what should be done if SEA uncovers subpopulation sensitivity.

2. A Brief Review of Fairness Assessment

Fairness concerns have been with us for centuries. In the context of standardized tests, fairness has been a major policy issue for the last five decades. During the late 1960s, the discussion about fair assessment was heated. Testing results had indicated for decades that noticeable disparities existed in average performance between different groups (e.g., White and Black students). In the 1960s, the tests were accused of measuring the wrong things. A common expectation was that if the tests measured the correct things then differences between subgroups would be smaller. Tests also were criticized because they were used inappropriately, a criticism that has continued through the remainder of the 20th century to the present.

In this section, we review fairness procedures that have been developed over the past decades. While some reference is made to events external to ETS, the focus is on ETS, which has been viewed as a leader in fairness assessment at the item level for the past 20 years. First we consider differential prediction and differential validity, procedures that examine how consistently tests scores predict criteria, performance on a job or in the classroom, across different subgroups. Then we spend time on the item level assessments, the quantitative differential item functioning, and the qualitative test fairness review process. Finally, we examine score equity assessment, the focus of this study, and review some of its uses over the past few years.

2.1 Fair Test Use as a Lack of Differential Prediction

The 1970s witnessed the beginning of a series of differential validity and differential prediction studies. The Supreme Court decision *Griggs v. Duke Power Co.* in 1971 included the terms *business necessity* and *adverse impact*, both of which affected employment testing. Adverse impact is a substantially different rate of selection in hiring, promotion, transfer, training, or other employment-related decisions for any race, sex, or ethnic group. Business necessity can be used by an employer as a defense when the employer has a criterion for selection that appears to be neutral but that excludes members of one sex, race, national origin, or religious group at a substantially higher rate than members of other groups. The employer must prove that its selection requirement having the adverse impact is job related and consistent with business necessity. In other words, in addition to appearing race/ethnic/gender neutral, the selection instrument had to have demonstrated validity for its use. Ideally, this validity would be the same for all subpopulations.

Young (2001) reviewed studies from as far back as 1974 that examined either differential validity (a difference in test/criterion correlations between, for example, males and females) or differential prediction (a difference in predicted grades for say males and females). Differential prediction analyses (DPA) are preferred to differential validity studies because differences in predictor or criterion variability can produce differential validity even when the prediction model is fair (Linn, 1975). Differential prediction analyses examines whether the same prediction models hold across different groups.

Petersen and Novick (1976) in the lead article in a special issue dedicated to fair selection compared several models for assessing fair selection, including the regression model (Cleary, 1968), the constant ratio model (Thorndike, 1971), the conditional probability model (Cole, 1973), and the constant probability model (Linn, 1973). They demonstrated that the regression, or Cleary, model, which is a differential prediction model, was a preferred model from a logical perspective in that it was consistent with its converse (i.e., fair selection of applicants was consistent with fair rejection of applicants). In essence, the Cleary model examines whether the regression of the criterion onto the predictor space is invariant across subpopulations.

Differential prediction as a group sensitive function. Grades in college are influenced by the student as well as by different teachers who teach different courses at different schools across different universities. Fair prediction or selection requires invariance of prediction equations across groups,

$$R(Y \mid \mathbf{X}, G=1) = R(Y \mid \mathbf{X}, G=2) = ... = R(Y \mid \mathbf{X}, G=g)\,, \tag{1}$$

where R is the symbol for the function used to predict Y, the criterion score, from $\mathbf{X}$, the predictor. G is a variable indicating subgroup membership.

Fair prediction is difficult to achieve, as demonstrated by Young (2001), who cited ample evidence of the underprediction of female grades from high school grades and test scores that occurred when the total group prediction equation was used in place of the female group equation. Is this evidence of unfairness in the predictor or in the criterion? Or is it simply evidence that the use of test scores to predict grades in college is sensitive to the gender of the examinees? Whenever the equality above fails to hold across groups, then invariance is violated and the regression is sensitive to group.

For fair prediction to hold, the prediction model must be the appropriate model. Otherwise misspecification of the model can give the appearance of statistical bias. The prediction model is appropriate if **X** contains all the predictors needed to predict *Y* and the functional form used to combine the predictors is the correct one. For example, grades in college are often predicted from high school grades and test scores, and in some cases, other variables. If high school grades or test scores are dropped as predictors, it is highly unlikely that the regression of college grades onto the remaining predictors will be invariant. In addition to identification of the proper predictors and functional form, the reliability of the criterion itself plays a role. As Linn and Werts (1971) demonstrated in a brief classic on test fairness, replacing a reliable criterion with a less reliable version can result in a lack of invariance of prediction equations in a setting where invariance existed when it came to predicting the more reliable criterion. Linn (1976) in his discussion of the Petersen and Novick analyses noted that the quest to achieve fair prediction is hampered by the fact that the criterion in many studies may itself be unfairly measured.

Even when the correct equation is correctly specified in the full population and the criterion is measured well, invariance may not hold in subpopulations because of selection effects. Linn (1983) described this effect when he talked about predictive bias as an artifact of selection procedures. Linn used a simple case to illustrate his point. He posited that a single predictor X and linear model were needed to predict Y in the full population P. Samples drawn from P depend on a selection variable U that might depend on X in a linear way. Errors in the prediction of Y from X and U from X were also linearly related. Linn then showed that the sample R ($Y|X$, G) equaled the population R (Y|X) if the correlation between X and U were zero, or if errors in prediction of $Y|X$ and $U|X$ were uncorrelated. In other words, the slope of the relationship for predicting U from X must be zero or Y and U must be independent given X.

Achieving subpopulation invariance of regressions is difficult because of selection effects, misspecification errors, and criterion issues. Any attempt to assess whether a prediction equation is invariant across subpopulations such as males and females must keep these confounding influences in mind.

Finally to complicate validity assessment even more, there are as many external criteria as there are uses of a score. Each use implies a criterion against which the test's effectiveness can be assessed. Predictive validation is an unending and endless yet necessary task. Differential prediction studies are even more difficult to complete effectively because, as noted previously, there are so many threats to the subpopulation invariance of regression equations.

2.2 Differential Item Functioning (DIF)

Testing companies have at best indirect influence over how test scores are used. They can advise users, but they can’t prevent the user from using the scores incorrectly. On occasions misuse has been severe enough that companies have refused to sell tests to certain clients. Test companies do, however, have control over the test construction process. To meet societal demands for fair assessments, the companies have looked inward and focused on the basic building blocks of the test, the items.

In the 1970s, the item content review process at ETS was enhanced by the use of written guidelines about minorities and women. The reviews were voluntary and undocumented, however. In 1980, the reviews became mandatory, standardized, and documented with the publication of the *ETS Test Sensitivity Review Process: Guidelines and Procedures* (ETS, 1992). The reviews became the responsibility of all ETS test developers rather than being limited to minority staff. Rigorous training in performing sensitivity reviews and strict adherence to the documented guidelines for writing fair items were required of all test developers.

During the 1980s, the focus in the profession shifted to DIF studies. Although interest in *item bias* studies began in the 1960s (Angoff, 1993), it was not until the 1980s that interest in fair assessment at the item level became widespread. During the 1980s, the measurement profession engaged in the development of a wide array of item level models. DIF procedures developed as part of that shift in attention.

Moving the focus of attention from prediction of external criteria to prediction of item score, which is what DIF is about, represented a major change from a domain where so many factors could spoil the validity effort to a domain where analyses could be conducted in a relatively simple way. While factors such as multidimensionality can complicate a DIF analysis, they are negligible compared to the many influences that can undermine a test fairness study. In a DIF analysis, the item is evaluated against something designed to measure a particular construct and something that the test producer controls, namely a test score.

Differential item functioning (DIF) as a group sensitive function. Differential item functioning asks whether an item is measuring what it purports to measure in much the same way across important subpopulations given the same abilities. For most DIF methods, null DIF can be expressed as

$$\mathrm{E}(Y \mid \mathbf{X}, G=1) = \mathrm{E}(Y \mid \mathbf{X}, G=2) = \ldots. = \mathrm{E}(Y \mid \mathbf{X}, G), \qquad (2)$$

where Y is the item score, $\mathbf{X}$ is the matching variable, typically total score for observed score DIF methods, and E denotes the expectation operator.

Lack of DIF, like lack of differential prediction or the presence of test fairness, requires invariant regressions. In this case, the regression of item scores onto the matching variable needs to be invariant. DIF procedures differ with respect to whether the matching variable is explicitly an observed score (Dorans & Holland, 1993) or a latent variable (Thissen, Steinberg, & Wainer, 1993). They also differ with respect to whether the regression is parametric (Swaminathan & Rogers, 1990) or not (Shealy & Stout, 1993) and to how differences from null DIF are measured (i.e., deltas for the Holland & Thayer [1988] adaptation of the Mantel-Haenszel odds-ratio to DIF versus differences in expected values for standardization [Dorans & Kulick, 1986]). As mentioned in different chapters of the DIF booked edited by Holland and Wainer (1993) there are several pitfalls associated with doing a DIF analysis.

One fair criticism of DIF is that it is difficult to figure out why DIF occurs. With the exception of work completed by Schmitt and her associates (Dorans, Schmitt, & Curley, 1988; Schmitt & Dorans, 1991; Schmitt, Holland & Dorans, 1993) in which hypotheses gleaned from observational data led to experimental evaluations of the hypotheses, most DIF hypotheses are explanations based on a post-hoc study of items that have been selected for DIF evaluation, not predictions that are evaluated empirically.

Another criticism about DIF is that items are unreliable measures of the construct of interest. An item, by itself, can be used to support a variety of hypotheses about DIF. Performance on an item is susceptible to many influences that have little to do with the purpose of the item. In addition, an item on a reliable test is not likely to have much impact on total test performance. There is more to the test than the item.

Another fair criticism is that DIF-freeness is not a prerequisite for fair prediction. Score equity is, however, such a prerequisite. Score equity examines the total score.

2.3 Score Equity Assessment (SEA)

Score equity assessment (SEA) focuses on whether or not scores on different forms that are supposed to be used interchangeably are in fact interchangeable across different subpopulations. SEA uses subpopulation invariance of linking functions across important subpopulations to assess the degree of interchangeability of scores. Since 2000, invariance of linking functions has been studied with the Advanced Placement Program® (AP®), Law School Admissions Test (LSAT), and SAT®, and more recently with state accountability tests.

Subpopulation invariance of equating functions. Test score equating is a statistical process that produces scores considered comparable enough across test forms to be used interchangeably. There are five requirements that are often regarded as basic to all of test equating (Dorans & Holland, 2000). One of the most basic requirements of score equating is that equating functions, to the extent possible, should be subpopulation invariant (Dorans & Holland, 2000; Holland & Dorans, 2006). That is, they should not be strongly influenced by the subpopulation of examinees on which they are computed. The *same construct* and *equal reliability* requirements are prerequisites for subpopulation invariance. One way to demonstrate that two tests are not equatable is to show that the equating functions used to link their scores are not invariant across different subpopulations of examinees. Lack of invariance in a linking function indicates that the differential difficulty of the two tests is not consistent across different groups. The invariance can hold if the relative difficulty changes as a function of score level in the same way across subpopulations. If, however, the relative difficulty of the two tests interacts with group membership or there is an interaction among score level, difficulty and group, then invariance does not hold.

Note that subpopulation invariance is a matter of degree. No acceptable equating function can ever be completely subpopulation invariant, even in the best of circumstances. Instead, in the situations where equating is usually performed, subpopulation invariance implies that the dependence of the equating function on the subpopulation used to compute it is small enough to be ignored.

SEA focuses on whether or not test scores in different forms that are supposed to be used interchangeably are in fact interchangeable across different subpopulations. It uses the subpopulation invariance of linking functions across important subgroups (e.g., gender groups and other groups, sample sizes permitting) to assess the degree of score exchangeability. Compared to DIF analyses, SEA analysis is both less demanding and more relevant to reported scores. DIF analysis assesses whether the function relating item score to total score is invariant across subpopulations. One drawback of DIF analysis is that it tells little about the effects of DIF on reported scores. Another drawback with DIF is that it focuses on the item and ignores the reported score, which is used to make inferences about the examinee. In contrast, SEA focuses on invariance at the reported score level.

Differential prediction analysis is complicated by the many factors associated with a lack of control over the criterion. The quest for answers about differential prediction is endless. In contrast, SEA answers a more tractable question: For what subpopulations is this score interchangeable with scores from other tests built by the same process or with scores from tests meant to measure the same construct?

Previous applications of SEA methodology. Dorans and Holland (2000) included several examples of linkings that are invariant (e.g., SAT Math to SAT Math and SAT Verbal to SAT Verbal, and SAT Math to ACT Math) as well as ones that are not (e.g., Verbal to Math, and linkings between non-Math ACT subscores and SAT Verbal). Equatability indexes are used to quantify the degree to which linkings are subpopulation invariant.

Kolen (2004), in the special issue of *Journal of Education Measurement* on population invariance (Dorans, 2004a), traced the concept of population invariance in equating and linking from the 1950s to the early 2000s. Since 2000, several evaluations of population invariance have been performed. Yang (2004) examined whether the multiple choice to composite linking functions of AP exams remain invariant over subgroups by region. The study focused on two questions: (a) how invariant are cut-scores across regions and (b) whether the small sample size for some regional groups presents particular problems for assessing linking invariance. In addition to using the subpopulation invariance indexes to evaluate linking functions, the author also evaluated the invariance of the composite score thresholds for determining final AP grades. Dorans (2004b) used the population sensitivity of linking functions to assess score equity for two AP exams. Yin, Brennan, and Kolen (2004) looked closely at the issue of invariance of concordance results across subgroups, using concordances between ACT scores and scores on the Iowa Tests of Educational Development. Linear, parallel-linear, and equipercentile methods were used to conduct concordances for males, females, and the combined group. Gender invariance was evaluated both graphically and using group invariance statistics for each linking method. The different linkage methods were evaluated with respect to group invariance.

Dorans, Liu, and Hammonds (2008) used population sensitivity indexes with SAT data to evaluate how consistent linear equating results were across males and females. M. Liu and Holland (2008) examined the population invariance of parallel-linear linkings across different subpopulations of the Law School Admission Test. von Davier and Wilson (2008) examined the population invariance of IRT equating for an AP exam. Yang and Gao (2008) looked at invariance of linking computer-administered College-Level Examination Programs (CLEP$^{®}$) data across gender groups. Yi, Harris, and Gao (2008) examined the invariance of IRT equating across different subpopulations of a science achievement test.

SEA has also been used as a tool to evaluate score interchangeability when a test is revised. Liu, Cahn, and Dorans (2006) and Liu and Walker (2007) used SEA tools to examine the invariance of linkages across the old and new versions of the SAT using data from a major field trail conducted in 2003. This check was followed by SEA analyses conducted on operational data (Dorans, Cahn, Jiang, and Liu, 2006; Dorans, Liu, Cahn, & Jiang, 2006; Jiang, Liu, Cahn, & Dorans, 2006; Liu, Jiang, Dorans, & Cahn, 2006).

All these examples, as well as others such as Dorans, Holland, Thayer, and Tateneni (2003), are illustrations of SEA in which the fairness of a test score exchange process is assessed by the degree to which the linkage between scores is invariant across subpopulations. In some of these illustrations, such as one form of SAT Math with another form of SAT Math, the expectation of score interchangeability was very high since alternate forms of this test are designed to be parallel in both content and difficulty. There are cases, however, where invariance was expected but did not hold. Cook, Eignor, and Taft (1988), for example, found that the linking function between two biology exams depended on whether the equating was with students in a December administration, where most of the examinees were seniors who had not taken a biology course in a while, versus a June administration, where most of the examinees had just completed a biology course. This case, which has become an exemplar of lack of invariance, is discussed in detail by Cook (2007) and Petersen (2007). Invariance cannot be presumed to occur simply because tests are built to the same blueprint. The nature of the population can be critical, especially when diverse populations are involved.

Testing programs that link alternative versions of an exam often spend much time assessing fairness at the item level and presume that it holds at the test level. Granted, individual items contribute to the total test score. Test fairness, however, is being evaluated indirectly at the item level; SEA evaluates it directly. For most testing programs, SEA should confirm the fairness of the assembly process. For the few programs with SEA problems, knowledge of the problem is an essential first step to bring the program in compliance with professional standards.

3. Score Equity Assessment (SEA) Analyses Indexes

3.1 Subgroup Equatings

Each equating between a new form and an old form of a test has two components: a raw-to-raw equating function and a raw-to-scale scaling function. A raw-to-raw equating function $e((X \rightarrow Y)/V)$ is a transformation of raw scores on test *X*, to the scale of raw scores on test *Y* through the anchor test *V* (if applicable; otherwise *V* will be out of the equation). The second step is to convert the equated raw score of *X* to the reporting scale of *Y*, through a scaling function $s(Y)$ that maps the raw scores of *Y* to the scale. The first step of raw-to-raw equating function and the second step of raw-to-scale scaling function are composed to convert the raw scores of *X* onto the reporting scale of *Y* (Holland & Dorans, 2006).

$$s = s(e(X \rightarrow Y | V)) . \tag{3}$$

The subpopulation invariance usually refers to the raw-to-raw equating function. However, the reported or the scaled scores are the final scores that test users get, and most readers are familiar with and can easily interpret scaled score values (e.g., the College Board 200-to-800 scale). Hence, we examine the subpopulation invariance in the scaled score units, which is the concatenated result of the raw-to-raw equating and the raw-to-scale scaling functions.

Equating is usually conducted in the total group to produce a total group equating function. Then a total group scaling function is derived to place raw scores onto the score reporting scale. In a SEA analysis, equating and scaling functions are also produced for each subpopulation of interest. For example, one might use male and female examinees, and where sample sizes are sufficiently large, Asian-American, Black, Hispanic, and White examinees. Ideally, equating holds when the conversion functions are the same across these subpopulations. We use the SAT with its 200-to-800 score reporting scale to make the illustrations concrete. For the SAT, these conversions take raw scores on a new form to unrounded scaled scores on the 200-to-800 scale.

3.2 Difference Plots of Conversions

The difference plot, subgroup conversion minus total group conversion, is the most direct means of assessing population invariance. At each score point level, the subgroup conversion is compared to the total-group conversion.

There has been disagreement about which conversions should be compared. Dorans and Holland (2000) suggested examining the subpopulation linking functions versus the total population linking function, while Brennan (2007) argued that the differences of linking functions should be conducted between pair of subpopulations (e.g., males versus females). In this study, we use the total population linking function as the baseline to compare to in that it is the total population linking function that is used operationally. From a practical perspective, the decision reduces to a choice between the total population equating function and male equating function for male examinees, and a choice between the total population equating function and female equating function for female examinees. While of theoretical interest, it is unlikely that one will use male-equating function with female examinees or female-equating function with male examinees.

3.3 Equatability Indexes

Dorans and Holland (2000) and Dorans et al. (2003) suggested using the standardized root mean square difference (RMSD) to quantify the differences between the subpopulation linking functions and the total-population linking functions at a given score value. They also suggested using the root expected mean square difference (REMSD) to summarize overall differences between the linking functions. These formulas are adapted to comparisons of raw-to-scale functions.

Root mean square difference (RMSD). Let the total population *P* be composed of a set of subpopulation P_g . The two tests to be linked are denoted by *X* (new form) and *Y* (old form). At each *NF* raw score level *x*, the RMSD is defined as

$$RMSD_{(x)} = \sqrt{\sum_g w_g \left[s_{Pg}(x) - s_P(x) \right]^2} \, , \qquad (4)$$

where $s(x)$ represents the composed raw-to-raw equating function and the raw-to-scale scaling function that transformed raw scores of *X* to the reporting score scale of *Y*, and $w_g = \frac{N_g}{N}$ denotes the relative proportion of examinees from total population *P* that are in P_g so that $\sum_g w_g$ = 1. In the present study, the linkings convert the raw scores into scaled scores on the familiar College Board 200 to 800 scale.

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*: the REMSD. The analogue for raw-to-scale scaling functions is

$$REMSD = \sqrt{E_P\left\{\sum_g w_g\left[s_{Pg}(x) - s_P(x)\right]^2\right\}} = \sqrt{\sum_g w_g E_P\left\{\left[s_{Pg}(x) - s_P(x)\right]^2\right\}}\,, \qquad (5)$$

where $\boldsymbol{E_P}\{\cdot\}$ denotes averaging over this distribution, which in this case is the distribution of raw scores on *X* in population *P*.

Root expected square difference (RESD). We also computed the root expected square difference (RESD) statistic, which is

$$RESD_{(g)} = \sqrt{\sum_x f_{gx}\left[s_{Pg}(x) - s_P(x)\right]^2}\,, \qquad (6)$$

to evaluate how close the *g*th subpopulation's raw-to-scale function (s_{Pg}) is to the full population raw-to-scale scaling function. $RESD_{(g)}$ weights by the relative frequency of new form raw scores, f_{gx}, in the subpopulation P_g, and *x* is the index for score level.

Let's briefly summarize the indexes discussed above. $RMSD_{(x)}$ provides an average across groups at each score level. There is only one $RMSD_{(x)}$ across different partitions of *P*. In contrast, $RESD_{(g)}$ provides an average across score levels for each group. There is an $RESD_{(g)}$ for each subgroup. The REMSD is the average of $RMSD_{(x)}$ across score levels.

Difference that matters (DTM). To evaluate the relative magnitude of a difference in score conversions, Dorans and Feigenbaum (1994) proposed the notion of score differences that matter (DTM), in the context of SAT linking. On the SAT scales, scores are reported in 10-point units (200, 210, 220 . . . 780, 790, 800). For example, at a raw score of 53, the corresponding unrounded scaled scores might be 784.3 from the total-group conversion and 785.1 from the White group-only conversion. Due to the vagaries of rounding, the rounded reported scores would be 780 based on the total-group conversion and 790 based on the White group-only conversion when ideally the rounded reported scores should be identical. The DTM, in contrast, treats these two conversions at this raw score point (53) as being equivalent. Dorans et al. (2003) adapted the above indexes, 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 5, which is half of the SAT score unit. Note this difference is best thought of as an indifference threshold. Any differences less than the DTM are considered not big enough to warrant any concern since they are smaller than the smallest difference that might actually matter.

Percentage of scores exceeding differences that matter (DTM). In addition to using RMSD and RESMD, we made use of the percentage of raw scores for which the total and subpopulation raw-to-scale score conversions differed by more than 5 points and of the percentage of examinees for whom these conversions created scaled scores that differed by more than 5 points. The calculation of the two percentage indexes was

$$\begin{aligned} &D_x = 1 \; if \; | S_{Pg}(x) - S_P(x) | >= DTM \\ &\%FS = \frac{\sum_x D_x}{X_{\max} - X_{\min} + 1} \qquad , \qquad (7) \\ &\%Examinee = \sum_x f_g D_x \end{aligned}$$

where the notations have their usual meanings. These two indexes provided straightforward insights into lack of invariance as a percentage of score range and a percentage of test-takers.

Averages and differences in averages. In addition to these indexes, we also compute average scores that were obtained from use of the total group conversion versus average scores that would have been obtained from use of the subgroup conversion, as well as the difference in these average scores. The label *mean diff* is used in the tables to indicate the average difference that would have been obtained had the subgroup conversion been used instead of the total-group conversion.

Spread measures. Deviations based on the total and subgroup conversions were computed as percentages of scores above 700 and below 300, where the two scores were usually regarded as indicators of how many test-takers were located in the top region and the bottom region of the scale.

4. Score Equity Assessment (SEA) Analysis Using SAT Math Data From Year 2005 and 2006

For the purposes of this illustration of how SEA could be implemented operationally, we use data from SAT forms placed on scale in 2005 and 2006. Most of these forms were placed on scale via a nonequivalent groups with anchor test (NEAT) design in which each new form of the test is equated to multiple old forms via external anchors. Other forms were equated through an equivalent groups (EG) design.

Table 1 contains the list of all NEAT design equatings for the SAT Math in 2006. As can be seen in Table 1, a new SAT form goes back to multiple old forms in the case of a typical SAT equating that uses the NEAT design. The raw-to-raw equating and the raw-to-scale scaling functions can be concatenated as following:

$$s_{gijk} = s_{gj}(e_{gijk}((X_i \rightarrow Y_j) \mid V_k)). \tag{8}$$

where *i*, *j* and *k* are indices for the new form, old form, and the anchor test *V*. For the case of a typical total-group equating, for example, the subscript *i* takes on one value only, the subscript *g* takes on one value only (total group), and the subscripts *j* and *k* take on three or four values for each pair of old form and anchor test in the braiding plan. An average across the multiple old form/anchor test pairs defines the operational conversion,

$$s_{gi} = \sum\nolimits_{j=k=1}^{4} w_j s_{gijk} \ , \tag{9}$$

where w_j is the weight assigned to the *jth/kth* old form/anchor test pair.

A total of 21 linkings is depicted for total group. In addition, linkings are also conducted for males and females, so there are 63 links in total. Table 1 also contains samples sizes for total group, males and females.

The purpose of conducting SEA analyses is to detect meaningful violations of score equity. Examination of standard errors in data from previous studies (Dorans, Cahn, et al., 2006; Dorans, Liu, et al., 2006; Jiang et al., 2006; Liu, Jiang, et al., 2006) led the authors to conclude

Table 1

Math Nonequivalent Groups With Anchor Test (NEAT) Linkings and the Sample Sizes in 2006

Linkage	Total		Female		Male	
	NF	OF	NF	OF	NF	OF
ANF1 → AOF1	11,764	7,011	6,328	3,806	5,366	3,205
ANF1 → AOF2	10,690	9,984	5,829	5,361	4,803	4,623
ANF1 → AOF3	11,764	10,211	6,328	5,577	5,366	4,634
ANF2 → AOF1	8,919	6,453	4,705	3,608	4,154	2,845
ANF2 → AOF2	9,795	9,135	5,244	4,934	4,494	4,201
ANF2 → AOF3	8,919	9,456	4,705	5,257	4,154	4,199
ANF3 → AOF1	10,736	7,011	5,922	3,806	4,755	3,205
ANF3 → AOF2	9,709	9,984	5,354	5,361	4,307	4,623
ANF3 → AOF3	10,736	10,211	5,922	5,577	4,755	4,634
ANF4 → AOF1	11,327	15,744	6,347	8,928	4,942	6,755
ANF4 → AOF2	13,751	6,453	7,811	3,608	5,895	2,845
ANF4 → AOF3	12,552	9,135	7,018	4,934	5,494	4,201
ANF4 → AOF4	13,751	9,456	7,811	5,257	5,895	4,199
ANF5 → AOF1	7,030	9,880	3,891	5,395	3,076	4,378
ANF5 → AOF2	5,399	6,453	2,933	3,608	2,413	2,845
ANF5 → AOF3	6,060	9,984	3,419	5,361	2,612	4,623
ANF5 → AOF5	5,399	9,456	2,933	5,257	2,413	4,199
ANF6 → AOF1	5,287	8,425	2,903	4,522	2,346	3,821
ANF6 → AOF2	6,855	6,453	3,734	3,608	3,068	2,845
ANF6 → AOF3	5,867	9,135	3,147	4,934	2,681	4,201
ANF6 → AOF6	6,855	9,456	3,734	5,257	3,068	4,199

Note. A = anchor test design, NF = new form, OF = old form.

that SEA analyses with the NEAT data should be limited to male and female examinees. That is because the sample sizes for Asian, Black, and Latino examinees are typically too small to produce stable results. The instability of the results would be more likely to shed heat not light on the issue. The massive amount of processing would not be worth the effort given the expected sampling instability.

In addition to the NEAT design equating, the SAT also uses the equivalent groups (EG) equating design on a limited basis (note that V_k will be out of Equation 8). The EG design is superior to the NEAT design because the differences in performance are due to differences in tests, not tests and groups. In the SAT case, the EG design also has large enough samples that produce stable results for Asian American, Black, Hispanic, White, and Others groups of examinees as well as for male and female examinees. With the EG design, two forms are equated directly to each other after having been taken by two groups that are presumed to be equivalent as a result of a sampling plan that attempts to achieve stratified random samples. In 2006 this design was used four times. In 2005, it was also used 4 times. The 2006 data were supplemented with the 2005 data, so that we could perform more racial/ethnic SEA analysis.

The equivalent group links for total group and gender groups are shown in Table 2, and the links for ethnic groups are shown in Table 3. Sample sizes are included as well.

In sum, for the six forms equated in 2006 via the NEAT design, a total of 63 equatings (21 equatings × 3 groups) were examined. For the four forms equated in 2006 and the four forms equated in 2005 via the EG design, a total of 64 equatings (8 equatings × 8 groups) were conducted.

5. Results

The results are presented in the following manner: First, we showed the mean difference (mean diff) and RESD figures for all ethnic groups across all EG equatings, identified the best and the worst cases, and examined those two cases in details. Second, a set of similar analyses were conducted for gender groups across all the EG equatings. Third, a set of similar analyses were conducted for gender groups across all the NEAT equatings.

The appendix contains a complete set of detailed data displays for all equatings, including those discussed in the text, for documentation purposes and for the convenience of the reader who wishes to examine the results on a case-by-case basis.

5.1 Ethnic/Racial Results for the Equivalent Groups (EG) Linkings

Figure 1 depicts RESD values plotted against mean difference values for each of the eight equivalent groups equatings for each of the five ethnic/racial subgroups. Each panel of the figure presents the White group paired with one of the other four groups (Asian American, Black, Latino, or Other). Note that all RESD and mean difference values were based on comparing a specific subgroup conversion to the total-group conversion. The White group was included in all of the panels for comparison purposes.

Table 2

Math Equivalent Groups (EG) Linkings and the Sample Sizes in 2005 and 2006: Gender Groups

Linkage	Total		Female		Male	
	NF	OF	NF	OF	NF	OF
ENF1 → EOF1	92,181	188,949	49,591	101,460	42,590	87,489
ENF2 → EOF1	92,094	188,949	49,342	101,460	42,752	87,489
ENF3 → EOF2	94,774	194,643	52,501	107,059	42,273	87,584
ENF4 → EOF2	94,538	194,643	52,036	107,059	42,502	87,584
ENF5 → EOF3	154,278	158,846	83,166	85,354	70,212	72,555
ENF6 → EOF4	193,605	198,094	106,945	109,250	85,509	87,701
ENF7 → EOF5	122,503	126,239	67,244	69,222	54,321	55,999
ENF8 → EOF6	119,962	123,483	63,811	66,140	55,217	56,435

Note. E = equivalent groups design, NF = new form, OF = old form.

Table 3

Math Equivalent Groups (EG) Linkings and the Sample Sizes in 2005 and 2006: Ethnic Groups

Linkage	White		Other		Latino		Black		Asian American	
	NF	OF	NF	OF	NF	OF	NF	OF	NF	OF
ENF1 → EOF1	59,174	121,414	11,208	22,935	7,623	15,302	6,953	14,386	7,223	14,912
ENF2 → EOF1	59,145	121,414	11,177	22,935	7,599	15,302	6,953	14,386	7,220	14,912
ENF3 → EOF2	57,083	116,892	11,250	23,264	8,931	18,382	8,491	17,263	9,019	18,842
ENF4 → EOF2	56,734	116,892	11,200	23,264	8,921	18,382	8,435	17,263	9,248	18,842
ENF5 → EOF3	93,547	95,880	17,922	18,241	13,804	14,422	13,317	14,048	15,688	16,255
ENF6 → EOF4	116,131	118,419	20,680	21,436	20,521	21,179	18,103	18,575	18,170	18,485
ENF7 → EOF5	64,290	66,479	15,113	15,510	15,390	15,895	15,936	16,233	11,774	12,122
ENF8 → EOF6	58,313	60,122	12,912	13,221	16,991	17,454	17,881	18,488	13,865	14,198

Note. E = equivalent groups design, NF = new form, OF = old form.

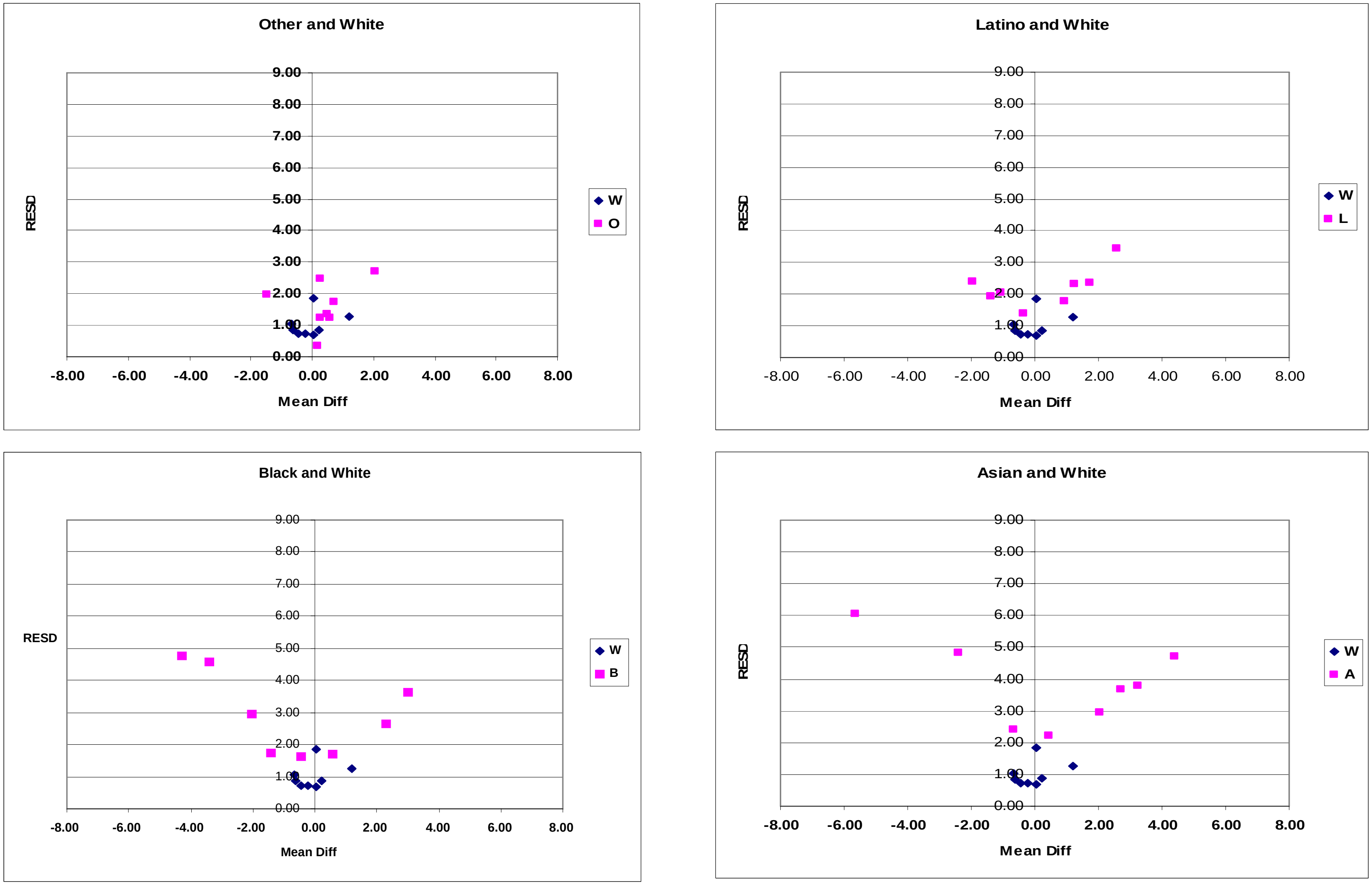


Figure 1. **Math ethnic/racial root expected square difference (RESD) and mean differences (mean diff) for equivalent groups (EG) linkings.**

Figure 1 clearly shows that the Asian American group exhibited the largest degree of group dependence, with some of the mean difference and RESD values either approaching or exceeding 5 (in absolute values). The group dependence shown from the Black group was also nonnegligible, with a couple of mean difference and RESD values close to 5. For White, Other, and Latino groups, most of the mean difference values were packed within a half-circle with a radius of 2. Most of the RESD values were a little larger than the mean difference values but still within a half-circle with a radius of 3. These small values of the mean difference and RESD indicate that subpopulation invariance pretty much holds for these three groups.

The direction of the mean difference is mixed for the White and the Latino groups. For example, half of the time the White examinees would have obtained a lower mean if the White-only conversion had been used instead of the Total group conversion; the other half of the time the white test-takers would have gotten a higher mean if a White-only conversion had been used, with mean difference alternating across the *Y* axis. The mean difference for the Black and the Asian American groups indicates a clearer tendency that the Black group would have obtained a lower mean if the Black-only conversion had been used more than half of the time. The negative mean differences suggest that relative to the total-group, the Black group found the new forms easier than the total group. The Asian American group, on the other hand, would have gotten a higher mean more than half of the time if the Asian American–only conversion had been used instead of the total-group conversion, implying that the Asian American group found the new form harder than the total group.

Among all the forms, we identified Form ENF1 as the worst case in that it exhibited the largest subgroup divergences from the total group. Form ENF5 was identified as the best case in that the subgroup divergence was the smallest.

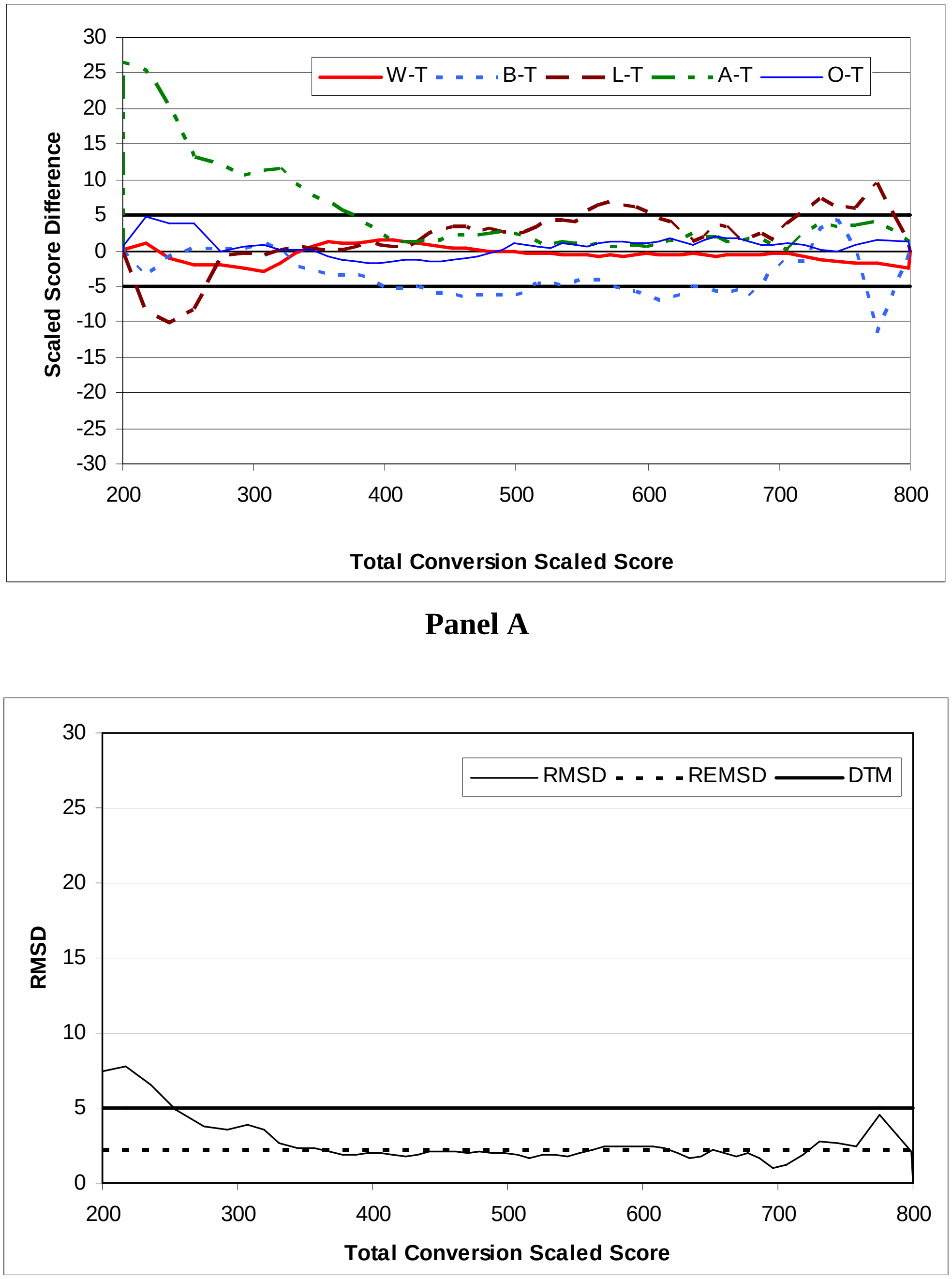
Worst case for ethnic groups: ENF1 to EOF1. Form ENF1 was linked to Form EOF1 through an EG design. The unsmoothed equipercentile linking was conducted for each of the following groups: Total, White, Other, Latino, Black, and Asian American.

Figure 2 contains the graphical summaries of the differences between a subgroup-specific linking and the total-group–specific linking. The difference plots in Panel A suggest that the Black-only conversion differed from the total-group conversion by approximately 5 points across the majority of the scale (between 400 and 700). The Asian American–only conversion fell outside of the ±5 band at the lower end of the scale (below 400), and the Latino-only conversion was lingering around the +5 line starting from the middle part of the scale. The White-only and the Other-only conversions were similar to the total-group conversion over the entire score range. The RMSD curve in Panel B fell below the DTM line except below scaled score 250. The REMSD line of approximately 2 was below the DTM line as well.

Table 4 summarizes the differences between each subgroup conversion and the total group conversion. Table 4 contains means and standard deviations as well as percentage of score above 700 and below 300 for the total group and subgroup conversions. Table 4 also contains the mean difference. In addition, it presents the RESD statistics, the percentage of raw scores for which the total and subgroup conversion differences exceed the DTM of 5, and the percentage of examinees affected by score differences that exceed 5 in absolute value.

In general, the White group would have obtained a slightly lower mean with the White-only conversion, whereas the Other group would have had a slightly higher mean with the Other-only conversion. The Latino and Asian American groups would have obtained a moderately higher mean with the Latino-only conversion and with the Asian American-only conversion, respectively. The Black group, on the other hand, would have had a much lower mean with the Black-only conversion. In terms of the magnitudes of the differences, the Black group showed the largest difference, with a mean difference of -4.3 and a RESD value of 4.8, both of which were close to the DTM threshold. The percentage of formula scores for which scaled scores between the total group conversion and Black-only conversion differed by more than 5 points was close to 40%, nearly half of the raw score points. The percentage of the examinees whose reported scores would have differed by more than 5 points was about 50%. In other words, half of the examinees would have had different reported scores if the Black-only conversion had been used. The percentage indexes associated with the Latino and the Asian American groups were not negligible, either. A different scaled score would have been reported at about 20% of the score levels if the Latino-only conversion or the Asian American-only conversion had been used. Approximately 14% of the Latino test-takers would have obtained a different scaled score if the Latino-only conversion had been placed.

Best case for Math: ENF5 to EOF3. In this particular linkage, the new Form ENF5 was linked to the Form EOF3 through an EG design. The linking used was the unsmoothed equipercentile linking for the Total group and each of the subgroups.



Panel A

Panel B

Figure 2. **ENF1 Math root mean square difference (RMSD) and root expected mean square difference (REMSD): Ethnicity.**

Note. W-T = White-total, B-T = Black-total, L-T = Latino-total, A-T = Asian American-total, O-T = other-total, DTM = difference that matters.

Table 4

ENF1 Math Ethnic Groups Results for the Equivalent Groups (EG) Equatings

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
92,181	T	TGL	530.1	110.2	1.8%	7.3%				
59,174	W	TGL	544.7	101.3	0.6%	7.5%				
		SGL	544.5	100.8	0.9%	7.5%	-0.2	0.7	0.0%	0.0%
11,208	O	TGL	521.1	120.3	3.4%	8.3%				
		SGL	521.4	120.9	3.4%	8.3%	0.3	1.2	0.0%	0.0%
7,623	L	TGL	476.5	102.1	3.4%	2.2%				
		SGL	479.1	103.8	3.4%	2.2%	2.6	3.4	21.2%	14.4%
6,953	B	TGL	440.7	102.6	8.0%	0.9%				
		SGL	436.4	101.3	8.0%	0.6%	-4.3	4.8	37.9%	49.7%
7,223	A	TGL	566.8	116.1	1.1%	15.5%				
		SGL	568.9	115.2	1.1%	15.5%	2.1	3.0	19.7%	4.8%

Note. T = total, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

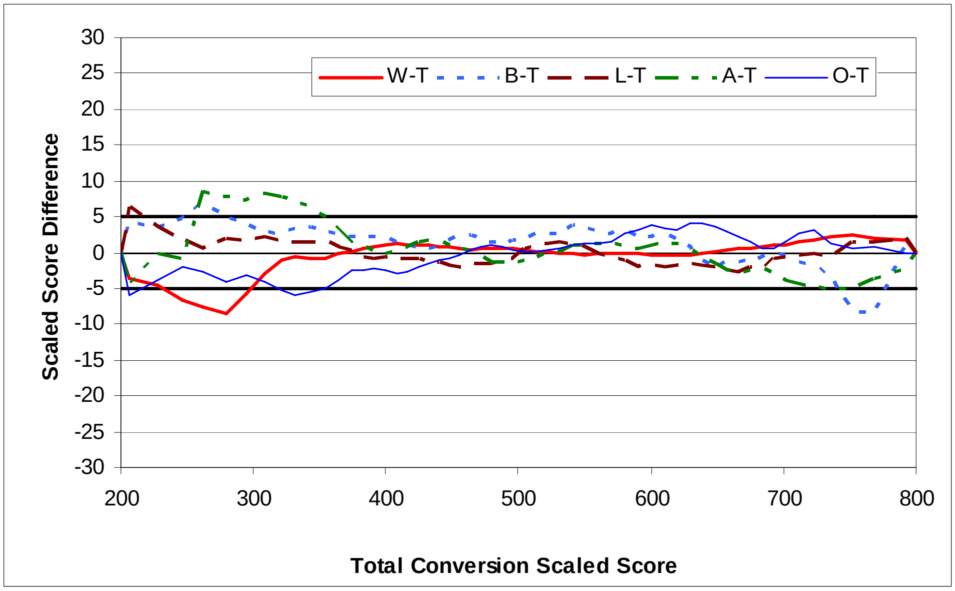
Figure 3 presents the differences between the subgroup-specific conversions and the total group conversion. The differences fell within the DTM band for each subgroup across the majority of the scale, except at the low end where the White-Total comparison had slightly larger difference than -5 and Asian American-Total had slightly larger difference than 5. The RMSD curve, shown in Panel B, was below the DTM curve across most of the score range except at the low end of the scale, where very few test-takers were located. The REMSD line was well below DTM line, with a value less than 2.

Table 5 summarizes the differences in linking results between each comparison. The data reveal that the White and Other groups would have had a slightly higher mean with the White-only conversion and with the Other-only conversion, respectively. On the other hand, the Latino and Asian American groups would have obtained a slightly lower mean with the Latino-only conversion and with the Asian American-only conversion, respectively. The Black group would have had a moderately higher mean with the Black-only conversion. The RESD values were smaller than 1 for the White group, smaller than 2 for the Latino group, and smaller than 3 for the Other, Black and Asian American groups. The two percentage indexes were also small for all the groups with one exception: The percentage of raw scores for which the scaled scores between the subgroup conversion and the Total-group conversion differed by more than 5 points was 12.1% for the Asian American group.

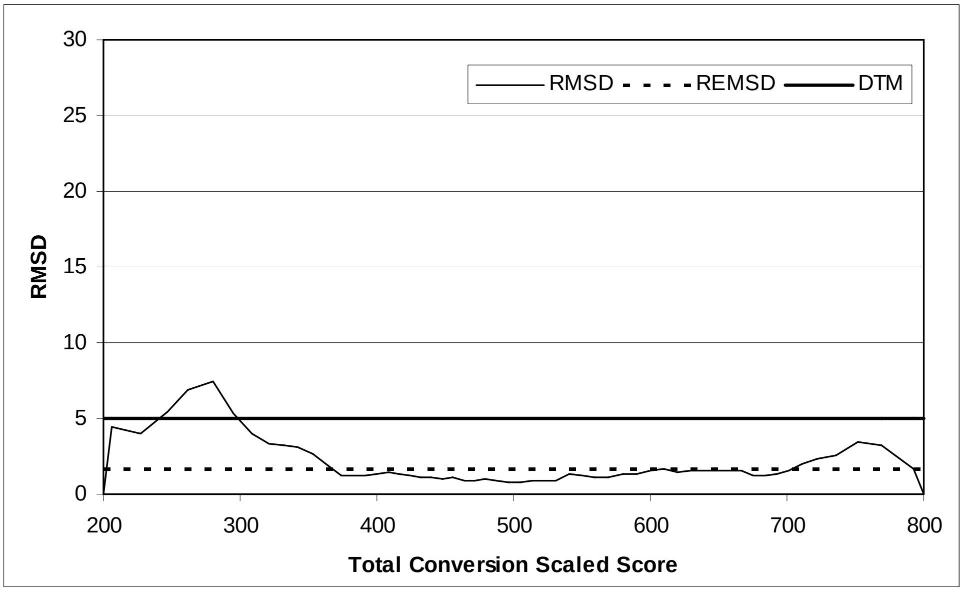
Overall, the SEA analyses across the worst case and the best case suggest that subgroup linking invariance did not seem to hold for the Asian American and Black groups on math. The conversions for the Latino group showed a smaller degree of subgroup sensitivity than those for the Asian American and Black groups. The White group, which is the largest component of the total group, and the Other group subgroup linkings were relatively close to the Total-group linkings.

5.2 Gender Results for the Equivalent Groups (EG) Linkings

Figure 4 depicts RESD values plotted against mean difference values for each of the eight EG equatings and for each of the six NEAT equatings. The filled diamonds and squares represent female and male EG linking results, respectively. All RESD values and all mean difference values for the EG equatings were less than 3; most were under 2. This indicates that population invariance had been obtained on Math across all the EG equatings for both gender groups. The filled diamonds and squares were evenly distributed above and below zero across male and female examinees, meaning female/male examinees sometimes found the new forms easier/harder and sometimes found the new forms harder/easier. On the basis of mean difference and RESD, we selected ENF1 as the worst case and ENF6 as the best case.



Panel A



Panel B

Figure 3. **ENF5 Math root mean square difference (RMSD) and root expected mean square difference (REMSD): Ethnicity.**

Note. W-T = White-total, B-T = Black-total, L-T = Latino-total, A-T = Asian American-total, O-T = other-total, DTM = difference that matters.

Table 5

ENF5 Math Ethnic Groups Results for the Equivalent Groups (EG) Equatings

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean	RESD	% FS	% examinees
							diff		|DIFF| >= 5	|DIFF| >= 5
154,278	T	TGL	532.3	112.6	2.0%	6.8%				
93,547	W	TGL	547.9	100.1	0.7%	6.3%				
		SGL	548.1	100.3	0.7%	6.3%	0.2	0.9	6.1%	0.5%
17,922	O	TGL	518.8	120.1	3.7%	6.7%				
		SGL	519.1	122.1	4.7%	6.7%	0.3	2.5	7.6%	4.9%
13,804	L	TGL	478.2	104.7	3.9%	2.0%				
		SGL	477.8	104.1	3.9%	2.0%	-0.4	1.4	1.5%	0.2%
13,317	B	TGL	426.2	98.7	9.1%	0.6%				
		SGL	428.5	98.1	9.1%	0.6%	2.3	2.6	6.1%	3.2%
15,688	A	TGL	592.3	114.5	0.7%	19.8%				
		SGL	591.6	112.9	0.7%	19.8%	-0.7	2.4	12.1%	3.8%

Note. T = total, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

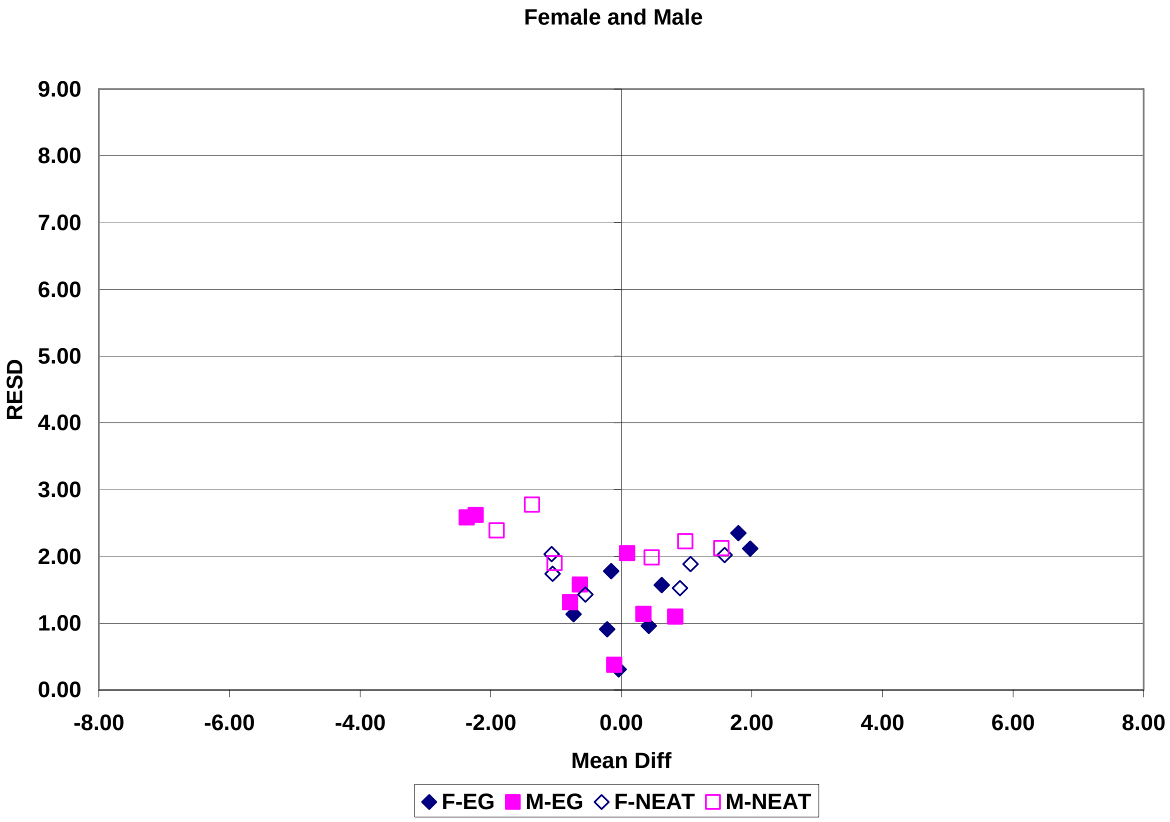


***Figure 4.* Math gender root expected square difference (RESD) and mean difference (mean diff) for equivalent group (EG) linkings and nonequivalent groups with anchor test (NEAT) linkings.**

Note. F = female, M = male.

Worst case for gender groups: ENF1 to EOF1. Form ENF1 was equated to Form EOF1 as indicated in Table 2. The chosen linking function was the equipercentile linking based on unpresmoothed data.

Figure 5 has two panels containing graphical summaries of the differences and similarities between subgroup-specific conversions and the Total group conversion. Panel A contains difference plots for the linking based on male-only and female-only conversions relative to the operational Total group conversion. With the exception of perhaps one score in the mid-700s and a few scores in the mid-200s, the dashed curve falls between the DTM bands, indicating little practical difference between the Total group and the female group conversion.

The solid curve falls between the DTM bands except at in the area below 300, indicating little practical difference between the Total group and the male group conversion. Panel B contains the equatability indexes, RMSD and REMSD. The solid curve is the RMSD as a function of score level and is less than the DTM of 5 throughout most of the score range. It approaches 5 in the mid-700s, and it exceeds the DTM just under a score of 300. The dashed horizontal line is the REMSD value, while the solid horizontal line is the DTM of 5. The REMSD, which is the square root of the average squared difference between an examinee's scores based on the Total group versus subgroup conversions, is close to 2. Even in this worst case, the subgroup conversion works in much the same way as the Total group conversions.

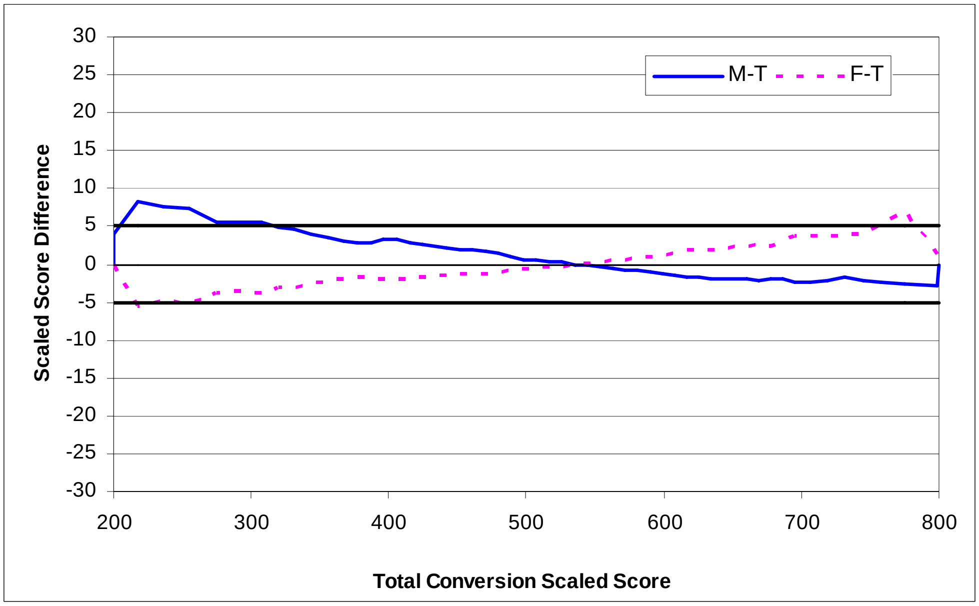
Further support for the inference that the Total group and subgroup conversions are working in a similar fashion can be found in Table 6. The results for males indicate that they would have received essentially the same mean, 548.1, if the male-only conversion had been used in place of the total conversion, which produced a mean of 548.0. For females, the conversion based on female-only would have produced a female average score of 514.6 instead of the mean of 514.7 based on the Total group conversion.

In addition to the mean difference, Table 6 contains the RESD statistics for males and females. Both of these statistics, 2.0 for males and 1.8 for females, are below the DTM of 5, indicating once again that the differences associated with using the operational versus gender specific conversions are small.

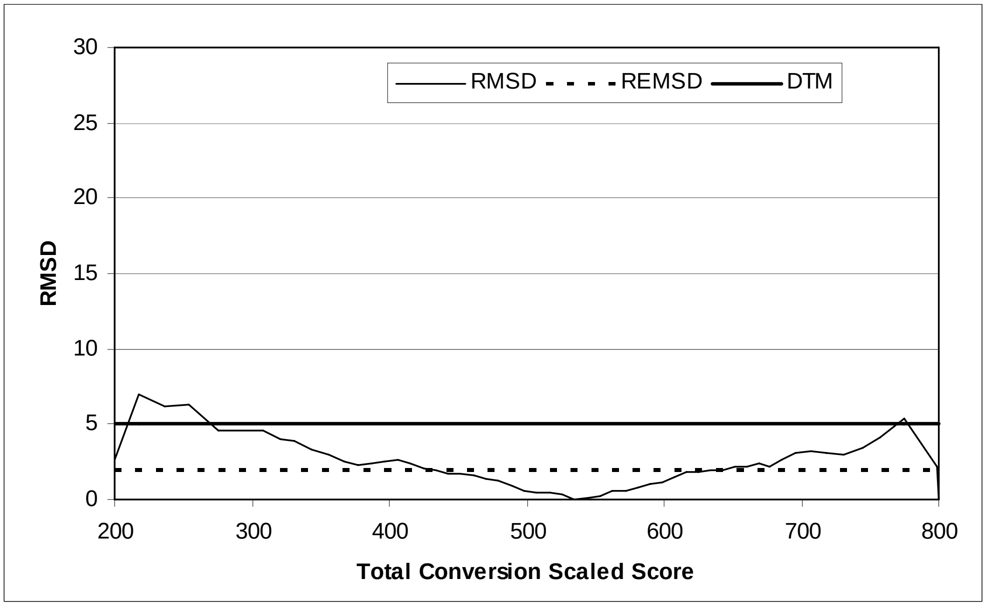
The numbers on the far right of Table 6 summarize the percentage indexes from a practical perspective. Of the formula scores, 9.1% had differences between the male-only and Total group conversions that exceeded 5 in magnitude; these occurred in the low score region. The comparable number for females was 6.1%, mostly in the lowest score region. The percentage of examinees that would have had scores affected by these differences was 2.0% for males and 1.2% for females.

In sum, the SEA analysis on the Math score for Form ENF1 indicates that population invariance holds reasonably well for males and females.

Best case for Math gender groups: ENF6 to EOF4. Form ENF6 was equated back to Form EOF4, as indicated in Table 2. The chosen linking function was the equipercentile linking based on unpresmoothed data.



Panel A



Panel B

Figure 5. **ENF1 Math root mean square difference (RMSD) and root expected mean square difference (REMSD): Gender.**

Note. M-T = male-total, F-T = female-total, DTM = difference that matters.

Table 6

ENF1 Math Gender Groups Results for the Equivalent Groups (EG) Equatings

N	Group	Linking	Mean	SD	% <3 00	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
92,181	T	TGL	530.1	110.2	1.8%	7.3%				
49,591	F	TGL	514.7	104.7	1.9%	4.5%				
		SGL	514.6	106.5	2.6%	4.5%	-0.2	1.8	6.1%	1.2%
42,590	M	TGL	548.0	113.7	1.7%	10.6%				
		SGL	548.1	111.7	1.7%	8.7%	0.1	2.0	9.1%	2.0%

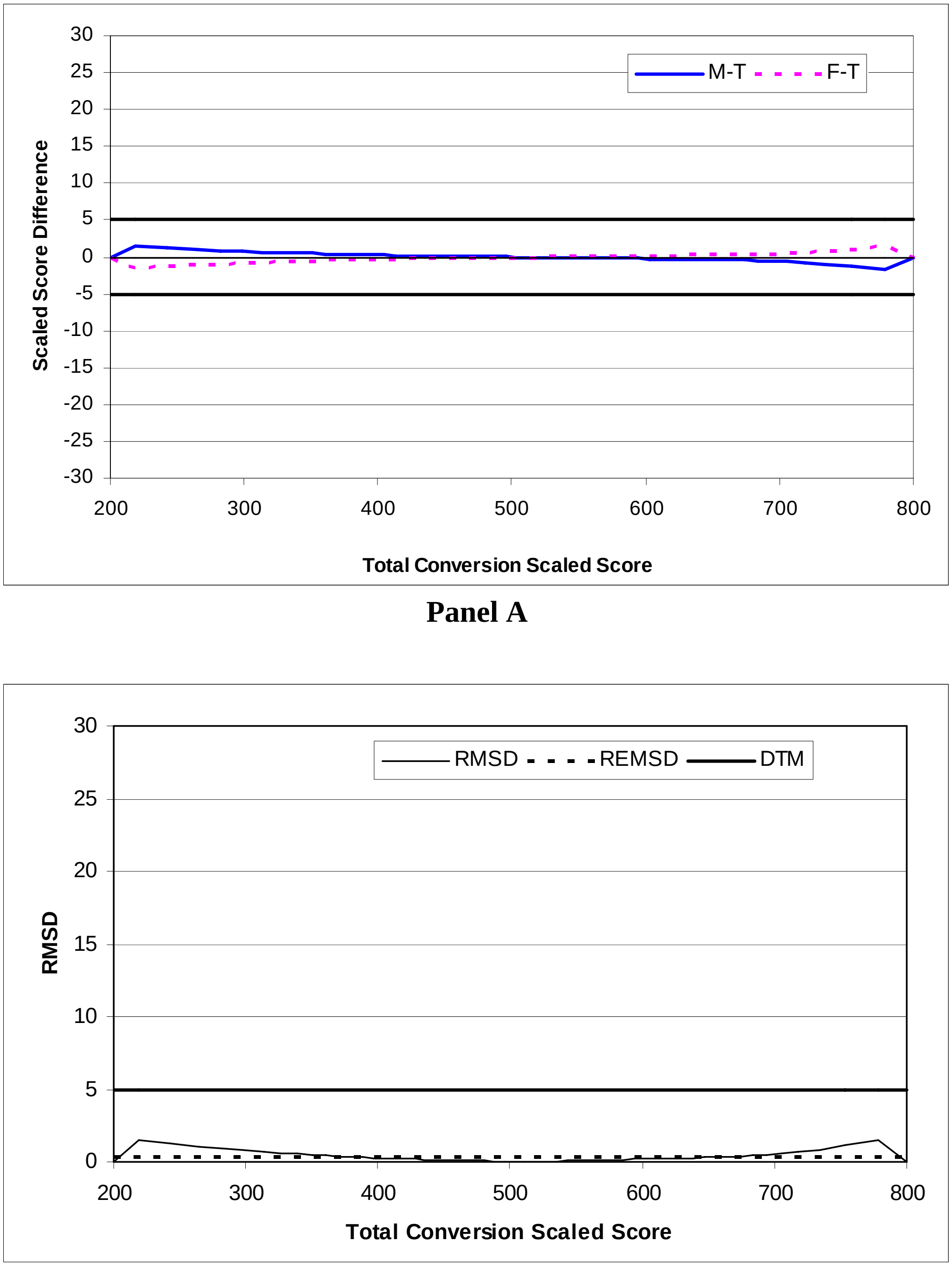
Note: T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Figure 6 has two panels containing graphical summaries of the differences and similarities between the subgroup-specific conversions and the Total group conversion and is similar in format to Figure 5. Both the dashed curve and the solid curve hug the no difference line in the upper panel. Panel B contains the equatability indexes, RMSD and REMSD. The solid curve is the RMSD as a function of score level, the dashed horizontal line is the REMSD value, and the solid horizontal line is the DTM of 5. Both these curves are very close to zero. Clearly, the subgroup conversion works in much the same way as the Total group conversions for Form ENF6.

Further support for the inference that the total group and subgroup conversions are working in a similar fashion can be found in Table 7, which has the same format as Table 6. The results for males indicate that they would have received essentially the same mean, 540.7, if the male-only conversion had been used in place of the total conversion, which produced a mean of 540.9. For females, the conversion based on female-only produced a female average score of 507.4 instead of the mean of 507.5 based on the Total-group conversion.

In addition to the mean differences, which are close to 0, Table 7 contains the RESD statistics for males and females. Both of these statistics, 0.4 for males and 0.3 for females, are close to zero, indicating once again that the differences associated with using the operational versus gender specific conversions are extremely small. Also as can be seen from Table 7, for both male and female examinees, 0.0% of the formula scores had differences between the gender-specific and Total-group conversions that exceeded 5 in magnitude.

In sum, the SEA analysis on the Math score for Form ENF6 indicates that population invariance was essentially achieved.



Panel A

Panel B

Figure 6. **ENF6 Math root mean square difference (RMSD) and root expected mean square difference (REMSD): Gender.**

Note. M-T = male-total, F-T = female-total, DTM = difference that matters.

Table 7

ENF6 Math Gender Groups Results for the Equivalent Groups (EG) Equatings

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF|>=5	% examinees |DIFF|>=5
193,605	T	TGL	521.9	107.6	2.0%	4.9%				
106,945	F	TGL	507.5	103.8	2.2%	3.3%				
		SGL	507.4	104.0	2.2%	3.3%	0.0	0.3	0.0%	0.0%
85,509	M	TGL	540.9	109.2	1.7%	6.9%				
		SGL	540.7	108.8	1.7%	6.9%	-0.1	0.4	0.0%	0.0%

Note: T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

5.3 Gender Results for the Nonequivalent Groups With Anchor Test (NEAT) Linkings

Figure 4 shows that as a set the NEAT equatings are very similar to the EG equatings in terms of being invariant. The hollow diamonds and squares, which represent NEAT linking for females and males, are mixed with the EG results. Again, the mean difference for females and males is evenly distributed above and below zero. Among the NEAT equatings, Form ANF1 has probably the best results, while Form ANF6 may have the worst.

Worst case for Math gender groups: ANF6. Form ANF6 was equated back to test Forms AOF1, AOF2, AOF3, and AOF6 through unique external anchor tests. The last row in Table 2 contains this set of linkings and the sample sizes. Each link was evaluated, and a conversion was selected. Operationally, the final equating decision for ANF6 involved a weighted average of the four equating functions. For the female and male subgroup linkings, the linking method used was identical to that used for the total group. In addition, the weight function used to average the four links was the same in both the subgroup and the total-group linkings.

Figure 7 has two panels containing graphical summaries of the differences and similarities for the average scaling functions and is identical in format to the figures seen with the EG design. With the exception of scores below 300, the dashed curve falls between the DTM bands. The solid curve falls between the DTM for scores above 350. Panel B contains the equatability indexes, RMSD and REMSD. The solid curve is the RMSD as a function of score level. It is less than the DTM of 5 for most scores above 350. The dashed horizontal line is the REMSD value, while the solid horizontal line is the DTM of 5. The REMSD is near 2.

Table 8 summarizes the differences between each subgroup conversion and the Total group conversion. The results for male examinees indicate that they would have received a mean of 508.3 if the male-only conversion had been used in place of the total conversion, which produced a mean of 507.3. For females, the female-only conversion produced a female average score of 475.9 instead of the mean of 477.0 based on the Total group conversion.

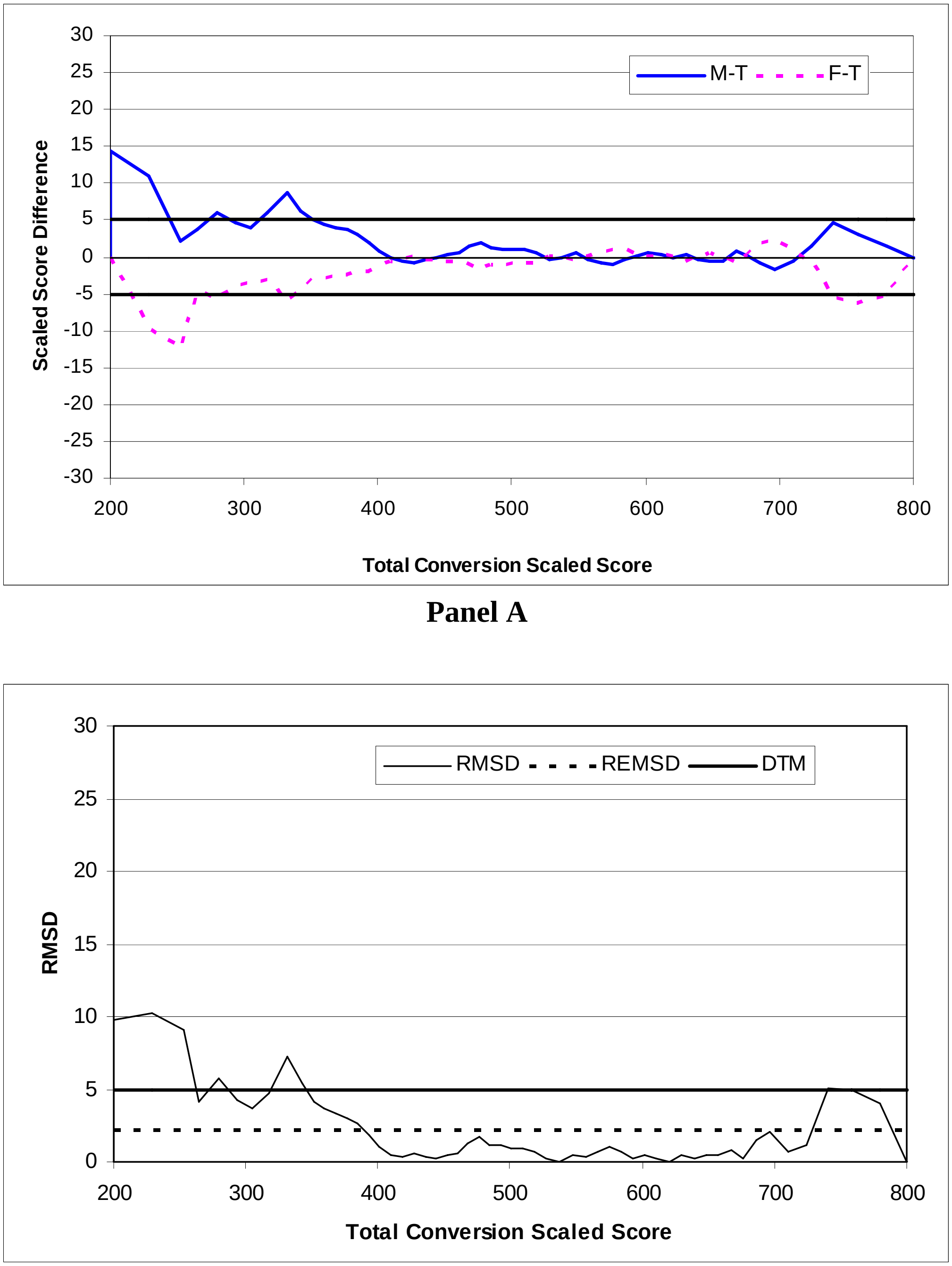
In addition to mean differences of about +/-1 point, Table 8 contains the RESD statistics for males and females. Both of these statistics, 2.2 for male examinees and 2.0 for female examinees, are below the DTM of 5, indicating that the differences associated with using the operational versus gender specific conversions are small.

The percentage indexes show that 10.6% of the formula scores had differences between the male-only and the total-group conversions that exceeded 5 in magnitude; these occurred mostly below 350. The same percentage was obtained for female examinees, mostly in the lower score region. The percentage of examinees that would have had been affected by these differences was 5.6% for male examinees and 3.8% for female examinees.

In sum, the SEA analysis on the Math score for Form ANF6 indicates that population invariance holds reasonably well for males and females but not as well as it did for the equatings based on the EG design.

Best case for gender groups: ANF1. Form ANF1 was equated to test Forms AOF1, AOF2, and AOF3 through unique external anchor tests. The top row in Table 1 contains the three ANF1 linkings and the sample sizes. Each link was evaluated, and a conversion was selected. Operationally, the final equating decision for ANF1 involved a weighted average of the three equating functions. For the female and male subgroup linkings, the linking method used for a link was identical to that used for the Total group. In addition the weight function use to average the three links was the same in both subgroup linkings and the Total group linking.

Figure 8 includes two panels containing graphical summaries of the differences and similarities for the average scaling functions. The dashed curve (female) is very close to the no difference line, while the solid curve (male) is also close except at the very bottom of the scale. Panel B contains the equatability indexes, RMSD and REMSD. The solid RMSD curve and the



Panel A

Panel B

Figure 7. **ANF6 Math root mean square difference (RMSD) and root expected mean square difference (REMSD): Gender.**

Note. M-T = male-total, F-T = female-total, DTM = difference that matters.

Table 8

ANF6 Math (Average) Gender Groups Results for the Nonequivalent Groups With Anchor Test (NEAT) Equatings

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
123,483	T	TGL	490.5	108.8	3.0%	3.9%				
66,140	F	TGL	477.0	103.9	3.3%	2.6%				
		SGL	475.9	104.9	4.5%	2.6%	-1.1	2.0	10.6%	3.8%
56,435	M	TGL	507.3	112.0	2.6%	5.5%				
		SGL	508.3	110.9	2.6%	4.2%	1.0	2.2	10.6%	5.6%

Note: T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

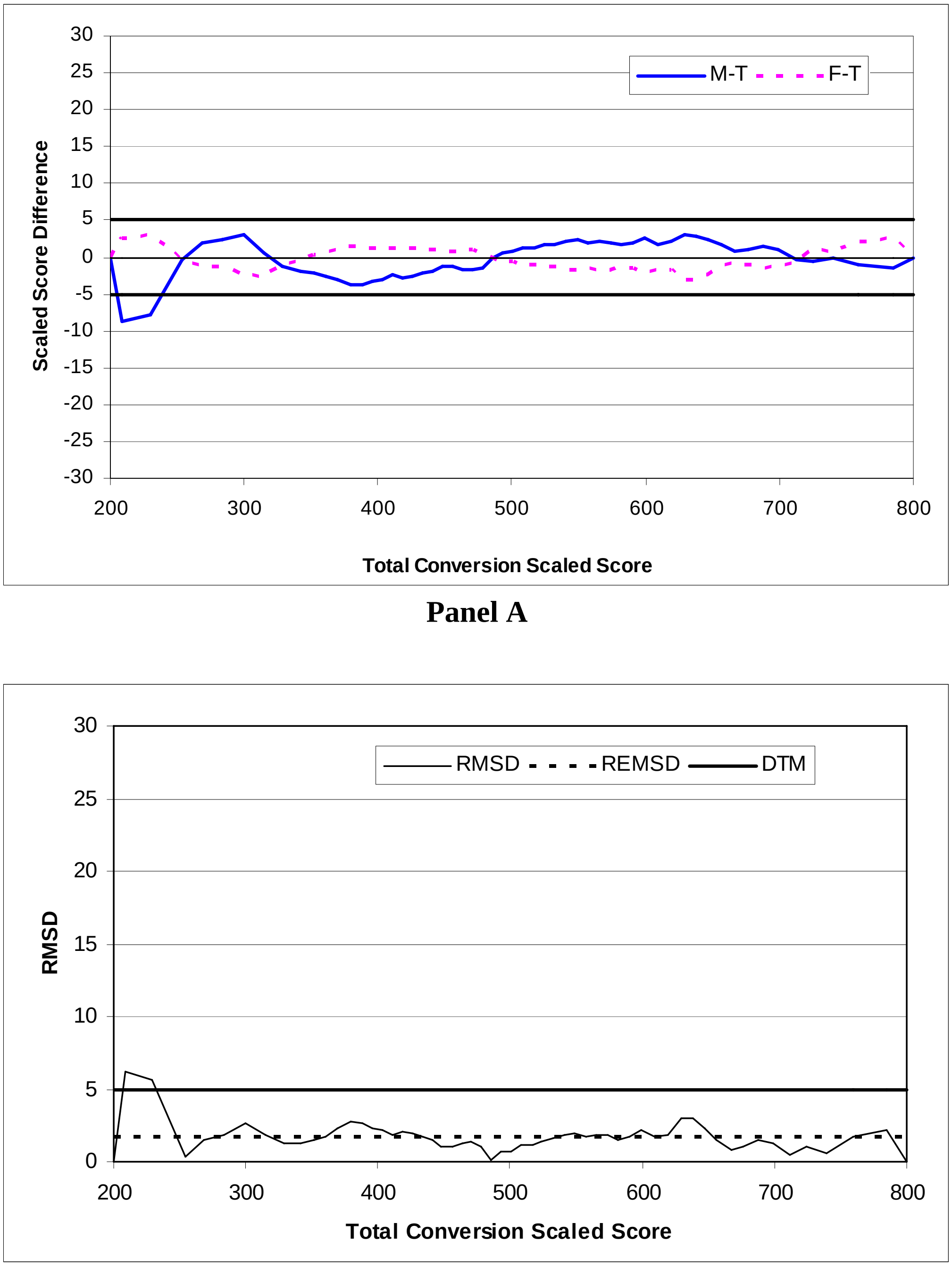
dashed horizontal REMSD line tend to be less than 2, between 250 and 800. Clearly, the Total group conversion works in much the same way as the subgroup conversions for Form ANF1.

Further support for the inference that the Total group and subgroup conversions are working in similar fashion can be found in Table 9, which has the same format as Table 8. The results for males indicate that they would have received essentially the same mean, 549.5, if the male-only conversion had been used in place of the total conversion, which produced a mean of 549.0. For females, the conversion based on female-only produced a female average score of 519.1 instead of the mean of 519.7 based on Total group conversion.

In addition to the mean differences, which are +/-.5, Table 9 contains the RESD statistics for males and females. Both of these statistics, 2.0 for males and 1.4 for females, are small, indicating that the differences associated with using the operational versus gender specific conversions are small.

The percentage indexes reveal that 0.0% of the formula scores had differences between the female-only and total group conversions that exceeded 5 in magnitude; for males 3% of the scores differ, but only .4% of the examinees would have had been affected, all at the low end.

In sum, the SEA analysis on the Math score for Form ANF1 indicates that population invariance was essentially achieved.



Panel A

Panel B

Figure 8. **ANF1 Math root mean square difference (RMSD) and root expected mean square difference (REMSD):Gender.**

Note. M-T = male-total, F-T = female-total, DTM = difference that matters.

Table 9

ANF1 Math (Average) Gender Groups Results for the Nonequivalent Groups With Anchor Test (NEAT) Equatings

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
158,846	T	TGL	532.5	112.4	2.4%	7.5%				
85,354	F	TGL	519.7	108.3	2.5%	5.3%				
		SGL	519.1	107.7	2.5%	5.3%	-0.5	1.4	0.0%	0.0%
72,555	M	TGL	549.0	114.4	2.1%	10.3%				
		SGL	549.5	115.4	2.1%	10.3%	0.5	2.0	3.0%	0.4%

Note: T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

6. Summary

This study investigated how to implement SEA as a routine operational practice for the SAT. We used SAT forms that were administered in 2005 and 2006. SEA was conducted on the Total group, the gender groups, and the ethnic groups on the forms that were placed on scale through the EG linking design. For the forms that were placed on scale through nonequivalent groups anchor test design linking, SEA was conducted on the Total group and gender groups.

The results were mixed. For the ethnic groups, the Black group and the Asian American group exhibited different degrees of subgroup linking sensitivity, ranging from negligible differences from the Total group linking, to a noticeable degree of subgroup linking dependence, and then to a substantial degree of subgroup linking deviation from the Total group. The results for the White and other groups, in contrast, were relatively stable, as was the case for the Latino group in all but one linking. In general, the subgroup linkages exhibited negligible subgroup sensitivity for these three groups.

As mentioned above, we observed a *direction switch* of subgroup linkage on the Black group and on the Asian American group. In the spring 2005 administrations, the Asian American group would have obtained a higher mean if the Asian American-only conversion had been used, implying that the Asian American group found the forms harder than the Total group. The direction of the mean differences went in the opposite direction in 2006, when the Asian American group would have received a lower mean (with one exception in December 2006 administration) if the Asian American-only conversion had been used, suggesting that the Asian American group now found the forms easier than the Total group. The Black group, on the other hand, exhibited an opposite pattern from the Asian American group: The group would have had a lower mean with the Black-only conversion in 2005, but would have received a higher mean in 2006 with the Black-only conversion (except December 2006 administration). Has the construct evolved for these two groups of test-takers? Or was it due to sampling variability? These questions are worth further consideration.

The results for the gender groups showed that population invariance was achieved across the gender groups for all forms regardless of data collection design. The degree of the subgroup linking sensitivity was negligible.

7. Discussion

The equating process links scores from different editions of the same test. For testing programs that build nearly parallel forms to the same explicit content and statistical specifications and administer forms under the same conditions, the linkings between the forms are expected to be equatings.

SEA analysis focuses directly on the statistical end product of the test development and scoring process—the score to be reported, and examines whether or not scores that are supposed to be used interchangeably are in fact interchangeable. The relationship among equating, SEA and subpopulation invariance can be illustrated as follows: equating ensures score comparability and interchangeability. SEA provides a useful tool to check score equatability. For a linking to be an equating, the assumption of subpopulation invariance must be met. Checking SEA via subpopulation invariance of equating functions could serve as a quality control check to ensure that well developed test assemblies remain within acceptable tolerance levels with respect to equatability (Deming, 1982, 1986).

How should SEA analyses be incorporated into statistical and psychometric practice? For example, when should an SEA analysis be conducted and on what groups? SEA analysis is performed on test level data and cannot be performed at the pretest stage. We advise against performing SEA prior to score reporting with the intent of ascertaining whether to use subgroup-specific equatings. First, the sample sizes are often small at this stage. Even a test like the SAT, which uses the NEAT design to place forms on scale, only has adequate sample sizes for males and females. Second, accommodating SEA checks is often impractical because of time considerations. Finally, aberrant results might be expected to occur with a single test form in any testing program. It is far more important to focus on systemic problems rather than idiosyncratic results.

The authors suggest conducting SEA across a number of administrations and forms, such as at the end of one testing year, or at the end of a cohort year, or after some other suitable interval. The examination of multiple forms should provide a better assessment of the fairness of the test assembly process. Unusual results will be more clearly seen as unusual in the context of SEA analyses across multiple forms.

The number of subgroups that can be studied will depend on circumstances such as test volume and how the data are collected. We expect that most testing programs should be able to examine whether linkings are invariant across gender.

Another related issue is sample size. SEA procedures should be performed on sample sizes that are adequate enough to detect meaningful effect sizes. Implementation of this principle requires an effect size, such as the minimal DTM and standard errors of equating or the standard errors of the difference of equating. The standard error formulae can guide us in determining which samples sizes are too small to support meaningful SEA analyses. For large samples, the effect size will be easier to interpret.

If it turns out linking results are not invariant, and the differences are consistent and large enough to have a practical impact on scores, then due diligence suggests that the test assembly, test administration, and statistical analysis processes should be scrutinized for possible causes.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brennan, R. L. (2007). Tests in transition: Synthesis and discussion. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). New York: Springer-Verlag.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Cook, L. L. (2007). Equating test scores: Practical problems, a practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73-88). New York: Springer-Verlag.

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

Deming, W. E. (1982, 1986). *Out of the crises.* Cambridge, MA: Massachusetts Institute of Technology Center for Advanced Engineering Study.

Dorans, N. J. (2004a). Editor's introduction: Assessing the population sensitivity of equating functions. *Journal of Educational Measurement, 41*(1), 1–2.

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

Dorans, N. J., Cahn, M., Jiang, Y., & Liu, J. (2006). *Score equity assessment of transition from SAT I Math to SAT Math: Gender* (ETS Statistical Rep. No. SR-06-63). Princeton, NJ: ETS.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

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

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,* 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 Research Rep. No. RR-03-27, pp. 79–118). Princeton, NJ: ETS.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., & Liu, J. (2007, April). *Are reported scores the same when test content and conditions of measurement change?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Dorans, N. J., Liu, J., Cahn, M., & Jiang, Y. (2006). *Score equity assessment of transition from SAT I Verbal to SAT Critical Reading: Gender* (ETS Statistical Rep. No. SR-06-61). Princeton, NJ: ETS.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97.

Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, March). Differential speededness: Some items have DIF because of where they are, not what they are. Paper presented in *ETS symposium: Advances in differential item functioning research with the SAT.* Princeton, NJ: ETS.

ETS. (1992). *ETS test sensitivity review process: Guidelines and procedures.* Princeton, NJ: Author.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jiang, Y., Liu, J., Cahn, M., & Dorans, N. J. (2006). *Score equity assessment of transition from SAT I Verbal to SAT Critical Reading: Ethnicity* (ETS Statistical Rep. No. SR-06-62). Princeton, NJ: ETS.

Kolen, M. J. (2004). Population invariance in equating: Concept and history. *Journal of Educational Measurement, 41*(1), 3–14.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43, 139–161.*

Linn, R. L. (1975). Test bias and the prediction of grades in law school. *Journal of Legal Education, 27,* 293–323.

Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13,* 53–58.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Liu, J., Cahn, M., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linking of new SAT to old SAT across gender groups. *Journal of Educational Measurement, 43,* 113–129.

Liu, J., Jiang, Y., Dorans, N. J., & Cahn, M. (2006). *Score equity assessment of transition from SAT I Math to SAT Math: Ethnicity* (ETS Statistical Rep. No. SR-06-64). Princeton, NJ: ETS.

Liu, J. & Walker, M. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109-134). New York: Springer-Verlag.

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three LSAT test administrations. *Applied Psychological Measurement, 32*(1), 27–44.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York: Springer-Verlag.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models of culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Schmitt, A. P., & Dorans, N. J. (1991). Factors related to differential item functioning for Latino examinees on the Scholastic Aptitude Test. In G. D. Keller, J. D. Deneen, & R. J. Magallan (Eds.), *Assessment and access* (pp. 105–132). Albany, NY: State University of New York Press.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating of hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (1971). Concepts of culture-fair selection. *Journal of Educational Measurement, 8,* 63–70.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of IRT true score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11–26.

Yang, W. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*(1), 33–41.

Yang, W., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based CLEP® examination. *Applied Psychological Measurement, 32*(1), 45–61.

Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32*(1), 62–80.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, *28*(4), 274–289.

Young, J. W., (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Rep. No. 2001-06). New York: The College Board.

Appendix

Table A1

ENF1 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
92,181	T	TGL	530.1	110.2	1.8%	7.3%				
49,591	F	TGL	514.7	104.7	1.9%	4.5%				
		SGL	514.6	106.5	2.6%	4.5%	-0.2	1.8	6.1%	1.2%
42,590	M	TGL	548.0	113.7	1.7%	10.6%				
		SGL	548.1	111.7	1.7%	8.7%	0.1	2.0	9.1%	2.0%
59,174	W	TGL	544.7	101.3	0.6%	7.5%				
		SGL	544.5	100.8	0.9%	7.5%	-0.2	0.7	0.0%	0.0%
11,208	O	TGL	521.1	120.3	3.4%	8.3%				
		SGL	521.4	120.9	3.4%	8.3%	0.3	1.2	0.0%	0.0%
7,623	L	TGL	476.5	102.1	3.4%	2.2%				
		SGL	479.1	103.8	3.4%	2.2%	2.6	3.4	21.2%	14.4%
6,953	B	TGL	440.7	102.6	8.0%	0.9%				
		SGL	436.4	101.3	8.0%	0.6%	-4.3	4.8	37.9%	49.7%
7,223	A	TGL	566.8	116.1	1.1%	15.5%				
		SGL	568.9	115.2	1.1%	15.5%	2.1	3.0	19.7%	4.8%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A2

ENF2 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
92,094	T	TGL	530.1	110.2	2.0%	6.6%				
49,342	F	TGL	514.0	105.1	2.2%	4.0%				
		SGL	514.6	106.5	2.2%	4.0%	0.6	1.6	0.0%	0.0%
42,752	M	TGL	548.7	113.0	1.8%	9.6%				
		SGL	548.1	111.7	1.8%	9.6%	-0.6	1.6	1.5%	0.2%
59,145	W	TGL	545.1	101.1	0.7%	6.9%				
		SGL	544.5	100.8	0.7%	6.9%	-0.6	0.9	0.0%	0.0%
11,177	O	TGL	520.9	120.2	3.7%	7.5%				
		SGL	521.4	120.9	3.7%	7.5%	0.6	1.2	3.0%	0.7%
7,599	L	TGL	477.8	105.2	4.2%	2.1%				
		SGL	479.1	103.8	4.2%	2.1%	1.2	2.3	7.6%	2.6%
6,953	B	TGL	438.5	100.9	8.6%	0.7%				
		SGL	436.4	101.3	8.6%	0.7%	-2.0	2.9	9.1%	3.3%
7,220	A	TGL	564.5	114.5	1.0%	13.6%				
		SGL	568.9	115.2	1.0%	13.6%	4.4	4.7	27.3%	39.2%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A3

ENF3 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
94,774	T	TGL	519.6	107.4	2.3%	4.6%				
52,501	F	TGL	506.2	103.1	2.5%	2.9%				
		SGL	506.6	103.9	2.5%	2.9%	0.4	1.0	0.0%	0.0%
42,273	M	TGL	536.3	110.3	2.1%	6.6%				
		SGL	535.5	109.5	2.1%	6.6%	-0.8	1.3	0.0%	0.0%
57,083	W	TGL	535.0	97.8	1.0%	4.4%				
		SGL	534.6	97.4	1.0%	4.4%	-0.5	0.7	1.5%	0.1%
11,250	O	TGL	504.2	116.3	4.5%	4.6%				
		SGL	504.9	115.3	3.5%	4.6%	0.7	1.7	1.5%	1.0%
8,931	L	TGL	474.9	100.6	4.1%	1.5%				
		SGL	473.6	99.6	4.1%	1.5%	-1.4	1.9	4.5%	0.9%
8,491	B	TGL	436.6	96.4	8.2%	0.5%				
		SGL	436.1	97.5	8.2%	0.5%	-0.4	1.6	1.5%	0.0%
9,019	A	TGL	563.7	113.0	1.1%	12.4%				
		SGL	566.5	114.8	1.1%	14.7%	2.7	3.7	16.7%	9.1%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A4

ENF4 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
94,538	T	TGL	519.6	107.4	2.3%	5.3%				
52,036	F	TGL	504.8	102.4	2.5%	3.2%				
		SGL	506.6	103.9	2.5%	3.2%	1.8	2.4	1.5%	0.2%
42,502	M	TGL	537.7	110.5	2.0%	7.7%				
		SGL	535.5	109.5	2.0%	6.3%	-2.2	2.6	0.0%	0.0%
56,734	W	TGL	535.2	98.0	0.9%	5.3%				
		SGL	534.5	97.4	0.9%	4.1%	-0.7	1.0	3.0%	0.5%
11,200	O	TGL	502.8	116.1	4.6%	5.0%				
		SGL	504.9	115.3	3.3%	5.0%	2.1	2.7	12.1%	10.8%
8,921	L	TGL	471.8	100.0	4.2%	1.3%				
		SGL	473.6	99.6	4.2%	1.3%	1.8	2.4	6.1%	0.8%
8,435	B	TGL	439.5	95.6	7.5%	0.4%				
		SGL	436.1	97.5	7.5%	0.7%	-3.4	4.6	36.4%	48.0%
9,248	A	TGL	563.3	113.6	1.2%	13.7%				
		SGL	566.5	114.8	1.2%	13.7%	3.3	3.8	10.6%	15.9%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A5

ENF5 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
154,278	T	TGL	532.3	112.6	2.0%	6.8%				
83,166	F	TGL	517.4	108.5	2.3%	4.6%				
		SGL	519.4	108.5	2.3%	4.6%	2.0	2.1	0.0%	0.0%
70,212	M	TGL	551.2	114.0	1.6%	9.6%				
		SGL	548.8	114.7	2.2%	9.6%	-2.4	2.6	3.0%	0.5%
93,547	W	TGL	547.9	100.1	0.7%	6.3%				
		SGL	548.1	100.3	0.7%	6.3%	0.2	0.9	6.1%	0.5%
17,922	O	TGL	518.8	120.1	3.7%	6.7%				
		SGL	519.1	122.1	4.7%	6.7%	0.3	2.5	7.6%	4.9%
13,804	L	TGL	478.2	104.7	3.9%	2.0%				
		SGL	477.8	104.1	3.9%	2.0%	-0.4	1.4	1.5%	0.2%
13,317	B	TGL	426.2	98.7	9.1%	0.6%				
		SGL	428.5	98.1	9.1%	0.6%	2.3	2.6	6.1%	3.2%
15,688	A	TGL	592.3	114.5	0.7%	19.8%				
		SGL	591.6	112.9	0.7%	19.8%	-0.7	2.4	12.1%	3.8%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A6

ENF6 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
193,605	T	TGL	521.9	107.6	2.0%	4.9%				
106,945	F	TGL	507.5	103.8	2.2%	3.3%				
		SGL	507.4	104.0	2.2%	3.3%	0.0	0.3	0.0%	0.0%
85,509	M	TGL	540.9	109.2	1.7%	6.9%				
		SGL	540.7	108.8	1.7%	6.9%	-0.1	0.4	0.0%	0.0%
116,131	W	TGL	537.9	98.4	0.8%	4.9%				
		SGL	539.1	98.1	0.8%	6.3%	1.2	1.3	1.5%	0.1%
20,680	O	TGL	510.9	113.9	3.3%	4.6%				
		SGL	511.1	113.7	3.3%	4.6%	0.2	0.3	0.0%	0.0%
20,521	L	TGL	474.6	101.4	4.0%	1.5%				
		SGL	472.6	102.5	4.0%	1.5%	-2.0	2.4	9.1%	5.2%
18,103	B	TGL	439.3	94.9	7.2%	0.4%				
		SGL	439.9	96.4	7.2%	0.7%	0.6	1.7	9.1%	0.6%
18,170	A	TGL	567.7	114.1	1.0%	13.2%				
		SGL	562.0	113.2	1.3%	13.2%	-5.7	6.1	40.9%	38.2%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A7

ENF7 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
122,503	T	TGL	504.1	108.2	3.1%	3.8%				
67,244	F	TGL	488.7	104.3	3.6%	2.4%				
		SGL	488.0	103.6	3.6%	2.4%	-0.7	1.1	3.0%	0.8%
54,321	M	TGL	524.3	109.3	2.3%	5.5%				
		SGL	525.1	109.8	2.3%	5.5%	0.8	1.1	0.0%	0.0%
64,290	W	TGL	526.6	95.5	0.9%	3.6%				
		SGL	526.6	97.2	0.9%	3.6%	0.0	1.8	12.1%	1.8%
15,113	O	TGL	486.5	115.4	5.7%	3.5%				
		SGL	485.0	115.4	5.7%	3.5%	-1.5	2.0	3.0%	1.5%
15,390	L	TGL	459.6	97.8	5.4%	0.9%				
		SGL	458.5	96.9	5.4%	0.9%	-1.1	2.0	6.1%	2.3%
15,936	B	TGL	429.7	93.4	8.5%	0.4%				
		SGL	432.7	93.2	8.5%	0.5%	3.0	3.6	13.6%	3.7%
11,774	A	TGL	562.4	118.0	1.4%	13.8%				
		SGL	560.0	114.7	1.0%	13.8%	-2.4	4.8	40.9%	45.0%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A8

ENF8 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
119,962	T	TGL	490.1	109.0	3.6%	3.4%				
63,811	F	TGL	476.8	104.8	4.0%	2.3%				
		SGL	476.5	104.0	4.0%	2.3%	-0.2	0.9	0.0%	0.0%
55,217	M	TGL	506.6	111.3	2.9%	4.7%				
		SGL	507.0	112.2	2.9%	4.7%	0.3	1.1	0.0%	0.0%
58,313	W	TGL	511.8	96.5	1.1%	2.9%				
		SGL	511.9	97.0	1.1%	2.9%	0.1	0.7	0.0%	0.0%
12,912	O	TGL	473.9	112.6	5.7%	3.0%				
		SGL	474.4	112.3	5.7%	3.0%	0.5	1.4	4.5%	1.4%
16,991	L	TGL	447.1	95.5	6.1%	0.7%				
		SGL	448.0	94.9	6.1%	0.9%	0.9	1.8	9.1%	1.7%
17,881	B	TGL	417.5	88.6	9.4%	0.2%				
		SGL	416.1	88.3	9.4%	0.2%	-1.4	1.7	15.2%	3.0%
13,865	A	TGL	559.9	120.1	1.4%	13.1%				
		SGL	560.3	118.9	1.4%	13.1%	0.4	2.2	9.1%	5.2%

Note. T = total, F = female, M = male, W = White, O = other, L = Latino, B = Black, A = Asian American, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A9

ANF11 Math (Average)

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
158,846	T	TGL	532.5	112.4	2.4%	7.5%				
85,354	F	TGL	519.7	108.3	2.5%	5.3%				
		SGL	519.1	107.7	2.5%	5.3%	-0.5	1.4	0.0%	0.0%
72,555	M	TGL	549.0	114.4	2.1%	10.3%				
		SGL	549.5	115.4	2.1%	10.3%	0.5	2.0	3.0%	0.4%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A10

ANF2 Math (Average)

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
179,539	T	TGL	513.2	105.4	2.1%	4.3%				
94,872	F	TGL	496.8	100.7	2.4%	2.6%				
		SGL	497.7	100.7	2.4%	2.6%	0.9	1.5	6.1%	1.2%
83,493	M	TGL	533.0	106.8	1.7%	6.4%				
		SGL	532.0	107.3	2.3%	6.4%	-1.0	1.9	7.6%	1.5%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A11

ANF3 Math (Average)

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
198,094	T	TGL	521.0	106.3	2.2%	5.5%				
109,250	F	TGL	506.6	102.9	2.7%	3.7%				
		SGL	507.1	102.0	1.9%	3.7%	0.5	1.5	4.5%	0.9%
87,701	M	TGL	539.9	107.3	1.6%	7.8%				
		SGL	539.2	108.7	1.6%	7.8%	-0.7	2.3	16.7%	7.1%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A12

ANF4 Math (Average)

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
502,848	T	TGL	528.3	105.6	1.9%	5.3%				
284,027	F	TGL	513.8	101.9	2.2%	3.5%				
		SGL	512.8	101.0	1.5%	3.5%	-1.1	1.7	6.1%	1.8%
217,094	M	TGL	548.0	107.1	1.5%	7.6%				
		SGL	549.5	107.8	1.1%	7.6%	1.5	2.1	7.6%	4.7%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A13

ANF5 Math (Average)

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
126,239	T	TGL	504.5	108.2	3.7%	4.3%				
69,222	F	TGL	488.4	103.6	4.4%	2.7%				
		SGL	490.0	103.8	3.1%	2.7%	1.6	2.0	4.5%	1.0%
55,999	M	TGL	525.5	109.9	2.8%	6.4%				
		SGL	523.6	109.9	2.8%	5.1%	-1.9	2.4	6.1%	1.7%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

Table A14

ANF6 Math

N	Group	Linking	Mean	SD	% < 300	% > 700	Mean diff	RESD	% FS |DIFF| >= 5	% examinees |DIFF| >= 5
123,483	T	TGL	490.5	108.8	3.0%	3.9%				
66,140	F	TGL	477.0	103.9	3.3%	2.6%				
		SGL	475.9	104.9	4.5%	2.6%	-1.1	2.0	10.6%	3.8%
56,435	M	TGL	507.3	112.0	2.6%	5.5%				
		SGL	508.3	110.9	2.6%	4.2%	1.0	2.2	10.6%	5.6%

Note. T = total, F = female, M = male, mean diff = mean difference, RESD = root expected square difference, TGL = total-group linking, SGL = subgroup linking, FS = formula score.

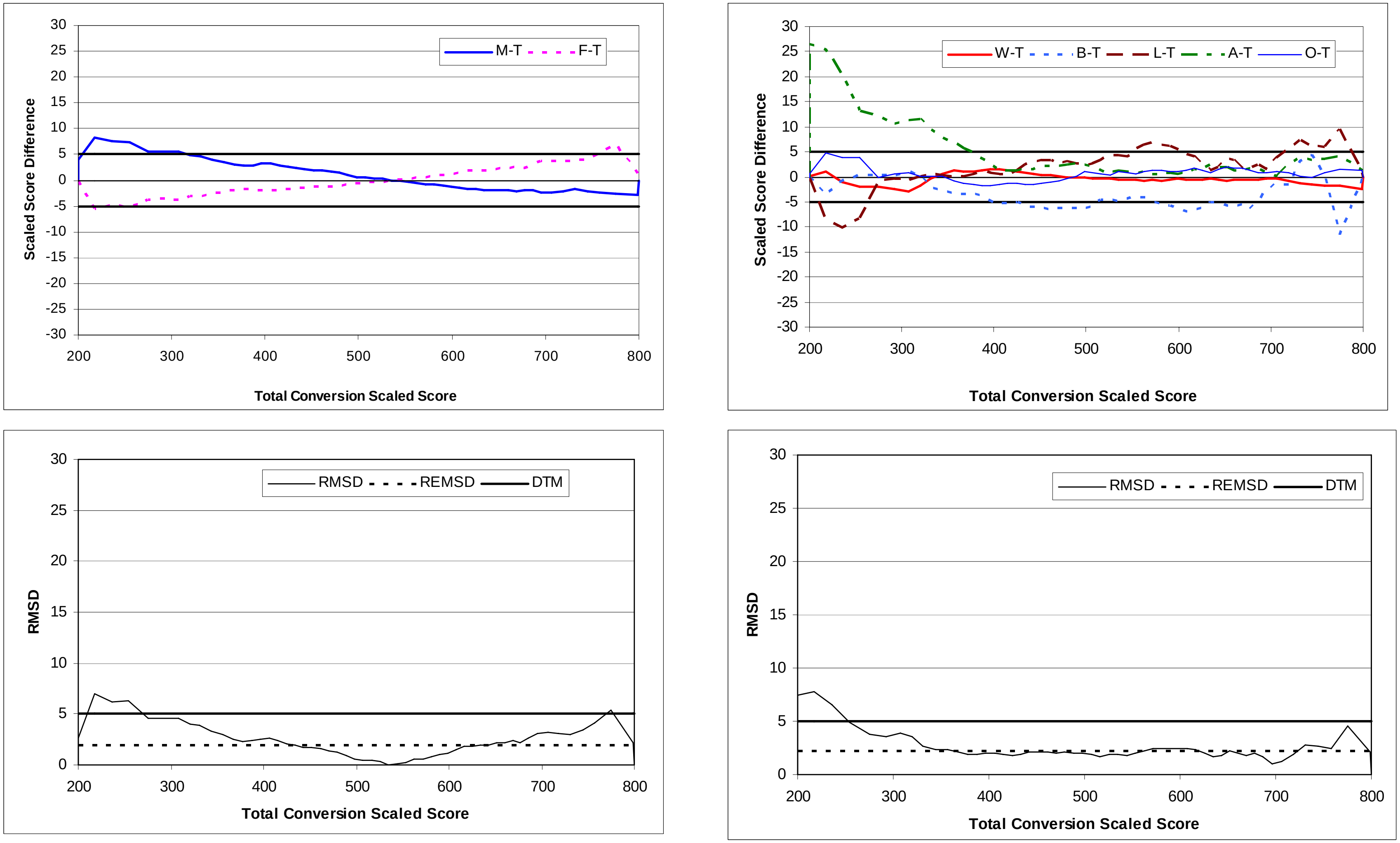


Figure A1. **ENF1 Math: Gender and ethnicity.**

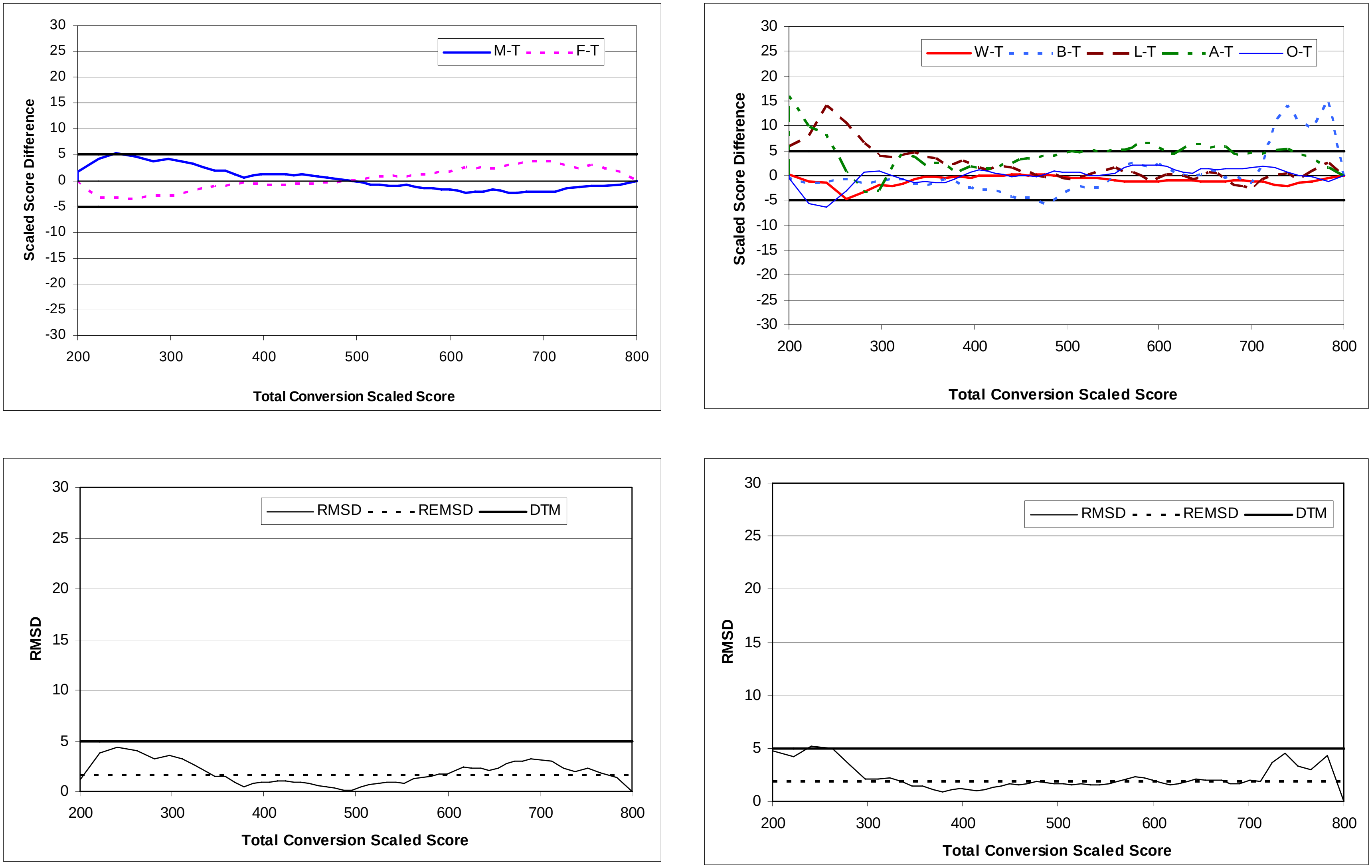


Figure A2. **ENF2 Math: Gender and ethnicity.**

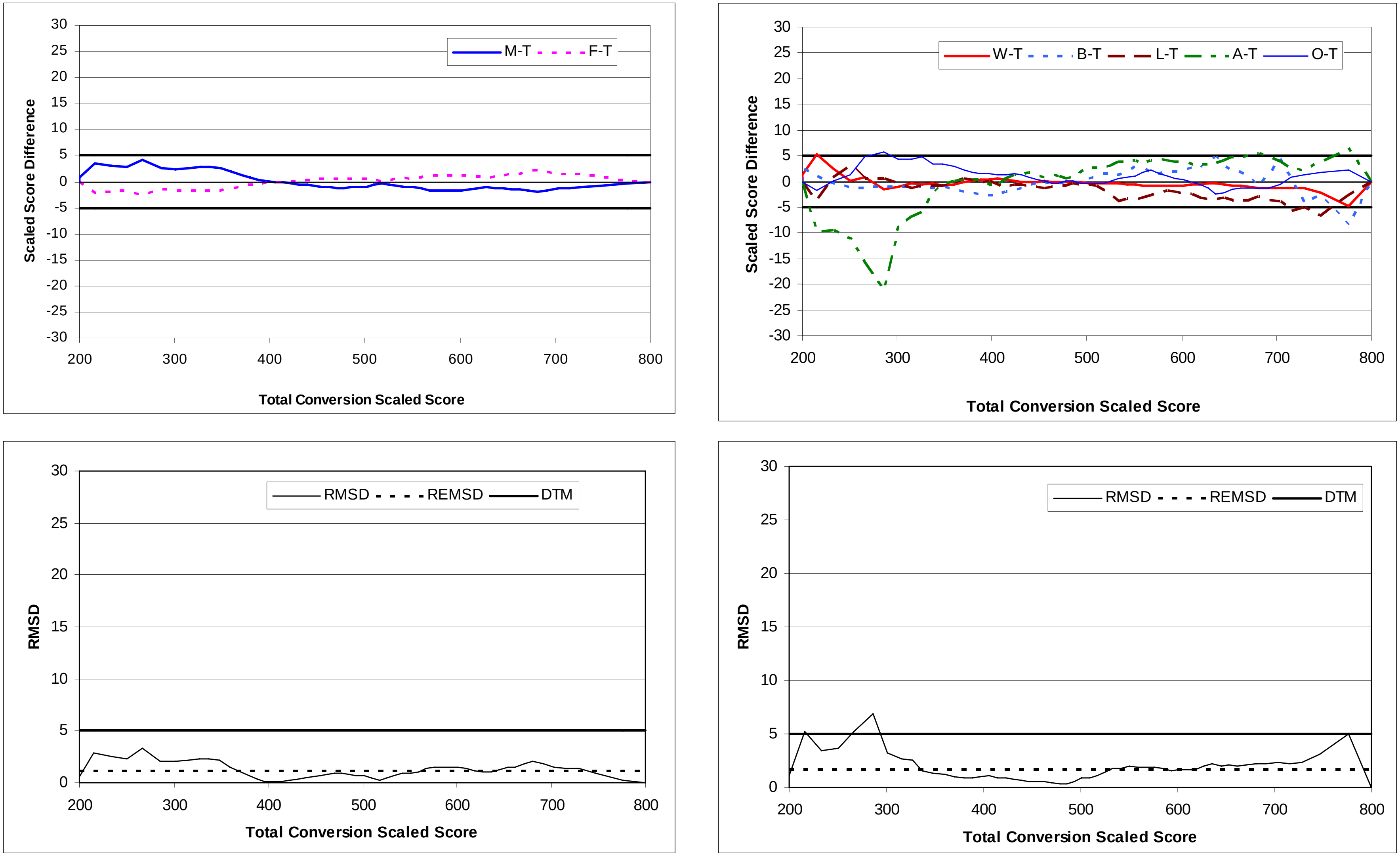


Figure A3. **ENF3 Math: Gender and ethnicity.**

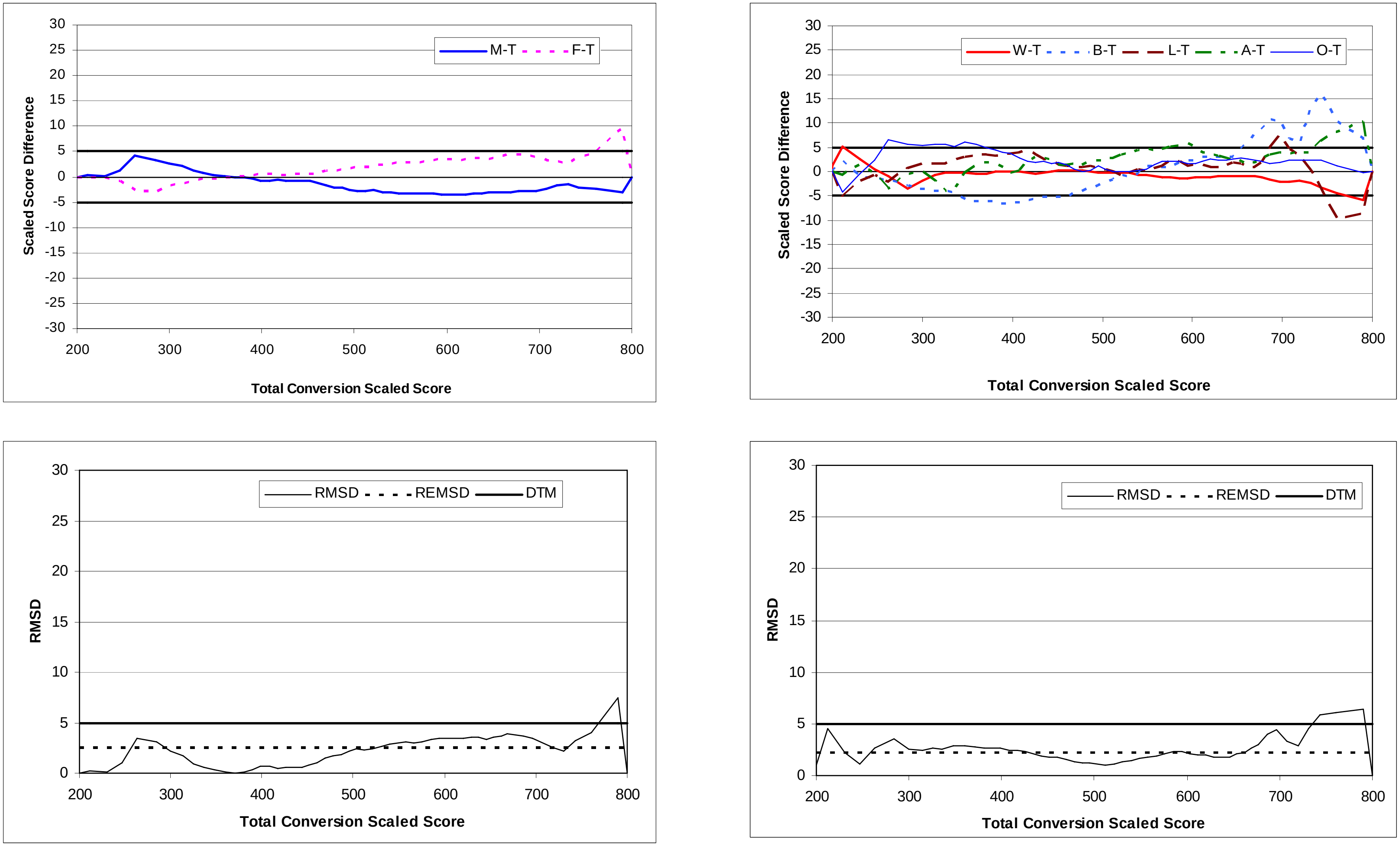


Figure A4. **ENF4 Math: Gender and ethnicity.**

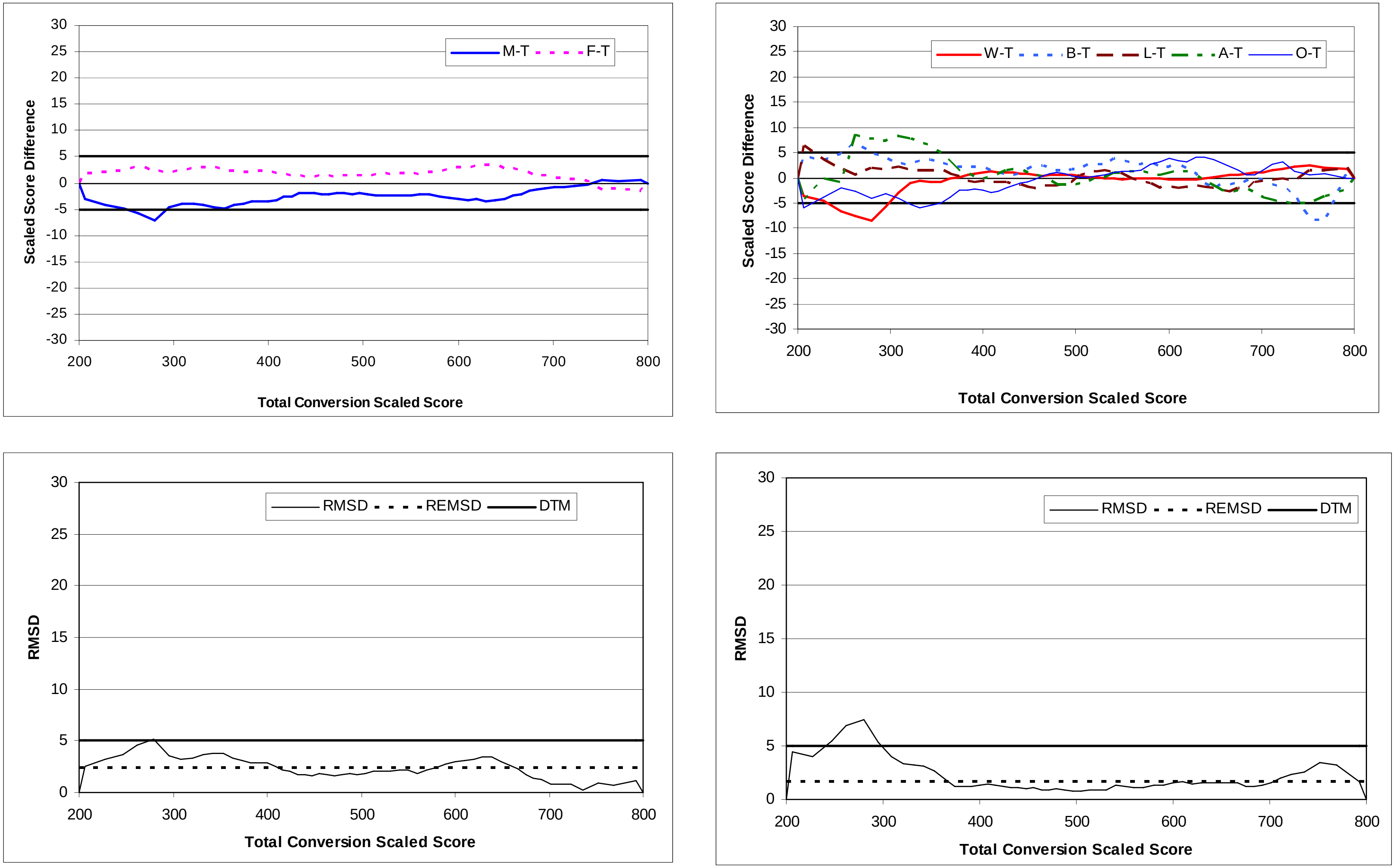


Figure A5. **ENF5 Math: Gender and ethnicity.**

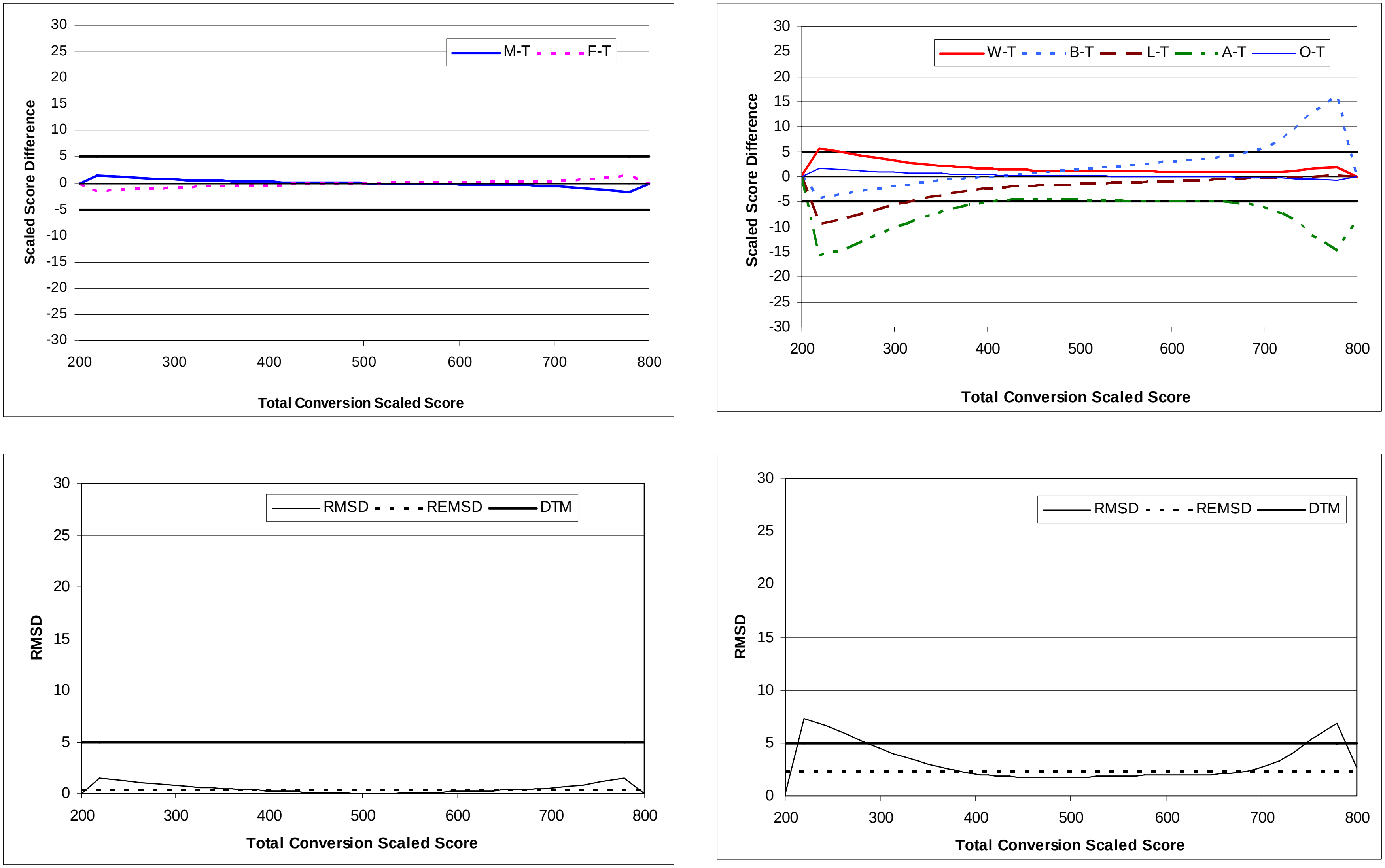


Figure A6. **ENF6 Math: Gender and ethnicity.**

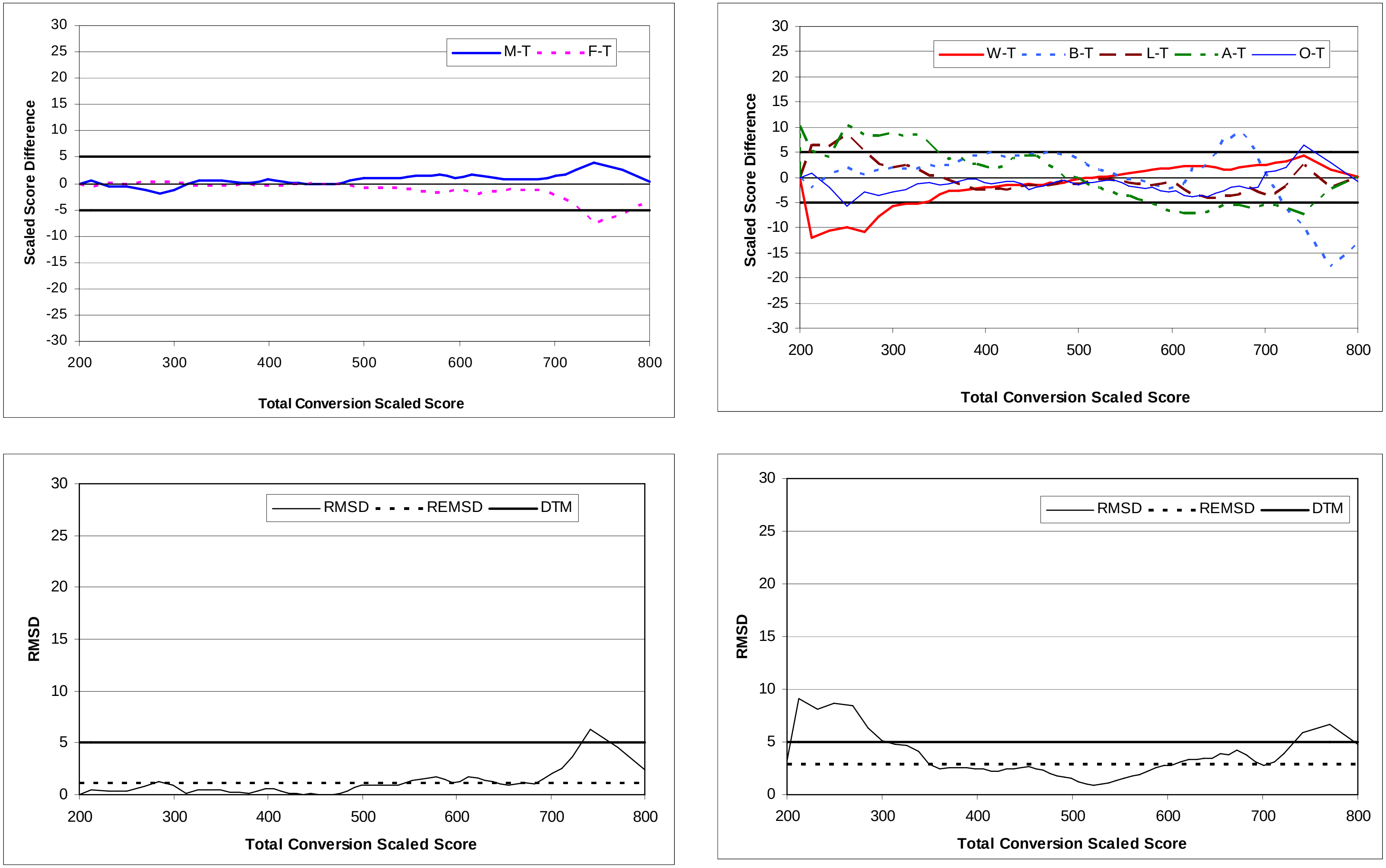


Figure A7. **ENF7 Math: Gender and ethnicity.**

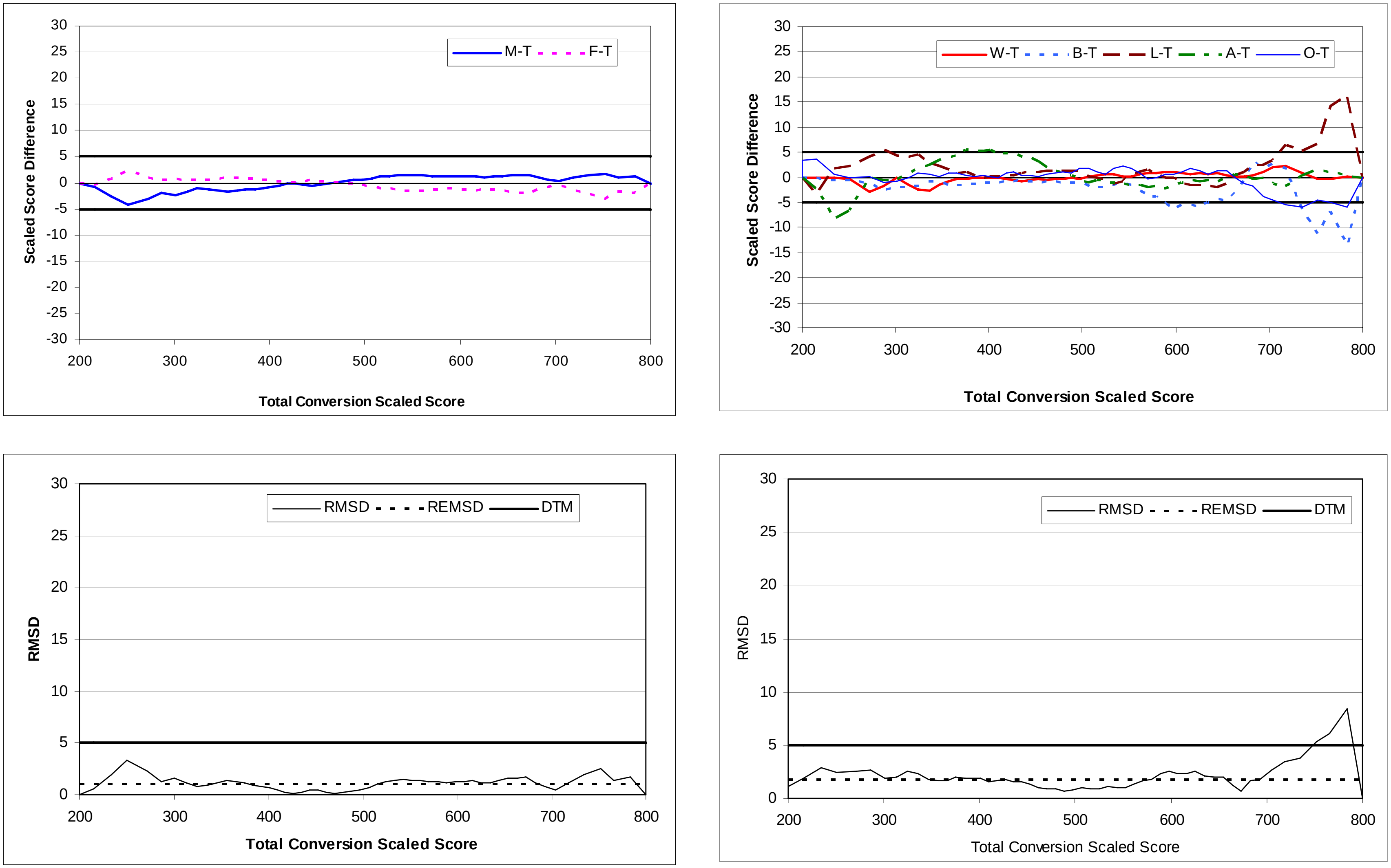
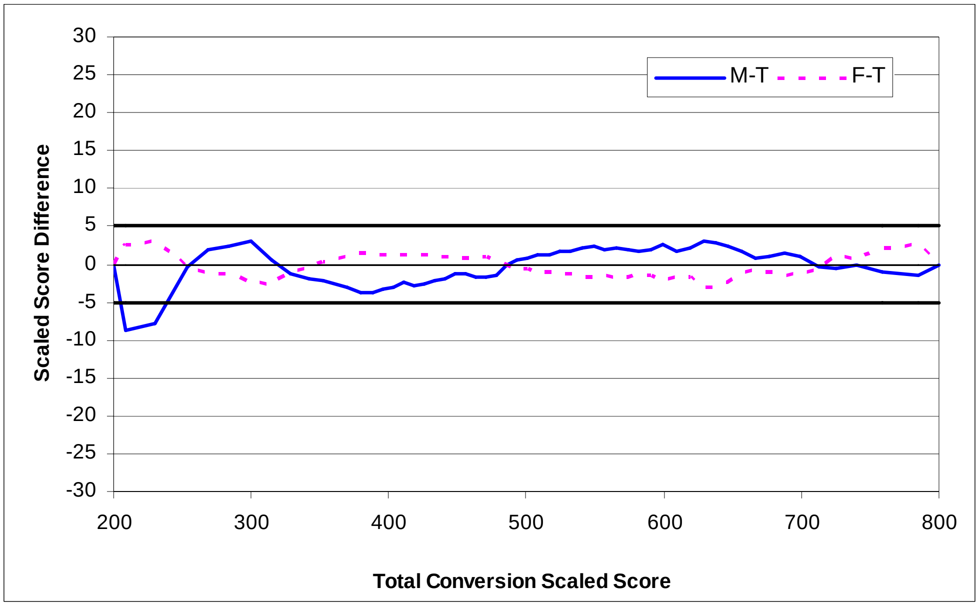
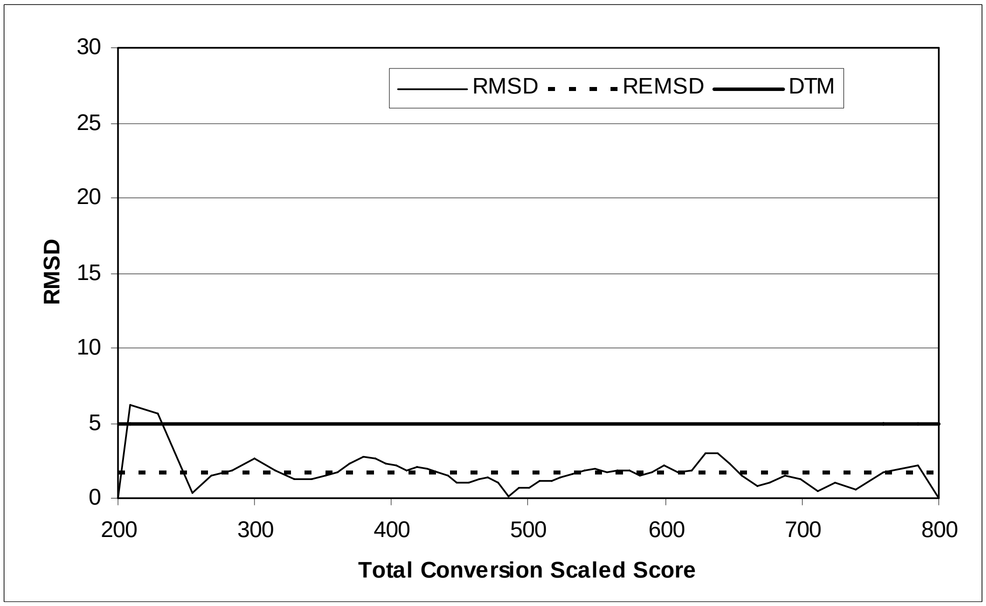


Figure A8. **ENF8 Math: Gender and ethnicity.**

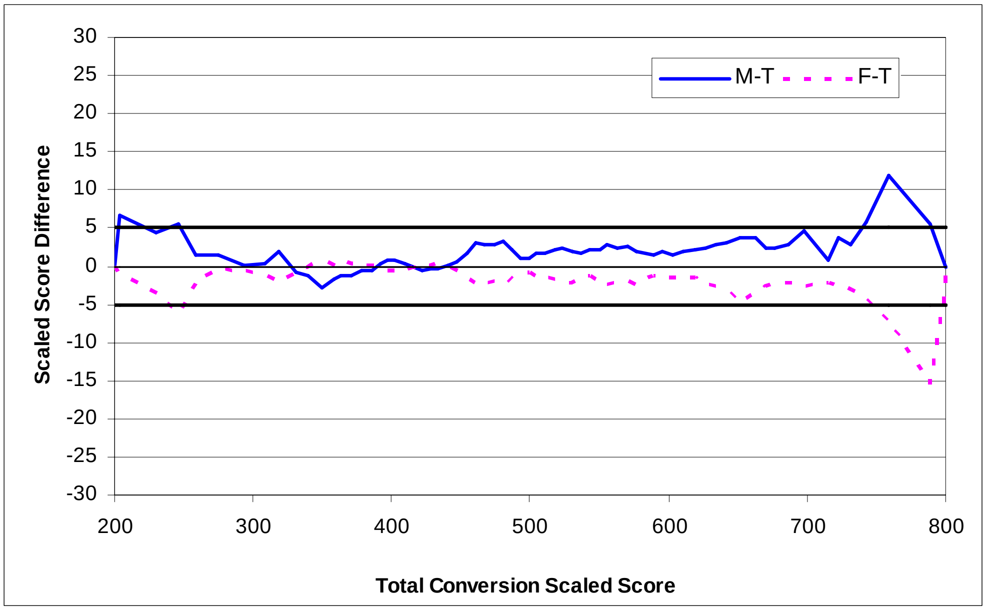


Panel A

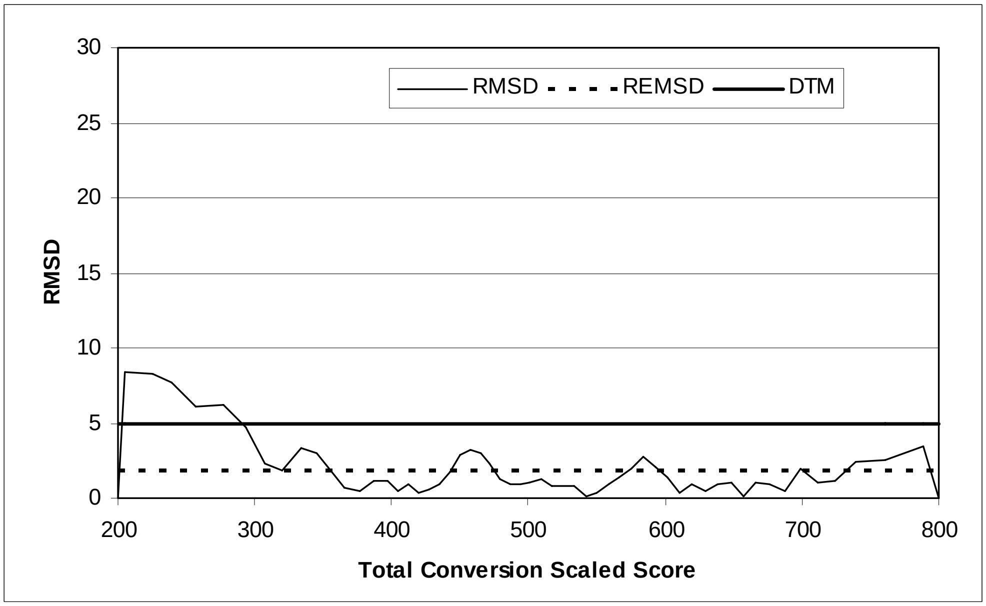


Panel B

Figure A9. **ANF1 Math.**

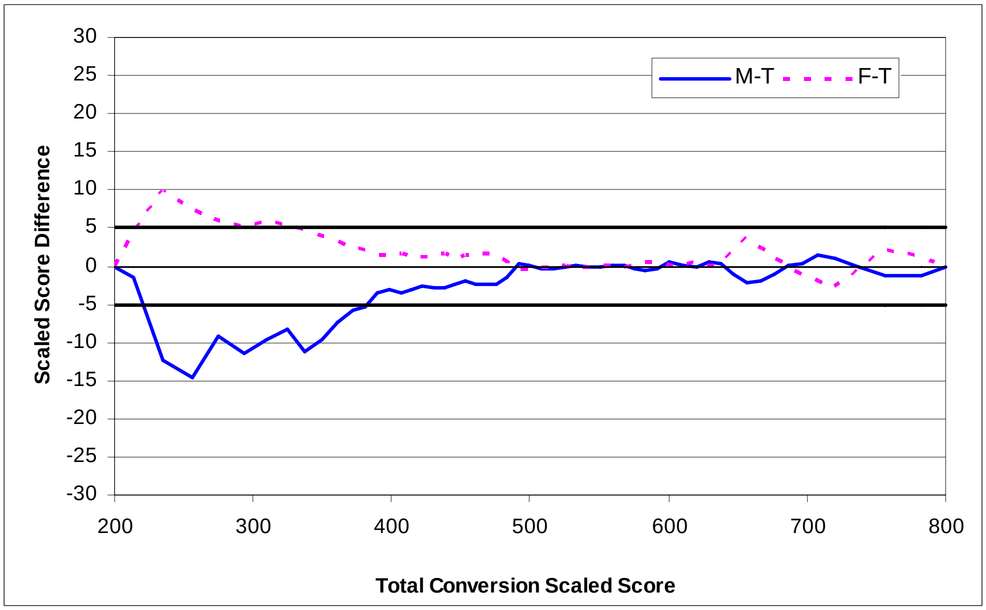


Panel A

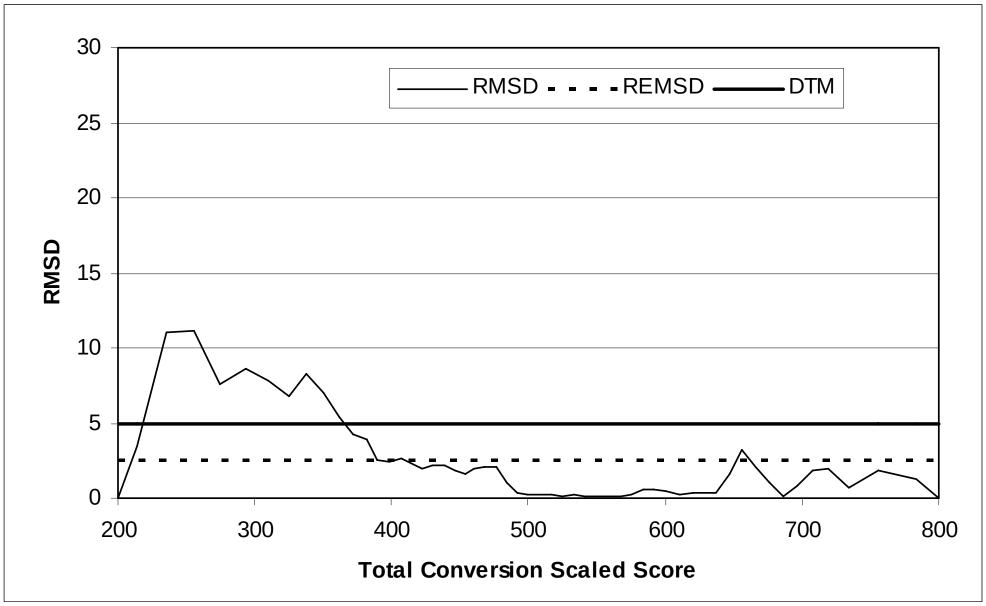


Panel B

Figure A10. **ANF2 Math.**

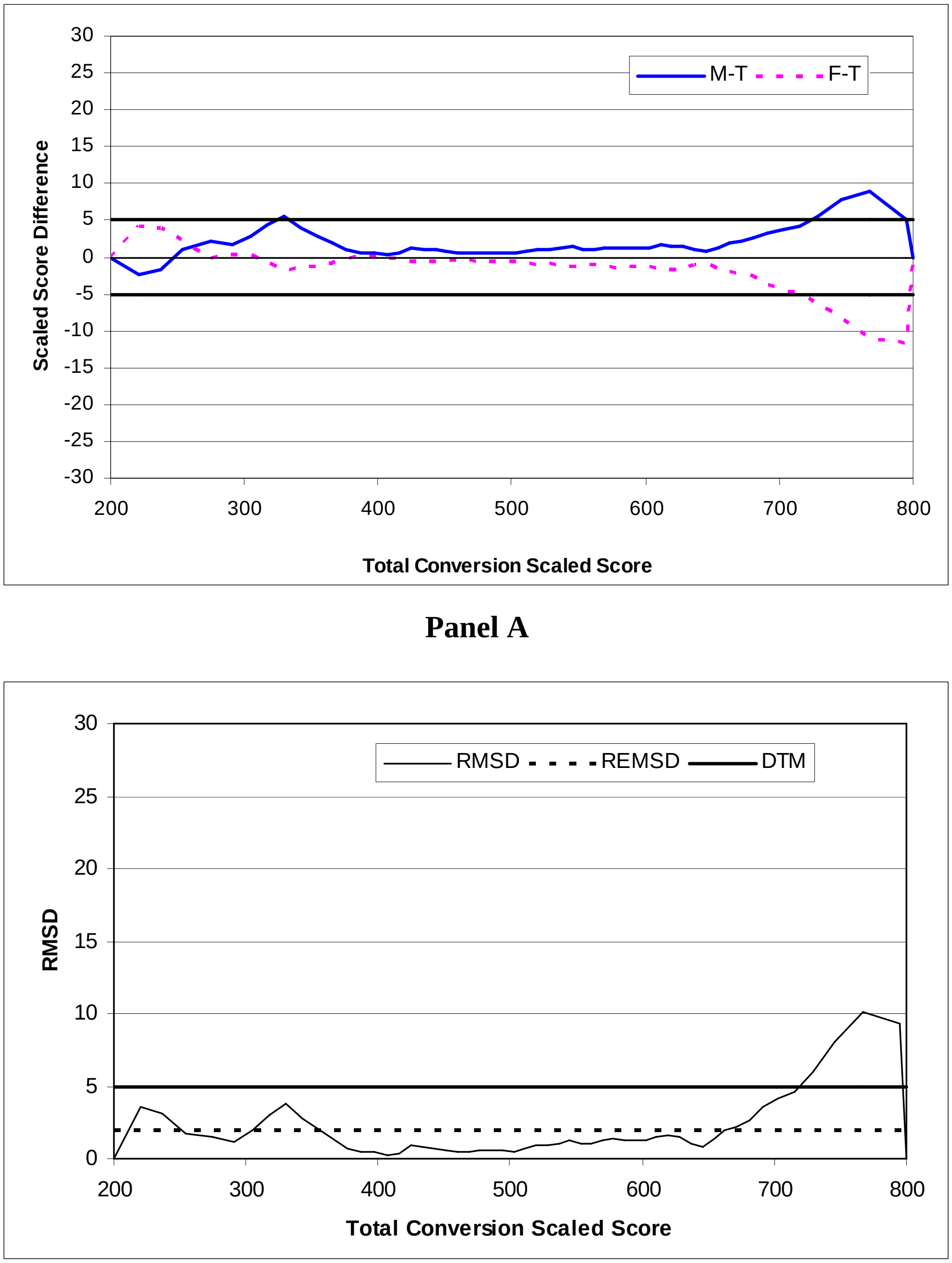


Panel A



Panel B

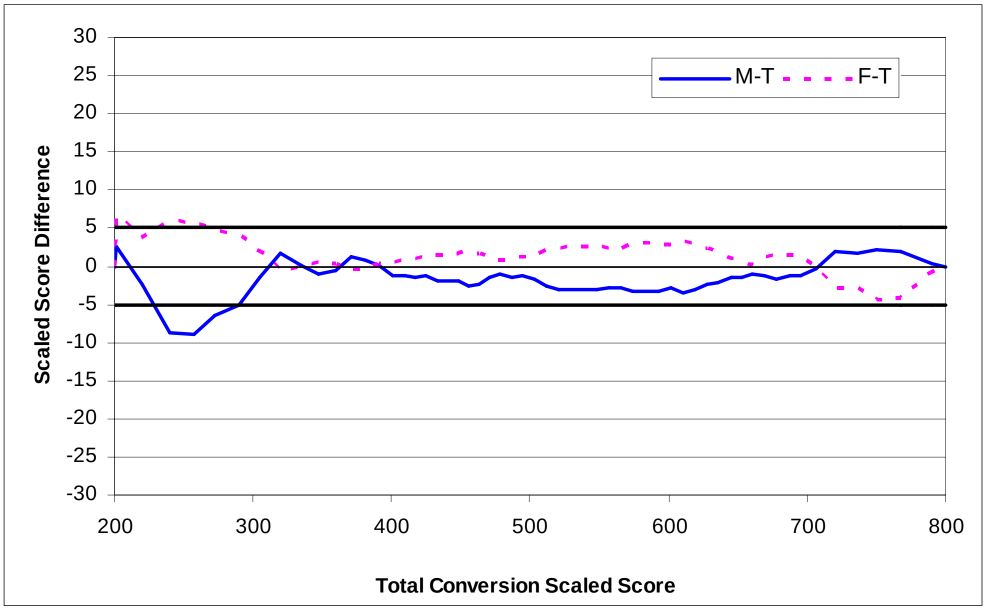
Figure A11. **ANF3 Math.**



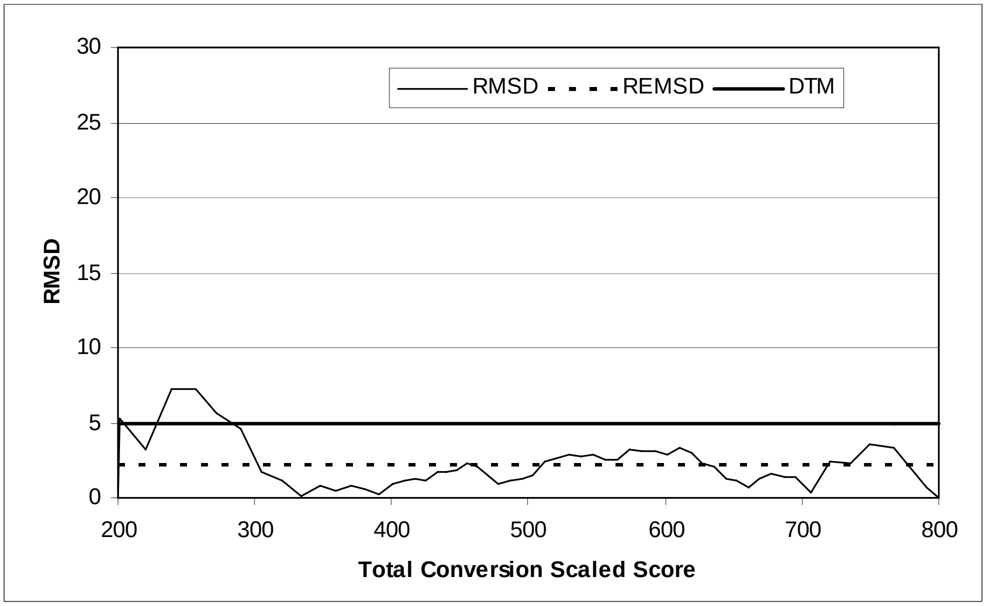
Panel A

Panel B

Figure A12. **ANF4 Math.**

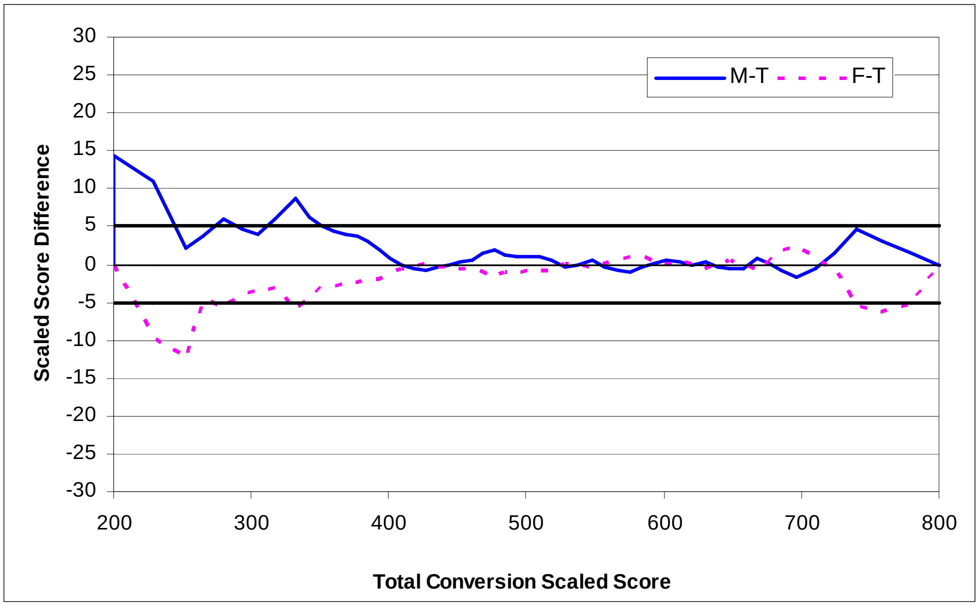


Panel A

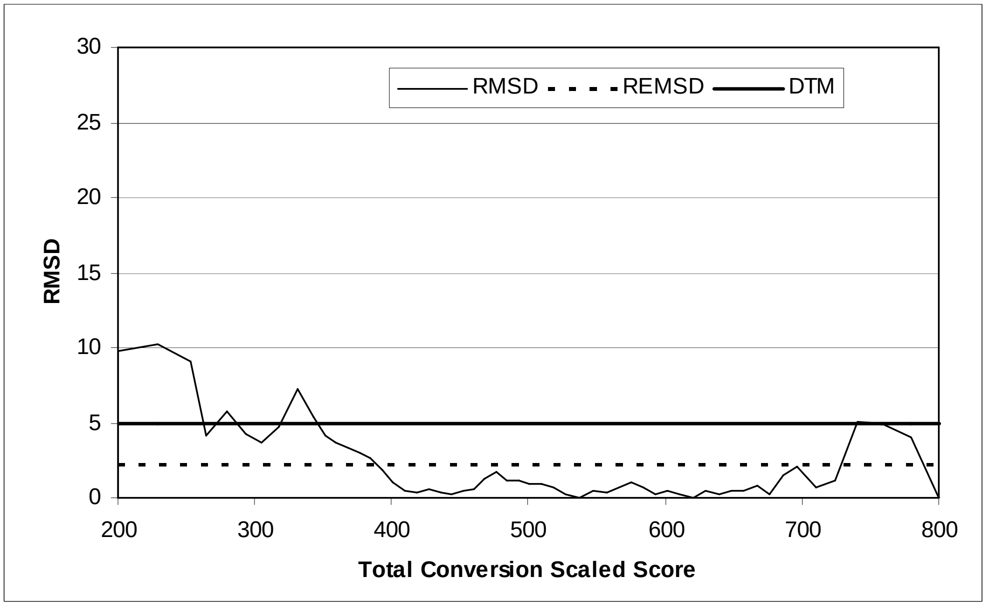


Panel B

Figure A13. ANF5 Math.



Panel A



Panel B

Figure A14. **ANF6 Math.**