Research Report

Subpopulation Invariance of Equating Functions

Gautam Puhan

Kevin C. Larkin

Stacie L. Rupp

Research & Development

August 2006
RR-06-25

Subpopulation Invariance of Equating Functions

Gautam Puhan, Kevin C. Larkin, and Stacie L. Rupp
ETS, Princeton, NJ

August 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study examined population invariance of equating functions over subgroups defined by ethnicity on a teacher certification test. Investigating subgroup equating invariance was important because the total group who took this test consists of two subgroups (i.e., Hispanic and non-Hispanic) and the Hispanic group is a distinctively more able group as compared to the non-Hispanic group on the construct being measured. The study used data collected during the 2003 and 2004 administration of a teacher certification test. The chained equipercentile and linear equating methods were used to derive the equating functions. The root expected square difference was used to compare equating functions derived using the total group with equating functions derived using either the Hispanic or non-Hispanic groups. Findings suggested lack of subgroup invariance in equatings for the first test form (Form *X*). Also, the Hispanic group equating was less invariant as compared to the non-Hispanic group equating. The second form of the test (Form *Y*) showed more subgroup invariance in equating. This difference may partly be attributed to the fact that Form *Y* had a much bigger sample size as compared to Form *X* and the difference in equating functions observed in Form *X* for the total group and the two subgroups may be due to sampling variability. Implications of these results on actual pass/fail rates are also presented and discussed.

Key words: Test equating, population invariance, DTM, pass/fail, root expected squared difference

Acknowledgments

The authors would like to thank Alina von Davier, Anna Kubiak, Wen-Ling-Yang, Dan Eignor, and David Wright for helpful comments and suggestions. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Theoretical Background

Test equating is a statistical procedure used to measure and adjust for difficulty differences across parallel forms of a test, which in turn allows for score comparisons across different groups of examinees regardless of the test forms they were administered. *Population invariance* is considered an important requirement for equating because if equating functions derived from different subpopulations (such as those defined by gender or ethnicity) are systematically different, then the interchangeability of alternate test forms becomes questionable. According to Angoff (1971), population invariance states "except for random error associated with the unreliability of the data and the method used to determine the score transformation, the equating function is independent of examinees or subgroups of examinees from whom the data are drawn to develop the conversion" (p. 563). However, from a practical point of view, strict population invariance is impossible to achieve. Although an equating function derived from a particular total group may be quite satisfactory for most subgroups within the total group, it is unlikely that it will be satisfactory if the subgroups have a markedly different mean and variance of ability as compared to the total group (Petersen, Kolen, & Hoover, 1989).

Research studies on subgroup invariance of equating relationships have found inconsistent results. For example, Cook, Eignor, and Taft (1988) have found differences in equating functions for a biology exam derived using examinees from a December administration, where most examinees were seniors and had not taken a biology course recently, and examinees from a June administration, where most examinees had recently completed a biology course. Van der Linden (2000) demonstrated that the equipercentile observed score equating function can differ when different ability level data are used to compute the equating functions. However, Dorans and Holland (2000) studied equating results for different subgroups based on gender, ethnicity, and language and found little evidence for population dependency of the equating relationships. Similarly, Yang and Gao (2004) found negligible differences between equating functions derived from subgroups based on gender and that for the total group on the College Level Examination Program (CLEP). As seen above, results of some studies show population dependency of equating functions while other studies do not show much population dependency of equating functions.

According to Yang (2004)**,** if equating functions are not invariant across subpopulations, then the new and old test forms are not equitable because if the equating derived for one

subpopulation is not invariant enough to be applied to other subpopulations, then generalizing the equating function to different subpopulations is difficult. Since distinct subgroups based on language and ethnicity (which are often markedly different in terms of ability) are increasingly becoming more noticeable in educational testing, it is important to examine whether an equating function derived for the total group is affected by the presence of different subgroups. For example, an equating function derived for a test intended to measure basic language skills may be adversely affected by the presence of a subgroup with very high language ability. Since the test is intended to measure basic language skills, it may be fair to exclude the high language ability subgroup from the total group while deriving an equating relationship.

The purpose of this study is to compare the equating functions derived using two different subgroups with the equating function derived using the total group on a large-scale certification test. Investigating subgroup invariance in equating relationships was especially important for this large-scale certification test because the total group who took this test consists of two subgroups (i.e., Hispanic & non-Hispanic) and the Hispanic group is a distinctively more able group as compared to the non-Hispanic group on the construct being measured. A negligible difference in the equating functions derived using the Hispanic and the non-Hispanic groups and the total group would indicate that the equating relationship derived using the total group is not significantly affected by the presence of these subgroups in deriving that function. The present study is designed to evaluate this assumption. It should be noted that throughout this paper the terms *subgroup invariance* and *subpopulation invariance* are used interchangeably.

Method

Teacher Certification Test and Examinee Sample

The study used test data collected during the 2003 and 2004 administrations of a large-scale certification test. The test consists of all multiple-choice items measuring knowledge and competencies necessary for a beginning-level teacher of Spanish and is used by states as a requirement for their teacher certification. Two alternate forms of the test were used (throughout this paper, the first form will be referred to as Form *X* and the second form will be referred to as Form *Y*). Form *X* consisted of 113 operational items and the reference form of the test (i.e., the form to which the Form *X* is equated) consisted of 139 operational items. Similarly, Form *Y* consisted of 120 operational items and the reference form of the test consisted of 140 operational items. Form *X* was equated to the old form using 22 internal anchor items. Form *Y* was equated

to the old form using 52 internal anchor items. Both Forms *X* and *Y* are right scored (i.e., total test score is simply the sum of all correct responses).

The sample for Form *X* consisted of 815 examinees of which approximately 16% were Hispanics and 82% were non-Hispanics or predominantly English-speaking examinees. Similarly, the sample for Form *Y* consisted of 1,952 examinees (accumulated across three testing administrations), of which approximately 18% were Hispanics and 68% were non-Hispanics or predominantly English-speaking examinees. The remaining test takers in Forms *X* and *Y* could not be classified into either the Hispanic or non-Hispanic category. However, they were retained for the analysis conducted with the total group of examinees.

This study was broken into three steps:

1. The equating relationships were derived for the new and reference forms using three different samples (i.e., the total sample, the non-Hispanic group, and the Hispanic group).
2. The difference in the equating functions derived using the three different samples was compared using statistical indices.
3. Finally, the impact of the three different conversions on actual pass/fail status of examinees was examined.

Analytical Procedure

The chained equipercentile and chained linear equating methods using anchor items were used to derive the equating relationships between the new and reference forms of the test. Under chained linear or chained equipercentile equating, the new form scores are equated to scores on the anchor items using the sample that took the new form. Then scores on the anchor are equated to scores on the reference form using the sample that took the reference form. Finally, these two conversions are chained together to produce a conversion of the new form scores to the reference form scores (see Livingston, 2004, and Kolen & Brennan, 2004). It should be noted that before the actual equating relationship was derived for the new and reference forms, a log-linear smoothing function (Holland & Thayer, 1987) was applied to the score distributions for adjusting the irregularities of the score distributions, which may cause problems for chained equipercentile equating (smoothing score distributions does not affect linear equating results).

The subgroup equating functions were compared to the total group equating function using the root expected squared difference (RESD), which describes a weighted average of differences between a subgroup equating function and total group equating function (see Yang & Gao, 2004). The RESD is computed by averaging the squared differences between the subgroup equating function and the total group equating function at each raw score level relative to the number of examinees at each raw score level and taking the square root of this value and dividing it by the standard deviation of the total group. The RESD statistic is defined as

$$RESD_j = \frac{\sqrt{\sum_{x=0}^{X} w_{xT}\{[e_T(x) - e_j(x)]^2\}}}{SD_T}$$

where x represents each raw score point, j denotes a subgroup, e_T represents the equated scores for the total group, e_j represents the equated scores for a particular subgroup, w_{xT} is the weighting factor indicating the proportion of examinees in the total group at each raw score level, and SD_T is an adjusted standard deviation of the reference form in the total group. Although one can use the SD of the reference form for the total group, use of the adjusted SD, which provides a better estimate of the SD in the target population, has been suggested by von Davier, Holland, and Thayer (2004). It is calculated as

$$SD_T = (\sqrt{w_n(\sigma_n)^2 + w_r(\sigma_r)^2 + (w_n(1-w_n)(M_n - M_r)^2}) \ \frac{\sigma_t}{\sigma_r},$$

where w_n and w_r are the weights for the new and reference forms, respectively, σ_n and σ_r are the standard deviations of the anchor for the new and reference forms in the total group, respectively, M_n and M_r are the means for the anchor for the new and reference groups, respectively, and σ_t is the standard deviation for the reference form for the total group. Smaller values of RESD indicate a negligible difference between two different equating functions.

Criterion for evaluating a large difference. Dorans and Feigenbaum (1994) proposed the notion of differences that matter (DTM) to evaluate when the value of the statistical indices discussed above is large enough to evoke questions about the equitability of tests. According to

the DTM notion, any difference that is within half of the reported score unit can be treated as close enough to ignore. This is particularly useful for tests using cut scores where one half of the reported score unit can potentially change the pass or fail status of examinees. Because scaled scores are reported in 1-point units for this particular Spanish test, half of this unit can make a difference in pass/fail status of examinees. Therefore a DTM of 0.5 will be used to evaluate differences between the total and subgroup equating functions for this test. To obtain a standard score equivalent, the DTM is divided by the adjusted standard deviation of the reference form for the total group. It should be noted that the equating relationships that are examined in the current study involve raw-to-raw score transformations and therefore the DTM derived for the scaled score may slightly differ for that derived from the raw score. Nevertheless, it will be used as an evaluative criterion for assessing difference between the subgroup and total equating functions.

Pass/fail decisions based on subgroup and total equating functions. Finally, the actual pass/fail status of examinees based on the different equating functions will be examined. This is important because statistically significant results may not be important practically. On the other hand, if the differences between two equating functions are found to be statistically insignificant but by using either of the equating functions, considerable differences in actual pass and fail decisions are observed, then the statistical results may be less useful. This can happen particularly in cases where scores derived using one equating chain as compared to another are very close to each other but would round to different scores. For example a difference between scores such as 150.4 and 150.6 would be undetected using the DTM criteria but 150.4 would round down to 150 and 150.6 would round up to 151. Consequently this may lead to a difference in pass/fail status of examinees who received such scores. As mentioned earlier, the equating results examined in the current study are for the raw-to-raw score transformation of the new and reference forms, although the actual scores for this test are reported in scaled score units after a linear transformation of the raw scores. Therefore the pass/fail decisions of examinees based on equating transformations derived from using different subgroups are based on scaled scores. It should be noted that the actual names of the user states and the cut scores associated with individual states are not identified in this paper for confidentiality reasons. Instead, hypothetical names for the user states (i.e., State 1 or State 2) have been used.

Why were raw scores used instead of scaled scores? Although this test reports scores on a 100 to 200 scale with increments of 1, all analysis (except the pass/fail decision results) were

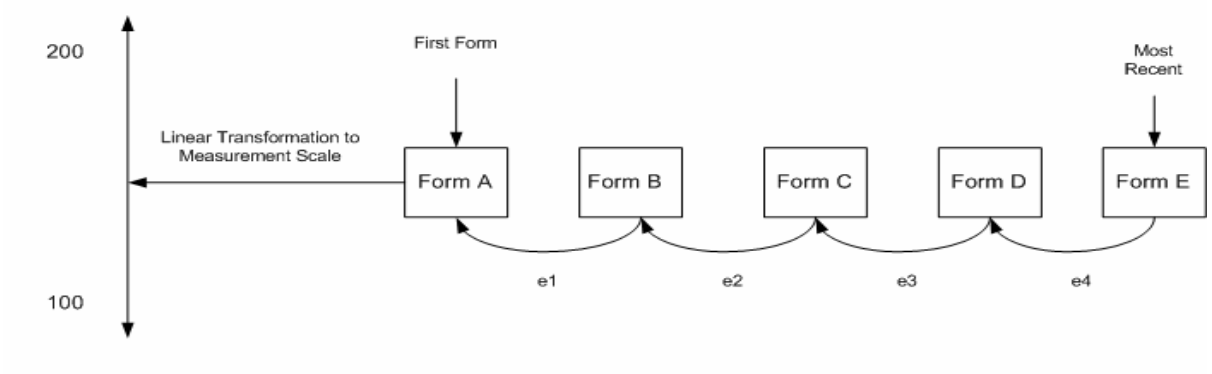

Figure 1. **Example of a hypothetical equating chain that links to the original base scale.**

Note. e1 through e4 represent equating errors.

conducted using raw scores. The primary reason for using raw scores instead of scaled scores is that the scaled scores for this test are linked to the original base scale (i.e., the first form of the test) through a series of intermediate equatings as shown in Figure 1.

In Figure 1, the raw–to-raw score transformation of scores on Form E to scores on Form D involve only one source of equating error (e4). However, if the comparisons of equating functions were made in scaled score units, then the intermediate equatings that lead to the measurement scale involve equating errors accumulated for the four equatings shown above (e1 through e4). To avoid this accumulation of error, which may make the comparison of equating functions for the total group and subgroups less precise, it was decided to use raw scores instead of scaled scores.

Determining the raw cut score range for the two test forms. Since equating differences for the total and subgroup equating functions were examined at the raw score instead of the scaled score level, giving an exact range of cut scores in raw score units for the two test forms is problematic because different user states set their cut scores in scaled score units. Nevertheless a raw cut score range was specified based on how many raw score points are needed to get a particular scaled score for any particular test form (Form *X* or Form *Y*). It should be noted that the raw cut score ranges may vary slightly depending on which equating function (total, non-Hispanic, or Hispanic) was chosen. For example, a scaled score of 150 may correspond to a raw score of 85 when the total group equating function is chosen. However, the same scale score may correspond to a slightly higher or lower raw score when the non-Hispanic or Hispanic group

equating functions are chosen. Therefore the final raw score range for a particular form was determined by taking the lowest and the highest raw cut score observed across the three conversions derived from the total group and the two subgroups. By doing so, the raw score range for Form *X* was determined as 60 to 90 and the raw score range for Form *Y* was determined as 68 to 90.

All analyses described above were first completed for Form *X* and then replicated for Form *Y*. Analysis of two test forms of the same test using the same subgroups and equating methods may help in finding converging evidence regarding subgroup invariance of equating functions. Furthermore, since Form *X* consisted of smaller sample sizes for the subgroups, replication of the study with Form *Y*, which has larger sample sizes for the subgroups, based on accumulated test data from different testing administrations may lead to more stable results.

Results

All results presented below were obtained using the chained linear and chained equipercentile method of equating using pre-smoothed score distributions. The results for Form *X* will be presented first followed by results for Form *Y*. The DTM criterion was used for evaluating differences between the equating functions derived using the total group and the different subgroups. The DTMs for Form *X* and Form *Y* are 0.022 and 0.023, respectively (calculated by dividing 0.5 by the adjusted SD estimated from using the synthetic group). Any equating differences between the total group and the different subgroups that are smaller than these values for the respective test forms were considered negligible.

Results for Form X

The means and standard deviations for the new and reference forms for the total, non-Hispanic, and Hispanic groups are presented in Table 1. As seen in Table 1, the mean of the total group for both the new and reference forms is slightly higher than the mean of the non-Hispanic group. However, the mean of the Hispanic group is considerably higher than the total and non-Hispanic groups, suggesting that the Hispanic group is a much higher ability group as compared to the total and non-Hispanic groups. Also, the Hispanic group showed less variability in scores as compared to the total and non-Hispanic groups for both the new and reference forms.

Table 1

Means and Standard Deviations for Form X

	Total	Hispanic	Non-Hispanic
New form (113 items) 22 anchor items			
N	815	136	667
Mean (T)	81.30	98.31	78.53
SD (T)	18.38	9.83	17.80
Mean (A)	16.55	19.87	15.98
SD (A)	3.90	2.10	3.84
Reference form (139 items)			
N	445	78	347
Mean (T)	96.17	113.88	91.22
SD (T)	21.52	18.76	20.22
Mean (A)	16.76	19.06	15.98
SD (A)	3.65	2.89	3.84

The chained linear and chained equipercentile equating functions derived using the total group and the two subgroups (Hispanic and non-Hispanic) are presented in Figures 2 and 3, respectively. As seen in Figure 2, the chained linear equating functions for the total group and the non-Hispanic group are very similar (especially in the cut or passing score ranges), indicating a negligible difference between these equating functions. However, the equating functions for the total group and the Hispanic group are slightly different from each other for most parts of the score scale, indicating a potentially non-negligible difference between these equating functions. Similarly, as seen in Figure 3, the chained equipercentile equating functions for the total group and the non-Hispanic group are very similar (especially in the cut or passing score ranges), indicating a negligible difference between these equating functions. However, the equating functions for the total group and the Hispanic group are slightly different from each other for most parts of the score scale, indicating a potentially non-negligible difference between these equating functions**.**

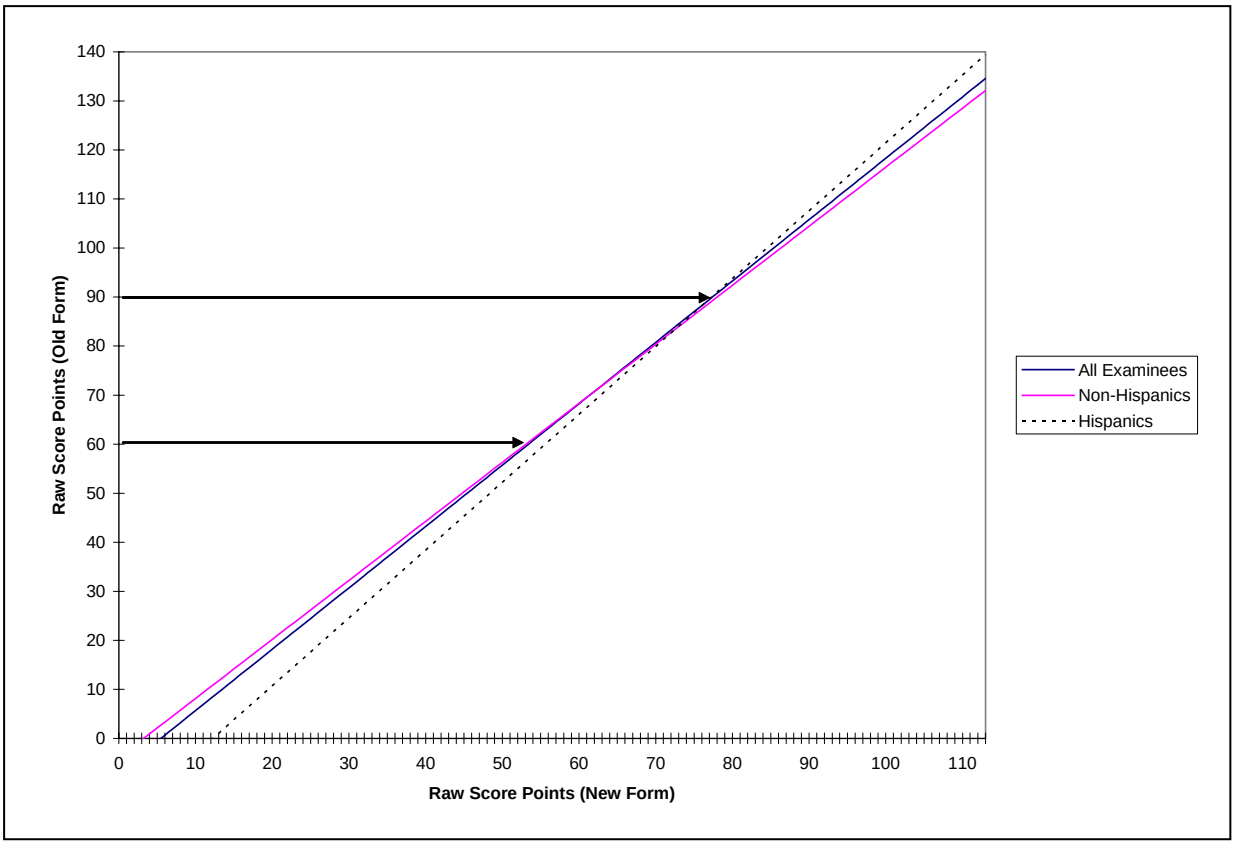

Figure 2. **Chained linear equating functions derived from the total group and two subgroups (Form *X*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

Actual standardized difference between total and subgroup equating functions. Results of the actual standardized difference[1] between the subgroup and the total group equating functions are presented in Figures 4 and 5. Since differences in equating functions may make differences in pass/fail decisions, the DTM criteria was used to evaluate the differences in the subgroup and total equating functions. It should be noted that although differences between equating functions were examined for the complete score scale for evaluating subgroup invariance of equating relationships, any differences that may exist between the total and subgroup equating functions beyond the cut score range would not be practically important (i.e., does not affect pass/fail status of test takers).

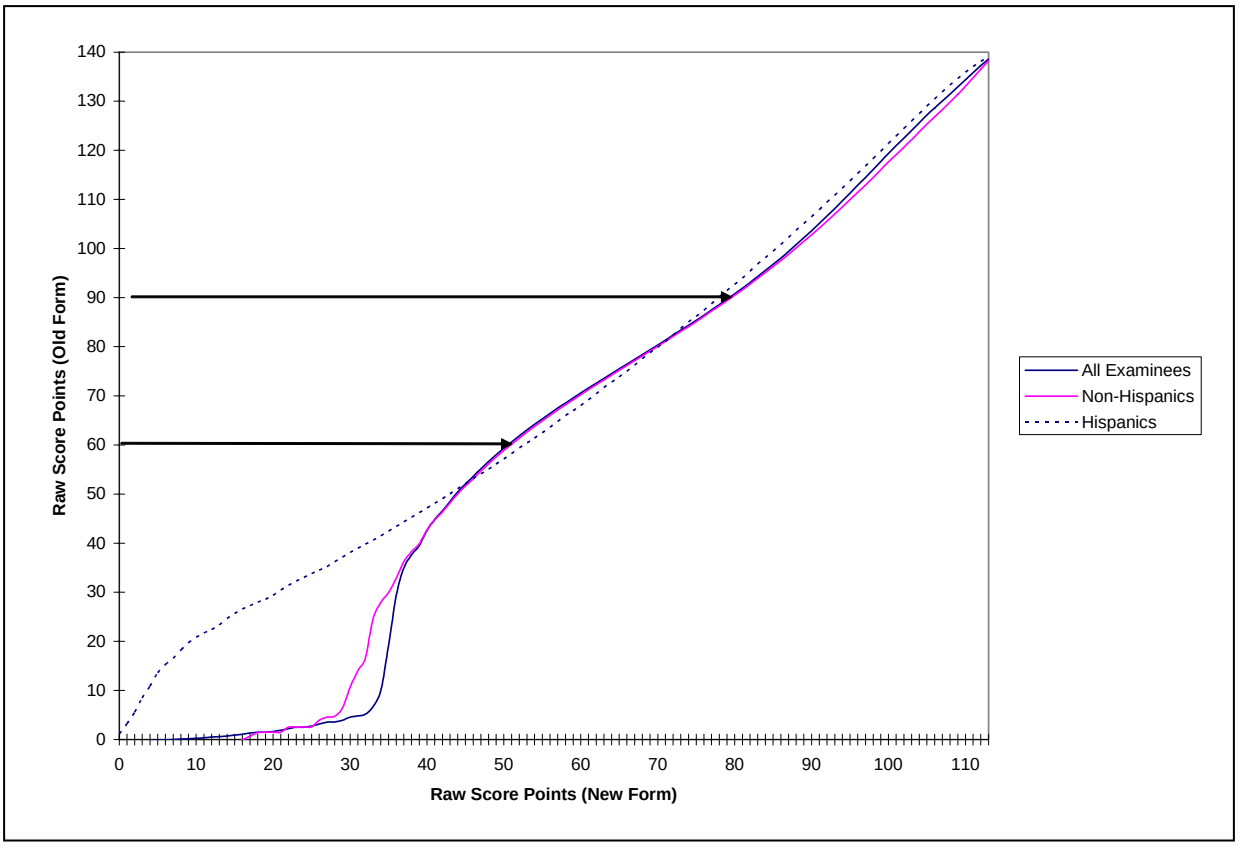

Figure 3. **Chained equipercentile equating functions derived from the total group and two subgroups (Form *X*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

As seen in Figure 4, the chained linear equating functions for the total group and the non-Hispanic group show negligible differences in the lower and middle points of the cut score region but potentially non-negligible differences in the higher end of the cut score region. However, the equating functions for the total group and the Hispanic group show larger differences between the lower and higher cut score regions and negligible differences in the middle range of the cut score region. Similarly, as seen in Figure 5, the chained equipercentile equating functions for the total group and the non-Hispanic group show negligible differences in almost all parts of the cut score range except for the very high cut score region, where there is a slight difference (0.03). However, the equating functions for the total group and the Hispanic group show larger differences between the lower and higher cut score regions and negligible differences in the middle range of the cut score region.

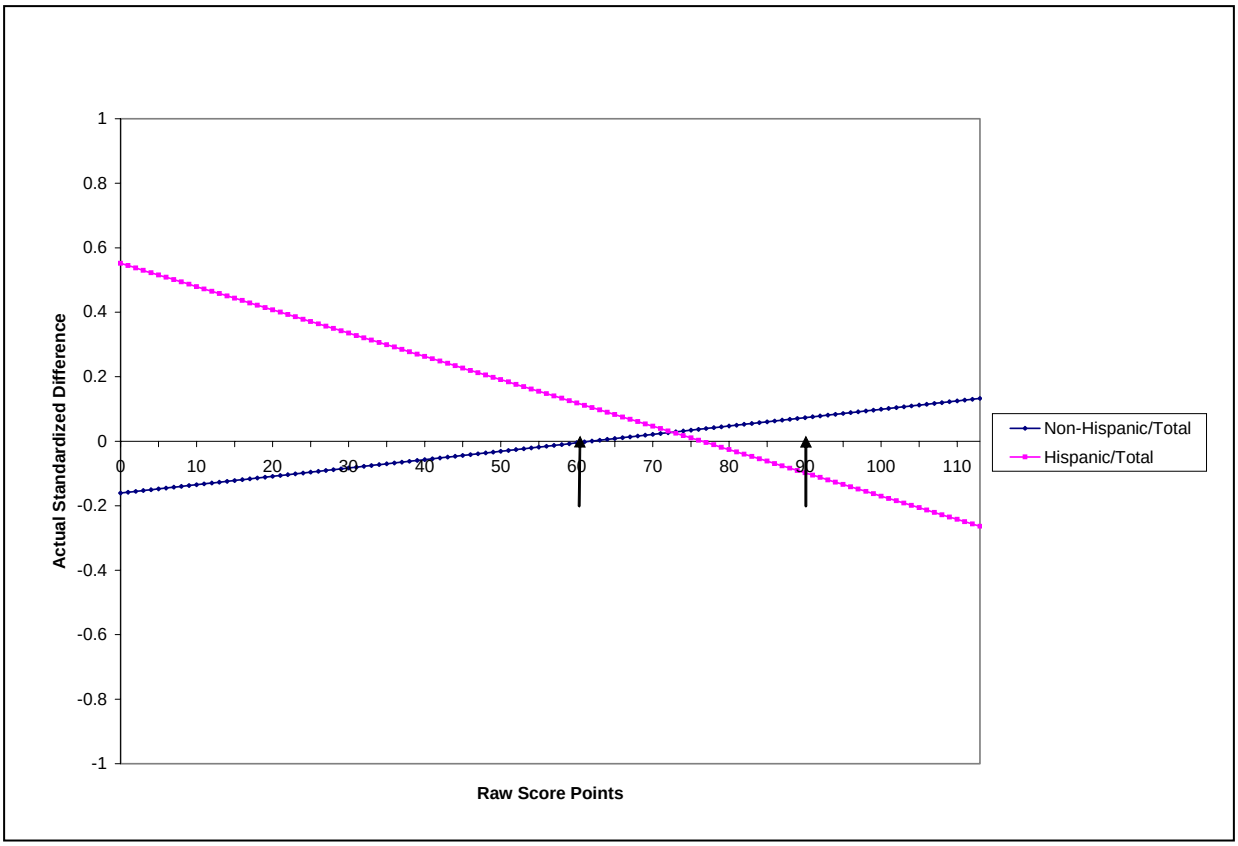

Figure 4. **Actual standardized difference between the total and subgroup linear equating functions (Form *X*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

Results from RESD analyses. For the chained linear equating functions, the RESD for the total group and non-Hispanic group comparison (0.05) is larger than the DTM (i.e., 0.022), indicating that the equating difference for the non-Hispanic group and the total group is potentially non-negligible. However, it should be noted that since the difference is quite small, it may not substantially affect pass and fail status of test takers. The RESD for the total group and Hispanic group comparison is considerably large (i.e., 0.11), indicating that the equating difference for the Hispanic group and the total group equating function is potentially non-negligible. Similarly, for the chained equipercentile equating functions, the RESD for the total group and non-Hispanic group (0.05) comparison is larger than the DTM (i.e., 0.022), indicating that the equating difference for the non-Hispanic group and the total group is potentially non-negligible, and the RESD for the total group and Hispanic group comparison is considerably

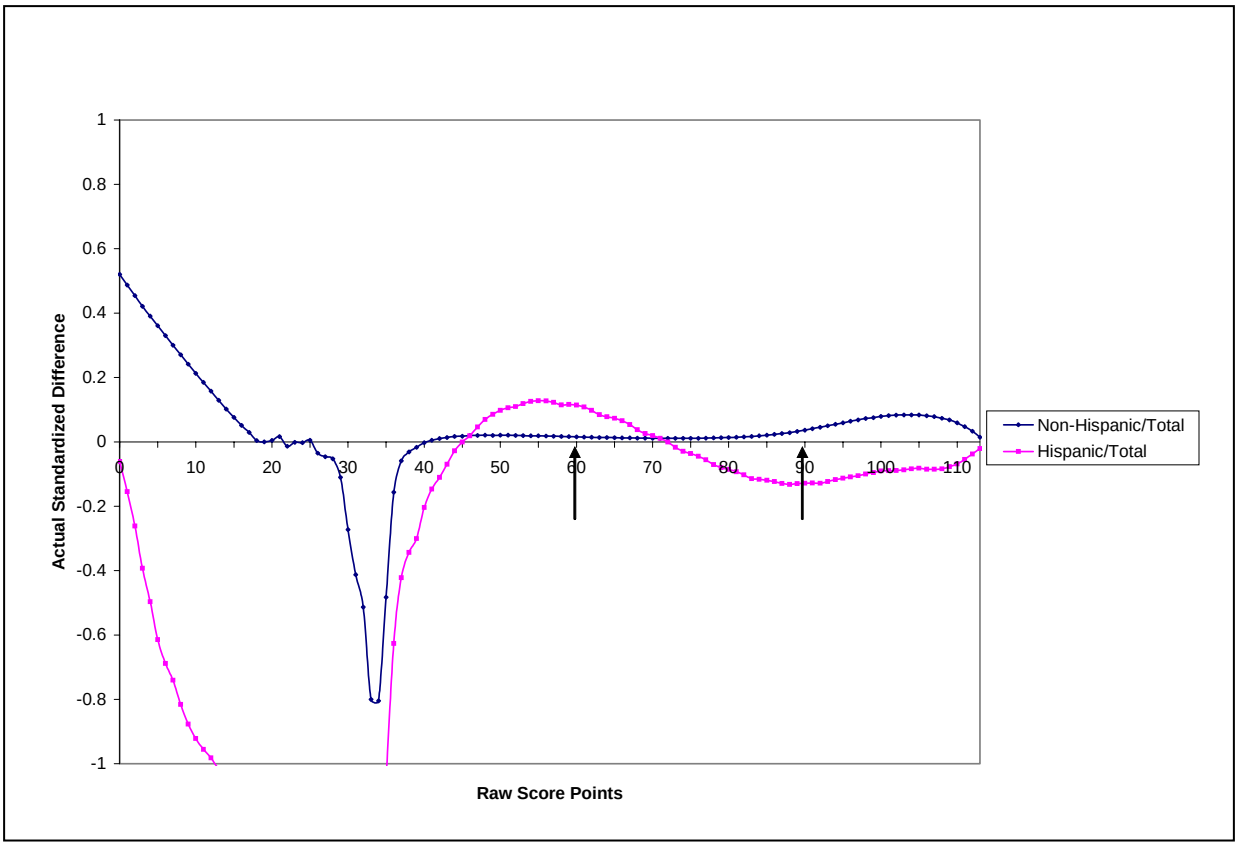

Figure 5. **Actual standardized difference between the total and subgroup equipercentile equating functions (Form *X*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

large (i.e., 0.13), indicating that the equating difference for the Hispanic group and the total group equating function is potentially non-negligible.

The RESD results indicate that the difference between the non-Hispanic group and the total group equating functions are smaller compared to the difference between the Hispanic group and the total group equating functions. Furthermore, the difference between the non-Hispanic group and the total group equating functions as compared to the Hispanic group and total group equating functions are negligible for a wider range of the cut score region.

Pass rates of test takers using the three different equating functions. The actual pass percentages of test takers from different user states with different cut scores are reported in Table 2. As seen in Table 2, the pass percentages for test takers using the three different equating functions are identical for some user states but slightly different for other user states using

different passing or cut scores. This is true for both the chained linear and chained equipercentile equating functions. It should be noted that the number of test takers whose pass or fail status may change could be large or small in absolute magnitude depending on how many test takers write this test for a particular user state.

Table 2

Actual Pass Percentages of Examinees Using Conversions Derived From the Three Groups (Form X)

State	*N*	Chained linear			Chained equipercentile		
		Total	Hispanic	Non-Hispanic	Total	Hispanic	Non-Hispanic
State 1	83	61.45	61.45	60.24	59.04	60.24	56.63
State 2	37	86.49	83.78	86.49	86.49	83.78	86.49
State 3	44	72.73	72.73	72.73	72.73	72.73	70.45
State 4	153	84.31	82.35	84.31	84.97	82.35	84.97
State 5	98	45.92	45.92	45.92	42.86	45.92	42.86
State 6	59	66.10	64.41	64.41	66.10	64.41	66.10
State 7	28	53.57	50.00	50.00	50.00	50.00	46.43

Results for Form Y

The means and standard deviations for the new and reference forms for the total, non-Hispanic, and Hispanic groups are presented in Table 3. As seen in Table 3, the mean of the total group for both the new and reference forms are slightly higher than the mean of the non-Hispanic group. However, the mean of the Hispanic group is considerably higher than the total and non-Hispanic groups, suggesting that the Hispanic group is a much higher ability group as compared to the total and non-Hispanic groups. As seen in case of Form *X*, the Hispanic group for Form *Y* also shows less variance in scores as compared to the total and non-Hispanic groups.

Table 3

Means and Standard Deviations for Form Y

	Total	Hispanic	Non-Hispanic
New form (120 items) 51 anchor items			
N	1,952	341	1,308
Mean (T)	89.06	101.39	84.13
SD (T)	17.58	11.77	17.19
Mean (A)	38.44	44.56	36.04
SD (A)	8.26	4.89	8.13
Reference form (140 items)			
N	2,109	385	1,378
Mean (T)	99.64	115.51	93.24
SD (T)	21.32	14.91	20.21
Mean (A)	37.56	44.16	34.91
SD (A)	8.57	5.22	16.24

The chained linear and chained equipercentile equating functions derived using the total group and the two subgroups (Hispanic and non-Hispanic) are presented in Figures 6 and 7, respectively. As seen in Figure 6, the chained linear equating functions for the total group and the non-Hispanic group are very similar (especially in the cut or passing score ranges), indicating a negligible difference between these equating functions. Also, the equating functions for the total group and the Hispanic group are very similar for most parts of the score scale, indicating a negligible difference between these equating functions. Similarly, as seen in Figure 7, the chained equipercentile equating functions for the total group and the non-Hispanic group are very similar (especially in the cut or passing score ranges), indicating a negligible difference between these equating functions. However, the equating functions for the total group and the Hispanic group are slightly different from each other in different parts of the score scale, indicating a potentially non-negligible difference between these equating functions.

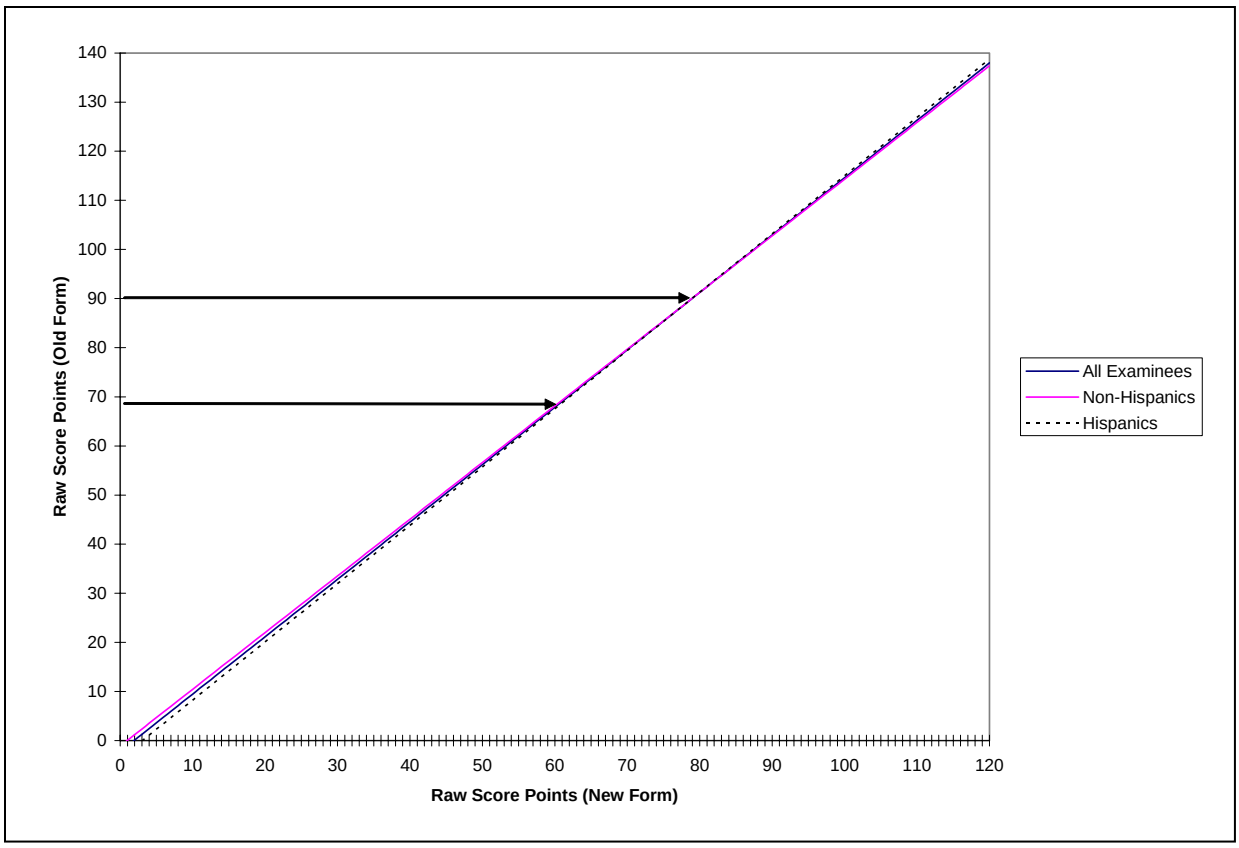

Figure 6. **Chained linear equating functions derived from the total group and two subgroups (Form *Y*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

Actual standardized difference between total and subgroup equating functions. Results of the actual standardized difference between the subgroup and the total group equating functions are presented in Figures 8 and 9. As seen in Figure 8, the chained linear equating functions for the total group and the non-Hispanic group show negligible differences across all parts of the cut score region (point of largest difference = 0.01) and very small but potentially non-negligible differences near the tails of the score scale. Also, the equating functions for the total group and the Hispanic group show negligible differences across all parts of the cut score regions (point of largest difference = -0.01) and very small but potentially non-negligible differences beyond the cut score regions. Similarly, as seen in Figure 9, the chained equipercentile equating functions for the total group and the non-Hispanic group show negligible differences across all parts of the cut score region (point of largest difference = -0.02) and very small but potentially non-negligible differences beyond the cut score regions. Also, the equating functions for the total

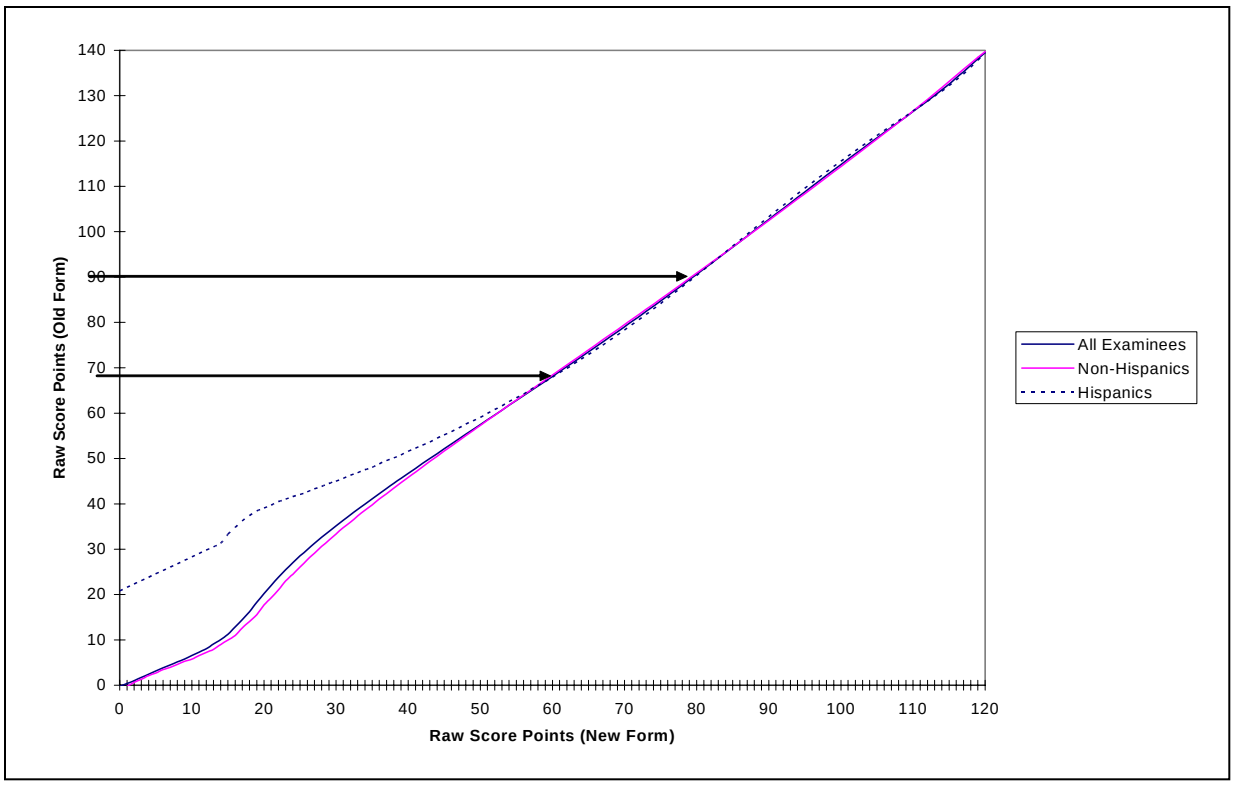

Figure 7. **Chained equipercentile equating functions derived from the total group and two subgroups (Form *Y*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

group and the Hispanic group show negligible differences across all parts of the cut score region except for a few points near the middle, where the largest difference is around 0.03. However, beyond the cut score region there is a large and potentially non-negligible difference in the equating functions. This large difference may be partially attributed to sampling error due to extremely low and often zero frequency counts towards the lower end of the score scale.

Results from RESD analyses. For the chained linear equating functions, the RESD for the total group and non-Hispanic group comparison (0.01) is small when compared to the DTM (0.023), indicating that the equating difference for the non-Hispanic group and the total group is negligible. The RESD for the total group and Hispanic group comparison (0.01) is also small, indicating that the equating difference for the Hispanic group and the total group equating function is negligible. Similarly, for the chained equipercentile equating functions, the RESD for the total group and non-Hispanic group comparison (0.01) is also smaller than the DTM (0.023),

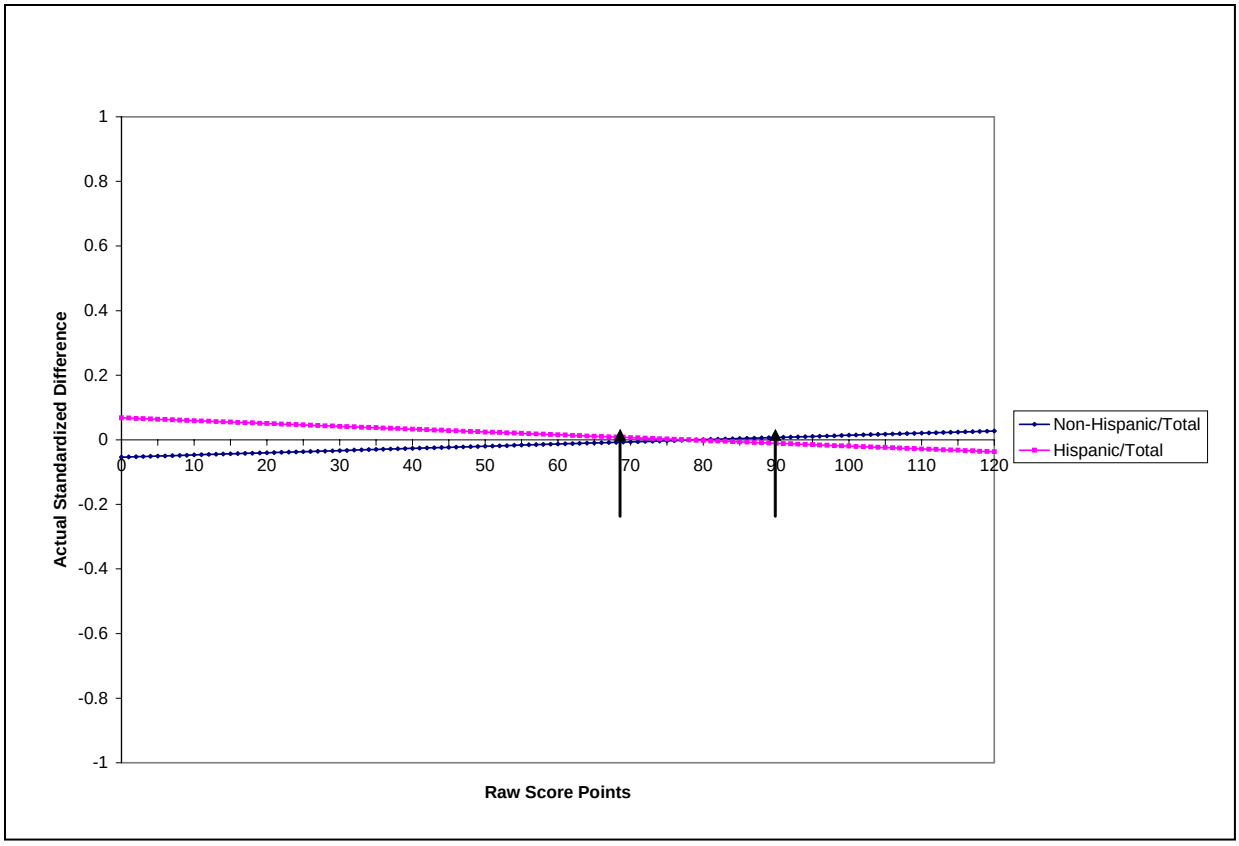

Figure 8. **Actual standardized difference between the total and subgroup linear equating functions (Form *Y*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

indicating that the equating difference for the non-Hispanic group and the total group is negligible. However, the RESD for the total group and Hispanic group comparison (0.03) is slightly larger than the DTM, indicating that the equating difference for the Hispanic group and the total group equating function is potentially non-negligible.

The RESD results indicate that the difference between the non-Hispanic group and the total group equating functions is negligible. Similarly, the difference between the Hispanic group and the total group equating functions is also small but potentially non-negligible, as shown by the chained equipercentile results.

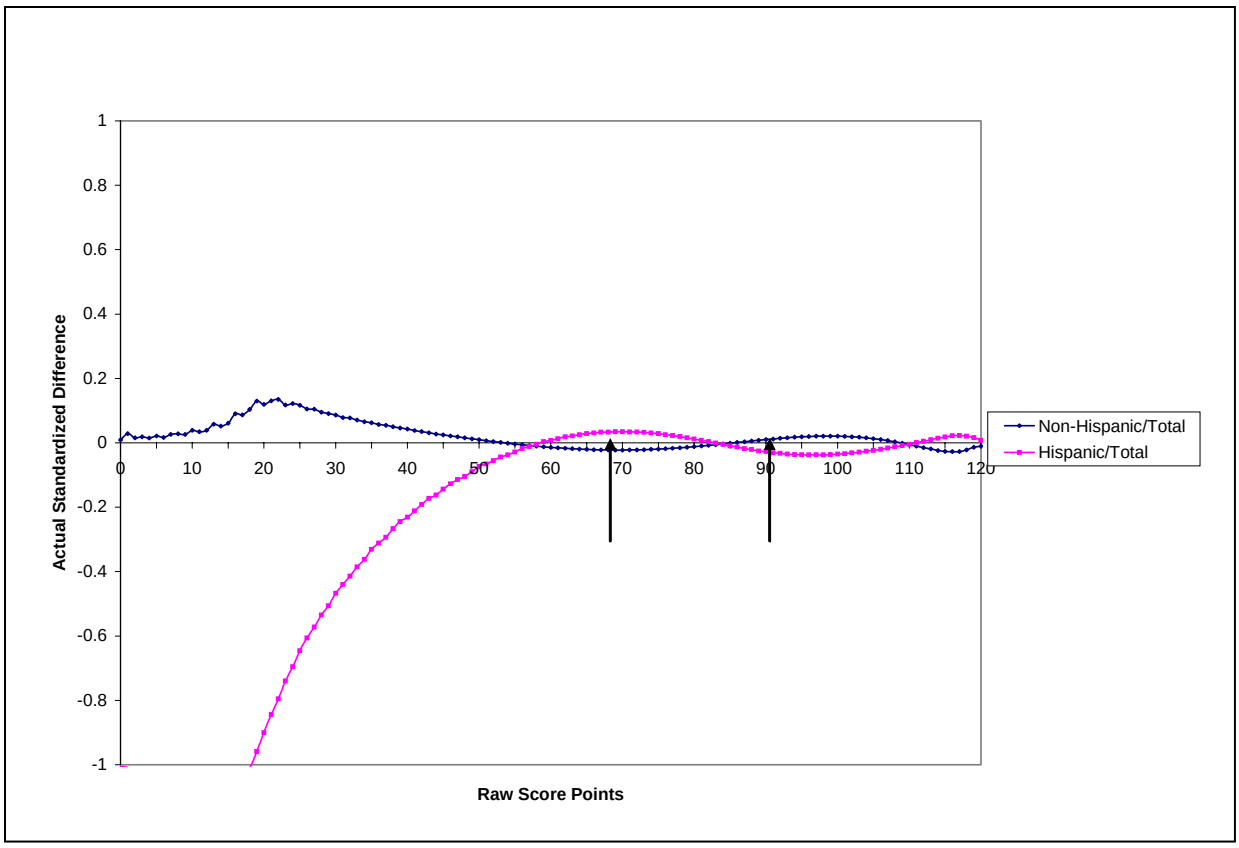

Figure 9. **Actual standardized difference between the total and subgroup equipercentile equating functions (Form *Y*).**

Note. The two arrows in the figures indicate the lowest and highest cut scores.

Pass rates of test takers using the three different equating functions. The actual pass percentages of test takers from different user states with different cut scores are reported in Table 4. As seen in Table 4, the pass percentages for test takers using the three different equating functions are identical for most user states. This is true for both the chained linear and chained equipercentile equating functions. Again, as pointed out earlier, the number of test takers whose pass or fail status may change could be large or small in absolute magnitude depending on how many test takers write this test for a particular user state.

Discussion and Conclusions

The purpose of the present study was to evaluate subgroup invariance of equating functions derived using the total group and two subgroups (i.e., Hispanic & non-Hispanic) for a large-scale certification test. Investigating subgroup invariance in equating relationships was

important for this test because the Hispanic group is a distinctively more able group as compared to the non-Hispanic group or the total group on the construct being measured, and the presence of this high-ability group may adversely affect the equating function derived using the total group. However, if subgroup invariance is shown to exist across the total group and for both the subgroups, then the presence of the two subgroups won't make any difference in pass/fail status of examinees.

Table 4

Actual Pass Percentages of Examinees Using Conversions Derived From the Three Groups (Form Y)

State	*N*	Chained linear			Chained equipercentile		
		Total	Hispanic	Non-Hispanic	Total	Hispanic	Non-Hispanic
State 1	154	70.13	70.13	70.13	70.13	70.13	70.13
State 2	72	80.56	80.56	80.56	79.17	79.17	80.56
State 3	83	89.16	89.16	89.16	85.54	85.54	85.54
State 4	390	88.46	88.46	88.46	87.69	87.69	88.46
State 5	314	54.78	54.78	54.78	54.78	54.78	54.78
State 6	58	79.31	79.31	79.31	79.31	79.31	79.31
State 7	79	59.49	59.49	62.03	59.49	59.49	59.49

The RESD was used to evaluate subgroup invariance of the equating functions. Overall, Form *X* results showed less subgroup invariance as compared to Form *Y* results. Form *X* findings are summarized as follows:

1. The actual standardized difference between the total group and non-Hispanic group linear equating functions showed negligible differences in the lower and middle areas of the cut score region and non-negligible differences in the higher cut score regions. However, the total group and Hispanic group equating functions showed negligible differences in the middle area of the cut score region and non-negligible differences

in the lower and higher areas of the cut score region. Results were similar for the chained equipercentile equating functions.

2. The RESD results for both the chained linear and chained equipercentile equating functions indicated a small but potentially non-negligible difference between the total group and non-Hispanic group equating functions but a large and potentially non-negligible difference between the total group and Hispanic group equating functions.
3. Finally, the pass percentages using the three different equating functions showed that for both the chained linear and chained equipercentile equating functions the pass percentages were more similar for the total group and non-Hispanic group equating functions as compared to the total group and Hispanic group equating functions.

These findings show lack of subgroup invariance for the total group and subgroups for Form *X*. Also based on the results, the Hispanic group equating appears to be less invariant as compared to the non-Hispanic group equating. Form *Y* findings are summarized next.

1. The actual standardized difference between the total group and non-Hispanic group linear equating functions showed negligible differences across all parts of the cut score region. Similarly, the total group and Hispanic group equating functions showed negligible differences across all parts of the cut score region. Results were similar for the chained equipercentile equating functions.
2. The RESD results for both the chained linear and chained equipercentile equating functions indicated a negligible difference between the total group and non-Hispanic group equating functions but a small but potentially non-negligible difference between the total group and Hispanic group equating functions.
3. Finally, the pass percentages using the three different equating conversions (total, Hispanic, and non-Hispanic groups) showed that for both the chained linear and chained equipercentile equating functions the pass percentages were identical across the three conversions for most cut score points.

As seen in the findings summarized above, the subgroup equating functions were less invariant for Form *X* as compared to the subgroup equating functions derived for Form *Y*. One possible reason for this difference in the results for Form *X* and Form *Y* is that the equating for

Form *X* was conducted using a smaller anchor set of items (i.e., 22 items) as compared to the equating of Form *Y*, which was conducted using a larger anchor set (i.e., 51 items). This may have led to some instability in the Form *X* equating. Although the anchor set for the Form *X* equating initially consisted of 28 items, 6 items were found to function quite differently for the new and reference form samples and therefore were dropped from the anchor set. Another possible reason for this difference in the results for Form *X* and Form *Y* is that Form *Y* has a much larger sample for the total group and subgroups as compared to Form *X* (see Tables 1 and 2). Since larger samples are more likely to produce more accurate equating results as compared to small samples under similar conditions, the differences that are noted for Form *X* may partly result from an equating error resulting from the small samples in Form *X*.[2] However, this supposition needs to be confirmed by additional research involving substantially larger sample sizes for the total group and subgroups. If other research with a larger sample also suggests that invariance holds well for different subgroups in large samples, then testing programs with fairly large sample sizes can be reasonably comfortable with the presence of subgroups in their equating sample. However, equating with smaller samples where distinct subgroups are present should be more carefully monitored for subgroup invariance. Since using statistical methods with small samples may often lead to misleading results, the decision regarding which group to use to conduct the equating can also be done on substantive grounds. For example, if a particular English language test is intended for test takers whose native language is not English, then excluding a subgroup that may be very proficient in English (e.g., native English language speakers) from the total group equating sample is reasonable and can be decided on substantive grounds alone. In such a case, the equating sample will always exclude the native English language test takers.

Another factor that has been pointed out by Green (2003) is that different forms often place slightly different emphasis on different parts of the test content and different groups of test takers often have differential exposure or interests to various aspects of the test content, which may result in different degrees of subgroup invariance across forms. Therefore, the authors suggest that the testing program should monitor such differences in content across these two forms to check for any consistent difference across different parts of the content.

Finally, it should be kept in mind that absence of differences as suggested by statistical indices can still lead to different pass/fail status of examinees if their scores are fairly close to the

cut or passing score. Therefore, absence of a statistical difference should not be used as the only indicator of invariance in equating. If the samples taking the test are fairly large, then even the slightest difference may affect pass/fail status of hundreds of examinees. On the other hand, large differences in equating functions across different parts of the score scale may have little impact on the pass/fail status when there are very few examinees in those parts or where the cut scores are further away from the points of difference of the equating functions. Therefore, both pass/fail rates and statistical indices should be used in conjunction to decide whether the invariance is practically important or not.

Since subgroups based on language and ethnicity are becoming more frequent in educational testing, and since teacher certification tests have extremely high stakes for test takers, it was important to examine whether the assumption of population invariance in equating may be compromised by the presence of markedly different subgroups in the total equating sample. Also including the actual pass/fail decisions based on the different equating functions was important because statistically significant results may not be important practically. On the other hand, for example, if the differences between two equating functions are found to be statistically insignificant but considerable differences in actual pass and fail decisions are observed by using either of the equating functions, then the statistical results may be less useful. As evident, the findings of this study suggest the lack of subgroup invariance in equating for the first test form with the smaller sample size. However, subgroup invariance in equating was observed for the second test form.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*(1), 15–32.

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

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

Green, B. F. (2003). Comments on population invariance of score linking. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR-03-27, pp. 127–130). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Program Statistics Research Tech. Rep. No. 87-79). Princeton, NJ: ETS.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York: Springer Science + Business Media.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65*(4), 437–456.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.

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. (2004, April). *Invariance of score linkings across gender groups for testlet-based CLEP*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Notes

[1] The actual standardized difference is simply the difference of two equating functions (i.e., Hispanic versus total) at each raw score level divided by the adjusted SD.

[2] To evaluate the effect of sampling error, one may consider using conditional standard errors of equating to evaluate differences between two equating functions. For example, if the difference between two equated scores at any score point is less than the (average) conditional standard error of equating at that score point, then one may argue that the difference is within the bounds of sampling error and therefore ignorable.