Research Report

First Language of Examinees and Its Relationship to Equating

Longjuan Liang

Neil J. Dorans

Sandip Sinharay

February 2009

ETS RR-09-05

First Language of Examinees and Its Relationship to Equating

Longjuan Liang, Neil J. Dorans, Sandip Sinharay

ETS, Princeton, New Jersey

February 2009

Abstract

To ensure fairness, it is important to better understand the relationship of language proficiency with the standard procedures of psychometric analysis. This paper examines how equating results are affected by an increase in the proportion of examinees who report that English is not their first language, using the analysis samples for a large-scale reading and mathematics test. The results indicate that equating is not affected much by an increase in the proportion of those examinees.

Key words: Fairness, language proficiency, equating

Acknowledgments

Any opinions expressed in this paper are those of the authors and not necessarily of ETS. The authors are grateful to Jennifer Zenna for her assistance with analyses of PSAT/NMSQT® data.

1. Research Question

Insufficient language proficiency might interfere with the measurement process. If some examinees do not possess the level of language proficiency needed to understand what the test questions are asking, the test may not measure what it is intended to measure for those examinees. Fair and valid measurement of the construct of interest may be adversely affected for these examinees when the construct of interest is not language proficiency itself. Including the examinees who are not proficient in English in an equating analysis might violate the basic equating requirement of construct similarity. Hence, it is important to better understand the relationship of language proficiency to the basic procedures used to ensure fairness. Ideally for this purpose, we would like to identify the population of test takers who possess at least the level of proficiency in English presumed to be necessary to provide a fair assessment of the construct of interest.

This report denotes this population as *sufficiently proficient in English* (SPE). However, very few large-scale testing programs collect information on the examinees' level of English proficiency. Instead, most testing programs ask examinees whether English is one of their first languages or English is one of their best languages—that information is inadequate to classify an examinee as SPE. In this paper, we are limited to studying *English first language* (EFL) and *not English first language* (NEFL) populations, instead of the SPE and *not sufficiently proficient in English* (NSPE) populations that are of real interest.

When equating procedures were initially implemented, the testing population was relatively homogeneous with respect to the test-takers' native language, which was English. As a consequence, equating was not likely to be affected by the inclusion of non-native speakers in the testing population, as this group constituted a small portion of the population. However, the composition of the U.S. population has changed since the above-mentioned practices were adopted. In particular, there has been an increase in the proportion of NEFL examinees. For example, *America's Perfect Storm* by Kirsch, Braun, Yamamoto, and Sum (2007) noted that, among other facts about a changing U.S. population, immigration has accounted for an increasing percentage of population growth over the past few decades and the Hispanic share of the U.S. population is expected to grow from 14% in 2005 to slightly more than 20% by 2030. As more non-native speakers take tests in English, the potential effects on equating are likely to grow in magnitude. Hence, there is a need to revisit the issue of choice of the examinee sample in equating.

Currently, there are no standard guidelines regarding how the NEFL examinees should be treated in different aspects of fairness procedures. Many tests remove the NEFL examinees from differential item functioning (DIF) analyses to prevent items from being flagged because of an examinee language issue, rather than a content issue. With equating, in contrast, the examinees for whom English was not their first language or best language are still included in the analysis. This phenomenon (of different policies regarding the choice of examinee sample for DIF and equating) was summarized in a survey of ETS tests (S. Sinharay & N. J. Dorans, personal communication, May 22, 2007). While the effects of language proficiency on fairness procedures, including both DIF and equating, should be addressed, this paper focuses on the sensitivity of equating to shifts in first language status of the samples employed. A companion report by Sinharay, Dorans, and Liang (in press) studies the effect of first language status on DIF analysis. The goal of this report is to examine whether the exclusion or inclusion of examinees from the analysis sample on the basis of first language status affects the results of equating and how equating results are affected by an increase in the proportion of NEFL examinees.

In order to achieve this goal, we examine data from a large sample of examinees obtained from the PSAT/NMSQT® (Preliminary SAT®/National Merit Scholarship Qualifying Test) administrations. In the PSAT/NMSQT, the examinees are asked, during the examination, about the language they first learned to speak. They can answer (a) *English*, (b) *English and another*, and (c) *Another*. However, they need not answer the question at all. Those examinees who choose either the first option or the second option to the question are referred to as EFL examinees. Those who choose the third option to the question are henceforth referred to as NEFL examinees. The PSAT/NMSQT equating for a new form of the test is performed on all sophomores and juniors (the Total group), regardless of how they answer the first language question. The reference form in each PSAT/NMSQT equating is an SAT form; the sample from this form is restricted to juniors and seniors regardless of how they answer the first language question. In this study, equating analyses are run on EFL and NEFL subgroups as well. The conversions are then compared to the one obtained from the Total group.

It is important to consider what happens if the proportion of NEFL examinees increases in the examinee population, especially in the light of the above-mentioned findings of Kirsch et al. (2007). Hence, we proceed to study the sensitivity of the equating results to the proportion of NEFL examinees by creating synthetic samples from the available data sets and running equating analyses on these synthetic samples. A synthetic sample is created by combining all the available NEFL examinees with a simple random sample of the EFL examinees, so as to have a specified proportion of NEFL examinees in the total.

Section 2 describes the data that are used. Section 3 describes our methods, including the analyses on the observed data and those on the synthetic samples. Section 4 reports the comparison of conversions across the EFL group, the NEFL group, and the Total group. This was done for both the observed sample and the synthetic samples. Section 5 provides discussion and conclusions.

2. Description of the Data

PSAT/NMSQT data are used for our analyses because the NEFL examinees comprise a significantly large and growing portion of the examinee population for this test. Two sections, critical reading and math, from two PSAT/NMSQT forms (one administered on Wednesday and one administered on Saturday), are included in the analyses. For the rest of this report, WCR is used to designate the critical reading section of the Wednesday PSAT/NMSQT form. Similarly, SCR designates the critical reading section of the Saturday form, WMA designates the mathematics section of the Wednesday form, and SMA designates the mathematics section of the Saturday form. Operationally, the WCR and SCR sections were equated to the critical reading sections of two SAT parent forms through 27 common items respectively. Similarly, the WMA and SMA sections were operationally equated to the mathematics sections of the same two SAT forms through 22 common items respectively. The PSAT/NMSQT writing section was excluded from this research because of its small sample size.

Tables 1 to 4 contain the number of examinees, means and standard deviations of the raw scores on the PSAT/NMSQT forms, and their corresponding parent or reference SAT forms. Examination of Tables 1 and 2 reveals that for the critical reading sections, the NEFL groups have lower means than the EFL groups for both Wednesday and Saturday PSAT/NMSQT forms and also for each of the parent SAT forms. This suggests that the NEFL group is consistently less able than the EFL group on the critical reading tests. However, this relationship seems to be the opposite for the mathematics tests. From Tables 3 and 4, the NEFL group outperformed the EFL group on the two SAT forms and on the PSAT/NMSQT SMA form. Only on the PSAT/NMSQT WMA form did the EFL group score higher. This suggests that exclusion or inclusion of the NEFL group in the equating could have different effects on the critical reading and mathematics sections of the PSAT/NMSQT.

Table 1

Basic Statistics for the PSAT/NMSQT Wednesday Critical Reading Section (WCR) Test and the Reference SAT Test

		Total [a]	EFL	NEFL	NEFL% [b]
New form:	*N*	277,594	244,431	23,984	9.01
PSAT/NMSQT	Mean	16.50	17.09	12.78	
WCR	SD	10.97	10.89	10.57	
Reference form:	*N*	98,467	79,863	8,313	9.43
SAT	Mean	33.51	33.98	26.52	
	SD	17.06	16.69	17.56	

Note. EFL = English first language, NEFL = not English first language.

[a] Total group includes EFL examinees, NEFL examinees, and those who didn't respond to the EFL question. [b] NEFL% = NEFL/(EFL + NEFL).

Table 2

Basic Statistics for the PSAT/NMSQT Saturday Critical Reading Section (SCR) Test and the Reference SAT Test Form

		Total [a]	EFL	NEFL	NEFL% [b]
New form:	*N*	154,371	144,570	7,963	5.22
PSAT/NMSQT	Mean	21.78	21.89	20.28	
SCR	SD	9.82	9.71	11.30	
Reference form:	*N*	100,902	81,840	8,489	9.40
SAT	Mean	35.37	35.97	27.14	
	SD	17.04	16.63	17.69	

Note. EFL = English first language, NEFL = not English first language.

[a] Total group includes EFL examinees, NEFL examinees, and those who didn't respond to the EFL question. [b] NEFL% = NEFL/(EFL + NEFL).

Table 3

Basic Statistics for PSAT/NMSQT Wednesday Mathematics Section (WMA) Test and the Reference SAT Test Form

		Total [a]	EFL	NEFL	NEFL% [b]
New form:	*N*	277,594	244,431	23,984	9.01
PSAT/NMSQT	Mean	13.84	14.15	12.76	
WMA	SD	9.79	9.67	10.53	
Reference form:	*N*	98,467	79,863	8,313	9.43
SAT	Mean	28.01	27.65	28.57	
	SD	13.93	13.53	15.87	

Note. EFL = English first language, NEFL = not English first language.

[a] Total group includes EFL examinees, NEFL examinees, and those who didn’t respond to the EFL question. [b] NEFL% = NEFL/(EFL + NEFL).

Table 4

Basic Statistics for PSAT/NMSQT Saturday Mathematics Section (SMA) Test and the Reference SAT Test Form

		Total [a]	EFL	NEFL	NEFL% [b]
New form:	*N*	154,371	144,570	7,963	5.22
PSAT/NMSQT	Mean	19.65	19.58	21.56	
SMA	SD	8.60	8.48	10.16	
Reference form:	*N*	100,902	81,840	8,489	9.40
SAT	Mean	27.22	26.89	27.46	
	SD	13.90	13.51	15.94	

Note. EFL = English first language, NEFL = not English first language.

[a] Total group includes EFL examinees, NEFL examinees, and those who didn’t respond to the EFL question. [b] NEFL% = NEFL/(EFL + NEFL).

3. Methods

3.1 Analysis of the Observed Data

Operationally, the unsmoothed chained equipercentile method was used to equate the two PSAT/NMSQT forms to their two parent SAT forms respectively, using the Total group. In this study, the same equating method was used for both the EFL and the NEFL groups. The scores of the EFL group of sophomores and juniors on the PSAT/NMSQT test were equated to the scores of the EFL group of juniors and seniors on the SAT test, and likewise the scores of the NEFL group on the PSAT/NMSQT test were equated to the scores of the NEFL group on the SAT test. The conversion based on the EFL group was then compared to the conversion based on the Total group, and the conversion based on the NEFL group was also compared to the conversion based on the Total group. To compare any two conversions, we used the root expected square difference (RESD) index (Dorans, Cahn, Jiang, & Liu, 2006),

$$RESD = \sqrt{\sum_m f_m \left[s_1(m) - s_2(m) \right]^2} \ ,$$

where m represents each raw score level. The quantity $s_1(m)$ represents the scaled score at raw score level m from Conversion 1 and $s_2(m)$ represents the scaled score at score level m from Conversion 2. The weight f_m is the relative frequency at score level m and is used so that scaled scores with higher frequencies receive larger weights. A small value of RESD indicates a negligible difference.

3.2 Analysis of the Synthetic Samples

The above analyses, after being performed on each of the data sets, provided results for the currently observed percentage of NEFL examinees, but did not indicate how equating results would differ if the percentage of NEFL examinees increased to another value. To go beyond this limitation, synthetic samples with varying percentages of NEFL examinees were created. Since the proportion of NEFL examinees was only around 9% for the Wednesday administration and only around 5% for the Saturday administration, all NEFL examinees were retained in these synthetic samples. Simple random samples of different sizes were drawn from the EFL examinees and combined with the NEFL examinees to form the synthetic samples. Our goal was to determine if a threshold could be found so that if the proportion of NEFL examinees goes beyond this value, the equating results will be significantly affected.

To simulate a consistent but gradual growth in the NEFL proportions, the NEFL proportion for the new form (PSAT/NMSQT) was chosen to be 5% higher than that of the reference form (SAT). The composition of the different synthetic samples is described in Table 5. The table shows the ratio of percentages of EFL and NEFL examinees in the new form and the reference form in each synthetic sample.

Synthetic Sample 9 in Table 5, in contrast to the other samples, involves equating with a new form sample of 50% EFL examinees and 50% NEFL examinees and a reference form sample of 90% EFL examinees, which creates a large difference in proportions of NEFL examinees for the new form and the reference form.

Table 5

Composition of the Synthetic Samples

Synthetic sample	1	2	3	4	5	6	7	8	9
New form (EFL:NEFL)	85:15	80:20	75:25	70:30	65:35	60:40	55:45	50:50	50:50
Reference form (EFL:NEFL)	90:10	85:15	80:20	75:25	70:30	65:35	60:40	55:45	90:10

Note. EFL = English first language, NEFL = not English first language.

4. Results

4.1 Results for the Observed Samples

Figure 1 shows the differences in the equated scaled scores for the EFL, NEFL, and the Total groups for the critical reading and mathematics sections of the two PSAT/NMSQT forms. The two horizontal lines shown in Figure 1 correspond to the difference that matters (DTM) criterion (Dorans, Holland, Thayer, & Tateneni, 2003). In general, the DTM is considered to be half of a score unit for unrounded scores. The DTM is 0.5 in this study, since the reporting scale of the PSAT/NMSQT is 20–80 with 1 point increments. Any difference whose absolute value is less than the DTM is negligible for many practical purposes.

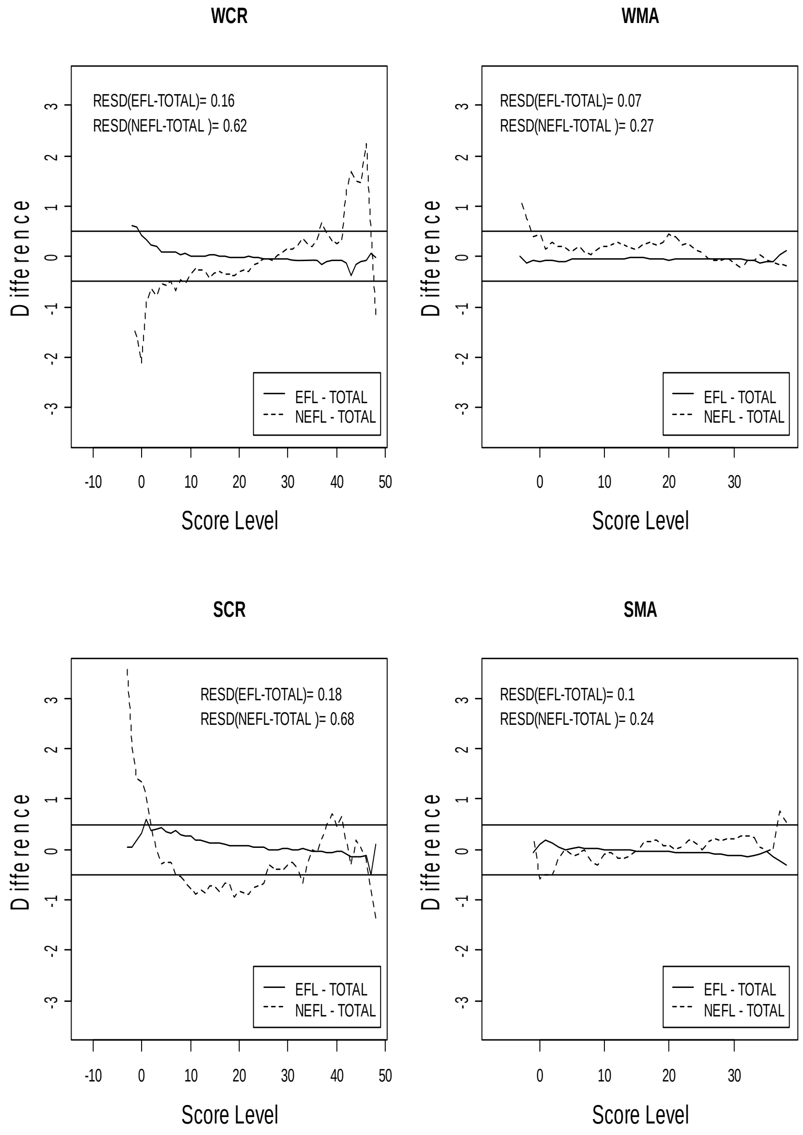

***Figure 1.* Differences in the equated scaled scores for the English first language (EFL) , not English first language (NEFL), and Total groups.**

Note. RESD = root expected square difference, SCR = Saturday critical reading section, SMA = Saturday mathematics section, WCR = Wednesday critical reading section, WMA = Wednesday mathematics section.

The four panels in Figure 1 show the results for the two sections from each of the two PSAT/NMSQT forms (WCR, WMA, SCR, SMA). The solid lines in all four panels represent the difference in the scaled scores based on the EFL group and the Total group. The dotted lines represent the difference in the scaled scores based on the NEFL group and the Total group. For the math sections, the conversion based on the EFL group is not much different from that based on the Total group. The absolute differences of the two conversions, as illustrated by the solid lines in the two right panels in Figure 1, are always less than the DTM. For the critical reading sections, the comparison of the conversions based on the EFL group and the Total group (as illustrated by the solid lines in the two left panels in Figure 1) shows that the differences at most score levels are still well within the DTM bandwidth, with only a few exceptions at the left tails. The differences for the critical reading sections seem to be a little larger than those for the math sections. These results suggest that for the data we analyzed, using the Total group or the EFL group would yield the same equating conversions, except for rounding. This is partly due to the large proportion of EFL examinees in the Total group. Tables 1 and 2 show that the Total group for the PSAT/NMSQT includes around 91% EFL examinees for the Wednesday form and around 95% for the Saturday form.

The differences in the conversions based on the NEFL group and on the Total group are also plotted and are shown as the dotted lines in all four panels in Figure 1. The two right panels show that for the math sections, only a few values at the tails are larger than the DTM. However, the two left panels show that the differences are larger for the critical reading sections, especially for the Saturday form, where the large differences occur not only at both the tails (where the number of examinees is small) but also in the middle of the score distribution (where the number of examinees is large). This suggests that the population invariance assumption for equating is violated on the critical reading sections on both forms.

Tables 6 to 9 show the extent to which the summary statistics of the equated scores of the EFL and NEFL groups are affected by whether the Total-group conversion or the subgroup-specific conversion is used to produce scaled scores. In each of these tables, TGL represents the total-group conversion and SGL represents a subgroup conversion. For the EFL group, SGL means that the conversion is based on the EFL group only; for the NEFL group, SGL means that the conversion is based on the NEFL group only. There are three blocks of rows in each of these tables corresponding to the statistics for the Total group, the EFL group, and the NEFL group. Columns 4–5 show the mean and SD of the equated scores of the groups for different conversions. For example, in the second block in Table 6, the results indicate that the EFL group has a mean of 45.30 if the EFL-only conversion is used instead of the Total-group conversion, which produces a mean of 45.27. The last four columns are some statistics based on the comparison of two conversions. For example, the percentage of the equated scaled score (ESS) differences (DIFF) that are greater than or equal to the DTM (% ESS |DIFF| > = 0.5), and the percentage of examinees with a difference this big (% Examinees |DIFF| > = 0.5) are shown in the last two columns.

The second row blocks of these tables indicate that the use of the EFL-only conversion or the Total-group conversion has little effect on the EFL examinees for all the forms. For example, the mean difference for the EFL group is 0.03 for the WCR test, 0.09 for the SCR test, and -0.05 for both the WMA and the SMA tests. The RESD indices in the four tables are all smaller than the DTM, with the differences for the math tests being smaller than those for the critical reading tests. The RESD for the EFL group is 0.07 for the WMA test, 0.10 for the SMA test, 0.16 for the WCR test, and 0.18 for the SCR test. The far right column in each table also shows that the EFL examinees are affected very little by the choice of the conversion. Only 1.8% of the examinees on the WCR test, 0.4% of the examinees on the SCR test, and no examinees (0%) on both the WMA and the SMA tests have absolute differences of the equated scaled scores that are greater than 0.5.

Table 6

Effect of First Language Status on Equating of the Wednesday Critical Reading Section (WCR) Test

Group	*N*	Linking	Mean	SD	Mean diff.	RESD	% ESS |DIFF| >=0.5 [a]	% examinees |DIFF| >=0.5
Total	277,594	TGL	44.66	11.47				
EFL	244,431	TGL	45.27	11.35	0.03	0.16	3.9%	1.8%
		SGL	45.30	11.25				
NEFL	23,984	TGL	40.77	11.32	-0.41	0.62	33.3%	36.8%
		SGL	40.36	11.72				

Note. DIFF = differences, EFL = English first language, ESS = equated scale score, NEFL = not English first language, RESD = root expected square difference, SGL = subgroup conversion, TGL = total-group conversion.

[a] % ESS |DIFF| > = 0.5 represents the percentage of absolute differences in equated scaled scores that are greater than or equal to the difference that matters (DTM) criterion.

Table 7

Effect of First Language Status on Equating of the Saturday Critical Reading Section (SCR) Test

Group	*N*	Linking	Mean	SD	Mean Diff	RESD	% ESS |DIFF| >= 0.5	% examinees |DIFF| >= 0.5
Total	154,371	TGL	49.32	10.33				
EFL	144,570	TGL	49.43	10.20	0.09	0.18	1.9%	0.4%
		SGL	49.52	10.09				
NEFL	7,963	TGL	47.83	12.13	-0.44	0.68	55.8%	62.4%
		SGL	47.40	12.15				

Note. DIFF = differences, EFL = English first language, ESS = equated scale score, NEFL = not English first language, RESD = root expected square difference, SGL = subgroup conversion, TGL = total-group conversion.

[a] % ESS |DIFF| >= 0.5 represents the percentage of absolute differences in equated scaled scores that are greater than or equal to the difference that matters (DTM) criterion.

Table 8

Effect of First Language Status on Equating of the Wednesday Mathematics Section (WMA) Test

Group	*N*	Linking	Mean	SD	Mean diff.	RESD	% ESS |DIFF| > = 0.5 [a]	% examinees |DIFF| > = 0.5
Total	277,594	TGL	44.75	11.73				
EFL	244,431	TGL	45.11	11.55	-0.05	0.07	0%	0%
		SGL	45.06	11.56				
NEFL	23,984	TGL	43.57	12.79	0.18	0.27	4.8%	3.0%
		SGL	43.75	12.68				

Note. DIFF = differences, EFL = English first language, ESS = equated scale score, NEFL = not English first language, RESD = root expected square difference, SGL = subgroup conversion, TGL = total-group conversion.

[a] % ESS |DIFF| > = 0.5 represents the percentage of absolute differences in equated scaled scores that are greater than or equal to the difference that matters (DTM) criterion.

Table 9

Effect of First Language Status on Equating of the Saturday Mathematics Section (SMA) Test

Group	*N*	Linking	Mean	SD	Mean diff.	RESD	% ESS |DIFF| > = 0.5 [a]	% examinees |DIFF| > = 0.5
Total	154,371	TGL	51.49	10.16				
EFL	144,570	TGL	51.40	9.99	-0.05	0.10	0%	0%
		SGL	51.35	9.94				
NEFL	7,963	TGL	53.94	12.41	0.08	0.24	10.0%	7.6%
		SGL	54.02	12.58				

Note. DIFF = differences, EFL = English first language, ESS = equated scale score, NEFL = not English first language, RESD = root expected square difference, SGL = subgroup conversion, TGL = total-group conversion.

[a] % ESS |DIFF| > = 0.5 represents the percentage of absolute differences in equated scaled scores that are greater than or equal to the difference that matters (DTM) criterion.

The third row blocks of Tables 6–9 show that the use of the NEFL-only conversion or the Total-group conversion has a larger effect on the NEFL group, especially for the critical reading tests. Use of the NEFL-only conversion in place of the Total-group conversion results in lower means for the NEFL group on the two critical reading sections, but higher means on the math sections. For example, the mean differences are -0.41 for the WCR test and -0.44 for the SCR test, but are 0.18 for the WMA test and 0.08 for the SMA test. The RESD indices are greater than the DTM for both critical reading sections (0.62 and 0.68 for WCR and SCR respectively) but are smaller than the DTM for both math sections (0.27 and 0.24 for WMA and SMA respectively). A large proportion of the NEFL examinees will be affected on the critical reading sections if NEFL-only conversion is used instead of Total-group conversion (36.8% for the WCR test and 62.4% for the SCR test), but a much smaller proportion of the NEFL examinees will be affected on the math tests (3.0% for the WMA test and 7.6% for the SMA test).

The above analyses were based on the observed population only. It is informative to see how big the differences would be with a growing NEFL group. Results from the synthetic samples, shown next, will address this question.

4.2 Results for Synthetic Samples

Figure 2 plots the RESD values for the synthetic samples for varying NEFL proportions. The horizontal lines in this figure correspond to the DTM criterion. Note that the X-axis is the NEFL proportion for the new form; the corresponding NEFL proportion for the reference form can be found in Table 5. Also note that the RESD is a summary statistic across all score levels. So a small RESD does not necessarily mean that the differences of two conversions are small at all score levels. To investigate further, Figures 3 to 6 plot the differences in equated scaled scores for EFL and Total and NEFL and Total at the individual score levels. Note that each panel in Figures 3 to 6 corresponds to a synthetic sample described in Table 5. For example, the top left panel in Figure 3, which shows results for the WCR test, corresponds to the synthetic sample for which the proportion of NEFL examinees is 0.15 for the new form and 0.10 for the reference form.

The circles in Figure 2 represents the RESD values calculated from the differences in the EFL-only conversion and the Total-group conversion. The circles show an increasing trend in general, with a few exceptions, in all four panels. This suggests that the difference between the EFL-only conversion and the Total-group conversion becomes larger in general as the NEFL proportion increases. But even when the NEFL proportion for the new form increases to 50%, the largest RESD value (Sample 8 for WCR test) only slightly exceeds the DTM value. Figures 3 to 6 provide detailed information on how the conversions change as the NEFL percentage increases.

The solid lines in these figures suggest that although the RESD values for most samples in all forms we tested are less than the DTM, the differences at individual score levels could still be large, especially for extreme scores.

The triangles in Figure 2 represents the RESD values calculated from the differences in the NEFL-only conversion and the Total-group conversion. The RESD values show a decreasing trend for the WCR and SCR tests. This is expected, since the extent of overlap in equating samples increases as the NEFL proportion increases. The RESD values for the WMA and SMA tests are in general small and are relatively stable as the NEFL proportion increases. This suggests that the math sections are less affected by different compositions of EFL and NEFL examinees in the equating sample than are the critical reading sections.

The solid circles and triangles are the RESD values for the Synthetic Sample 9 in Table 5, where the new form has 50% of NEFL examinees but the reference form has only 10% NEFL examinees. Although this sample combination has the largest difference in the NEFL proportions between the new form and the reference form, the differences in the two conversions are not much different from that for the synthetic sample where the reference form has 45% of NEFL examinees (Synthetic Sample 8), except for the WCR test where the RESD for Synthetic Sample 9 dropped from 0.51 (for Synthetic Sample 8) to 0.10.

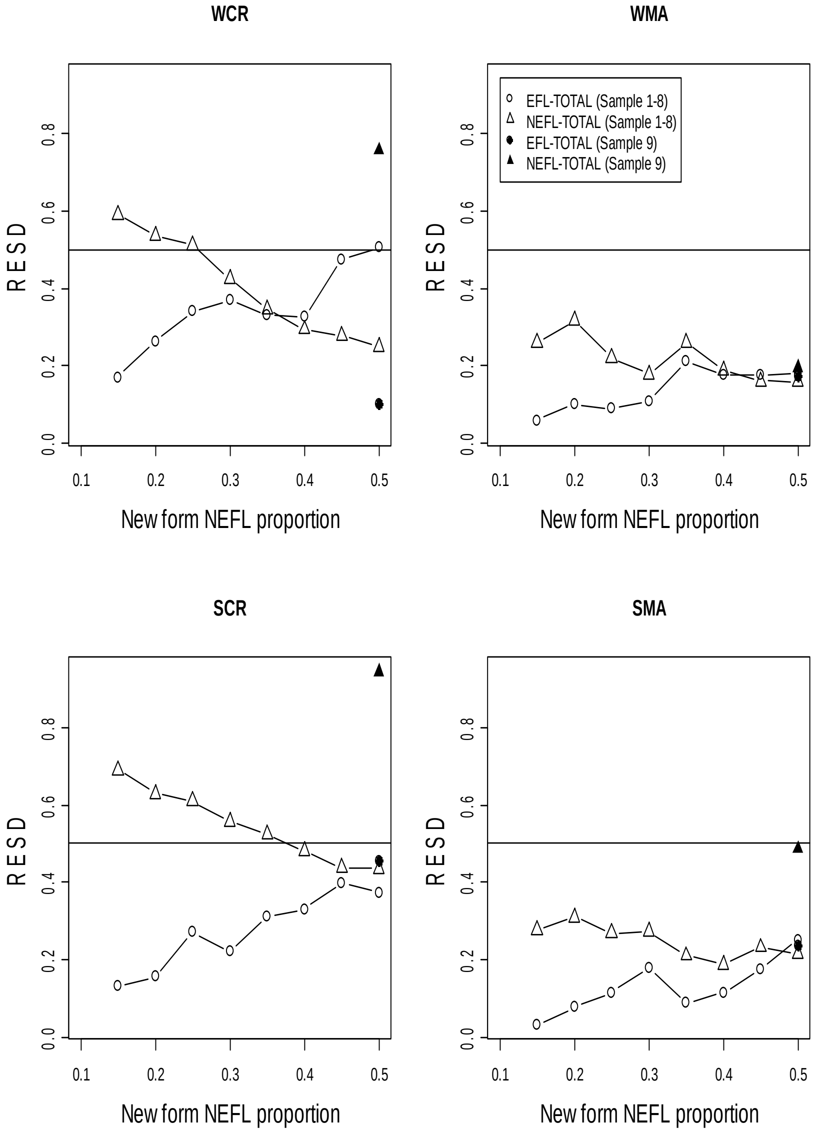

***Figure 2.* Root expected square difference (RESD) for synthetic samples for varying not English first language (NEFL) proportions.**

Note. EFL = English foreign language, RESD = root expected square difference, SCR = Saturday critical reading section, SMA = Saturday mathematics section, WCR = Wednesday critical reading section, WMA = Wednesday mathematics section.

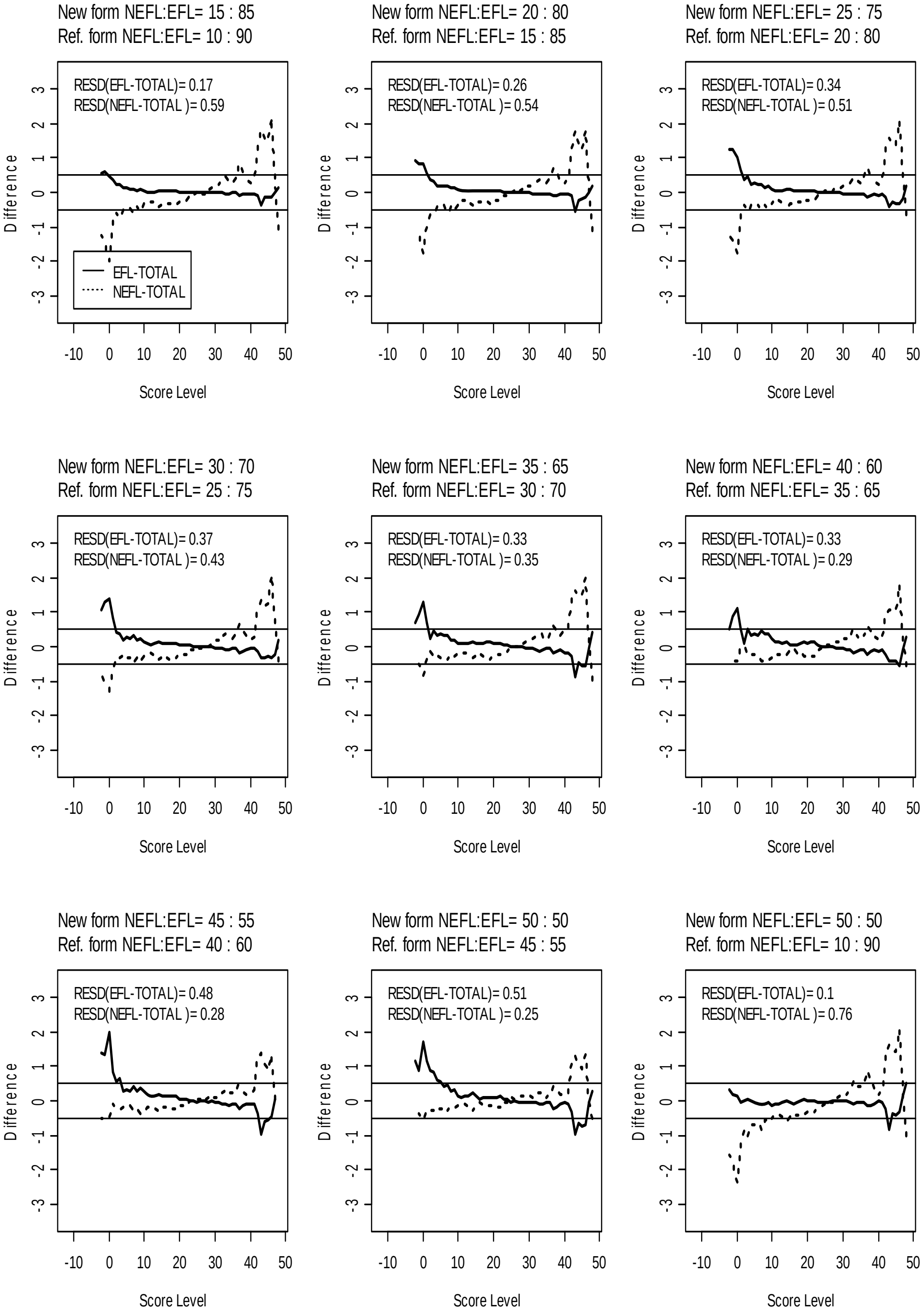

***Figure 3.* Differences of equated scaled scores for the Wednesday critical reading section (WCR) test.**

Note. EFL = English foreign language, NEFL = not English foreign language, Ref = reference, RESD = root expected square difference.

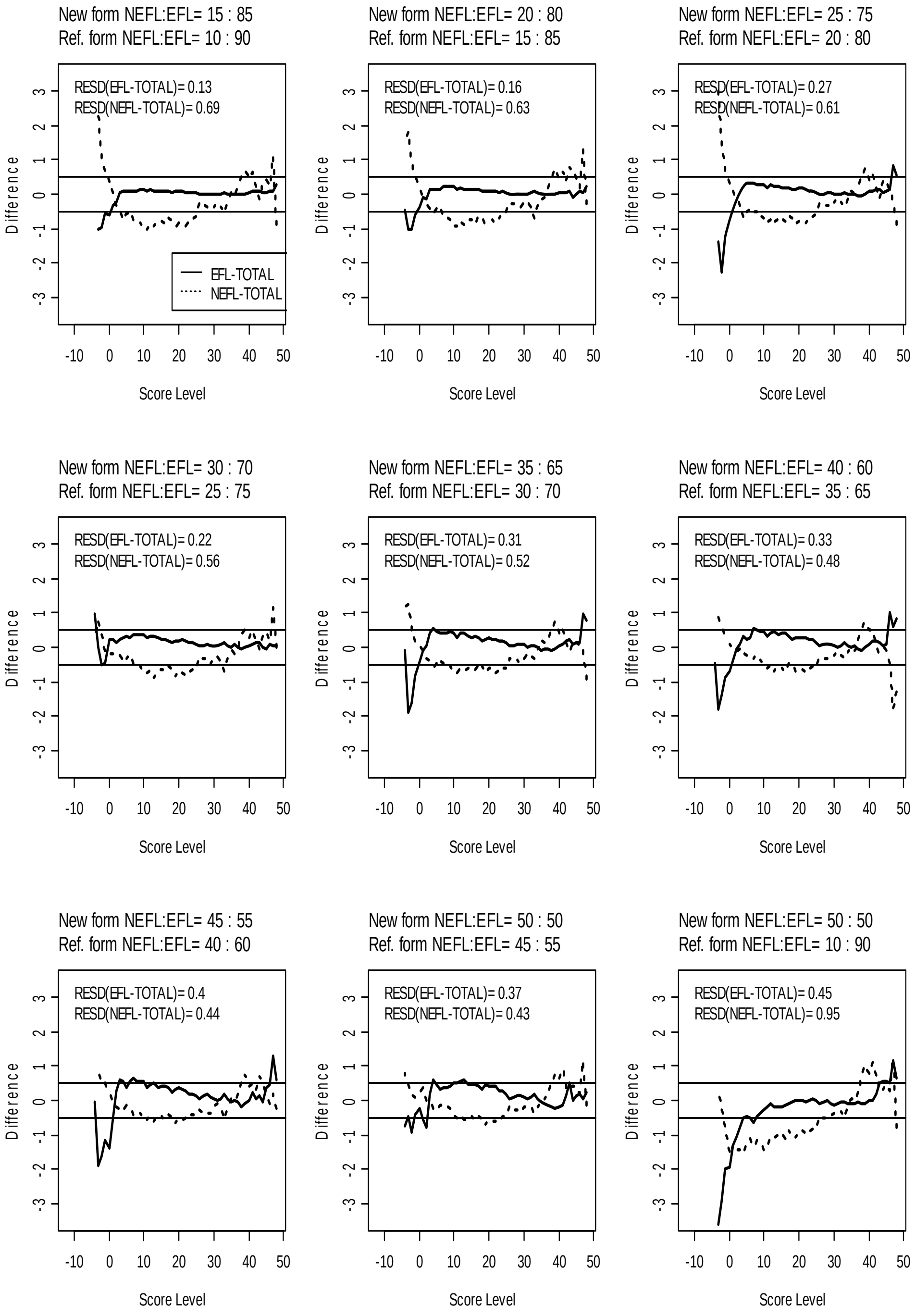

***Figure 4.* Differences of equated scaled scores for Saturday critical reading section (SCR) test.**

Note. EFL = English foreign language, NEFL = not English foreign language, Ref = reference, RESD = root expected square difference.

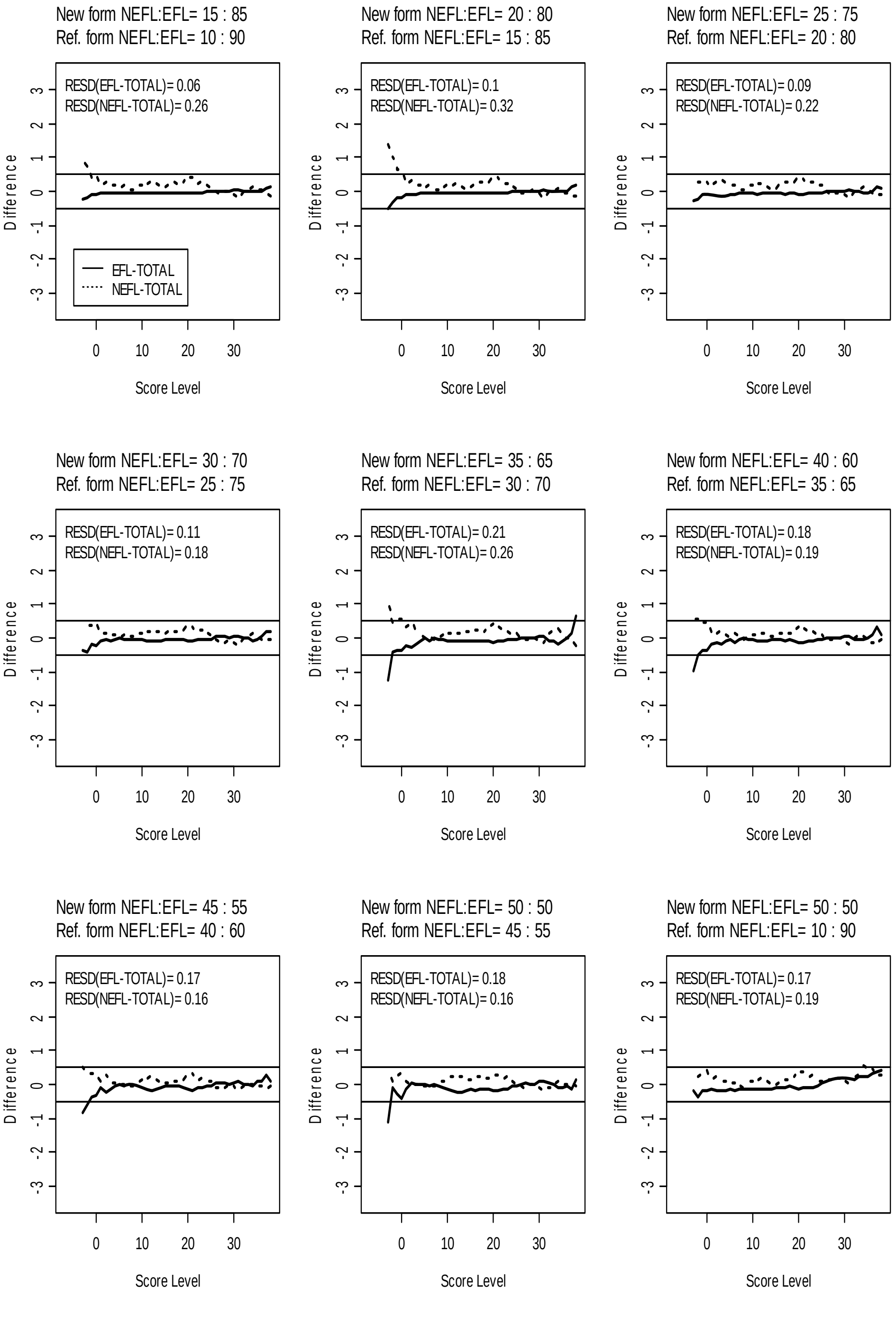

***Figure 5.* Differences of equated scaled scores for Wednesday mathematics section (WMA) test.**

Note. EFL = English foreign language, NEFL = not English foreign language, Ref = reference, RESD = root expected square difference.

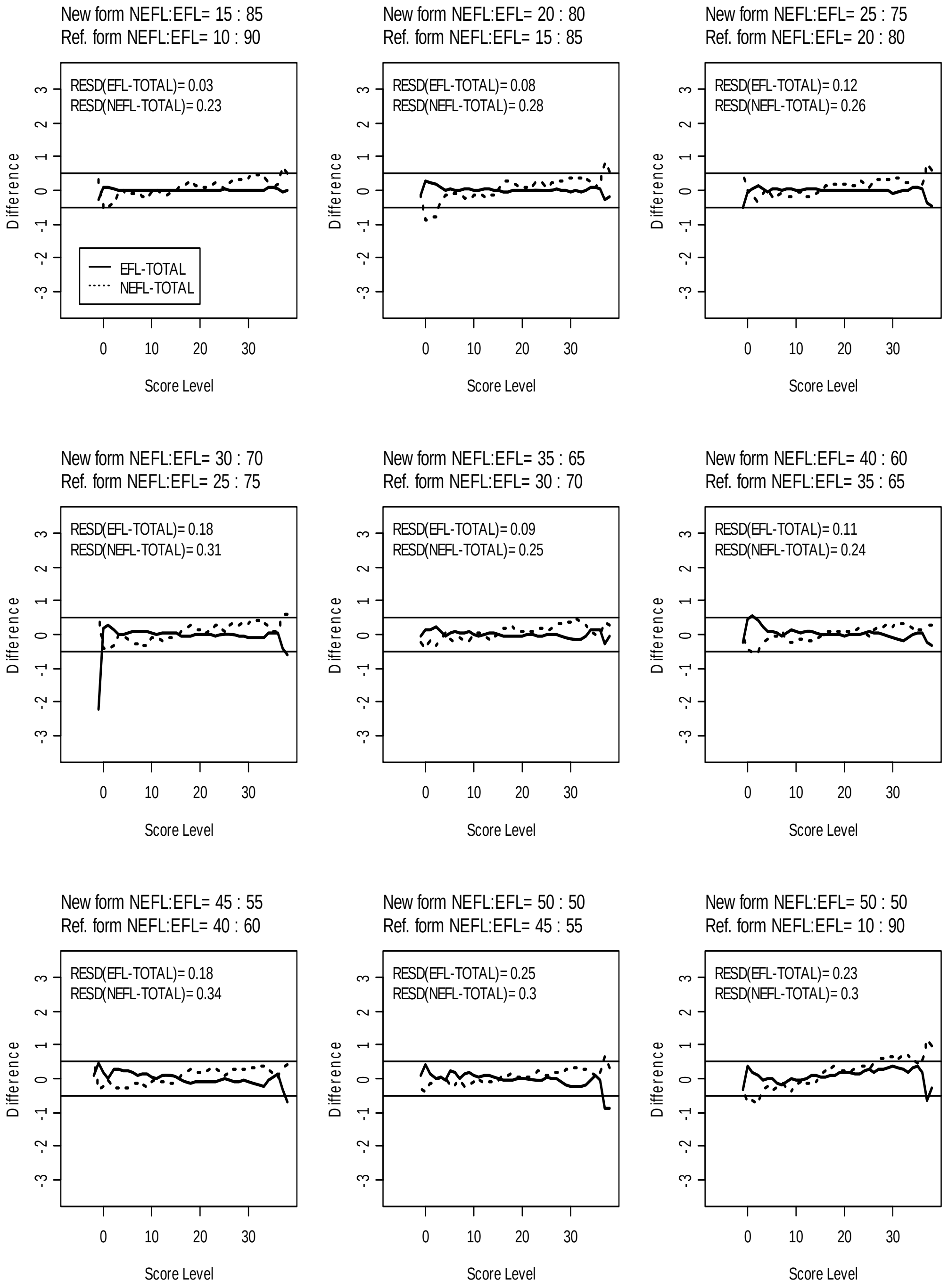

Figure 6. **Differences of equated scaled scores for Saturday mathematics section (SMA) test.**

Note. EFL = English foreign language, NEFL = not English foreign language, Ref = reference, RESD = root expected square difference.

Comparing the two left panels with the two right panels in Figure 2, we found that exclusion or inclusion of the NEFL examinees in equating has larger effects on the critical reading sections than on the math sections. This result is understandable. Since first language status is correlated with the critical reading scores, removing the NEFL group from the Total group should change the frequency distribution of the critical reading scores substantially. The comparison of the means for the EFL group and the NEFL group on the two critical reading sections and the two math sections in Tables 1 to 4 also shows that the EFL and NEFL groups are much more similar in ability on the math test than on the critical reading test. The more similar the two groups are, the less effect mixing the two groups is likely to have.

5. Discussion and Conclusions

Results from this study showed that from an operational point of view, when averaged across all score levels, it didn’t matter whether the equating sample was the EFL group or the Total group. This was particularly true for the math section. For the critical reading section, it was only when the NEFL proportion increased to about the same size as the EFL proportion that the equating results started to appear potentially worrisome. However, at certain score levels, especially for the low scores and high scores, there could be large differences between the equating conversion based on the EFL group and that based on the Total group.

A limitation of this study is that the criterion used to categorize examinees as EFL or NEFL was the examinees’ self-report on the question whether English was their first language and not a direct measure of their proficiency. The response to this EFL question provides ready-to-use information, because it is garnered from a question printed in the test form. The EFL question, however, is not likely to be an accurate measure of examinees’ true English proficiency status. For example, those who were born in another country but came to the United States at a very young age may indeed be proficient in English, even though English is not their first language.

English as best language (EBL) in some sense may be considered a better measure of examinees’ language proficiency level, because it directly asks about their language proficiency level. However, this index has deficiencies as well. For example, some bilingual examinees might claim the other language as their best language although they are proficient enough in English to understand the test instructions and questions. The above analyses were also run using EBL as a classification criterion and very similar results were obtained. In the future, we hope to remedy this limitation by studying a testing program that includes a more direct measure of English proficiency.

References

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., 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.

Kirsch, I., Braun. H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future*. Princeton, NJ: ETS.

Sinharay, S., Dorans, N. J., & Liang, L. (in press). *First language of examinees and its relationship to differential item functioning.* Princeton, NJ: ETS.