Research Report

First Language of Examinees and Its Relationship to Differential Item Functioning

Sandip Sinharay

Neil J. Dorans

Longjuan Liang

March 2009

ETS RR-09-11

First Language of Examinees and Its Relationship to Differential Item Functioning

Sandip Sinharay, Neil J. Dorans, and Longjuan Liang
ETS, Princeton, New Jersey

March 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

Abstract

To ensure fairness, it is important to better understand the relationship of language proficiency to standard psychometric analysis procedures. This paper examines how results of differential item functioning (DIF) analysis are affected by an increase in the proportion of examinees who report that English is not their first language in the analysis sample of a large-scale assessment. The results vary by group. In some combinations of focal/reference groups, the magnitude of DIF is not appreciably affected by whether the DIF is performed on examinees whose first language is not English. In other groups, the first language status matters. The results vary by type of test as well. In addition, the magnitude of DIF for some items is substantially affected by whether the DIF is performed on examinees whose first language is not English.

Key words: Fairness, Mantel-Haenszel, standardization

Acknowledgments

Any opinions expressed in this paper are those of the authors and not necessarily of ETS. The authors are thankful to Daniel Eignor, Hyeon-Joo Oh, Skip Livingston, Michael Zieky, and Shelby Haberman for their helpful comments and to Kim Fryer for her editorial help. The authors are grateful to Jennifer Zenna for her assistance with analyses of PSAT/NMSQT® data.

Table of Contents

Page

1. Research Question 1
    1.1 Differential Item Functioning (DIF) Procedures 1
    1.2 The Impact of Language Status and Examinee Groups 2
2. Description of the Data 3
3. Methods 6
    3.1 Differential Item Funtioning (DIF) Analyses on the Observed Population 12
    3.2 Differential Item Functioning (DIF) Analyses on the Synthetic Populations 13
4. Results: Impact of First Language Status on PSAT/NMSQT Data 13
    4.1 Differential Item Functioning (DIF) Results for the Observed Population 13
    4.2 Differential Item Functioning (DIF) Results for Synthetic Populations 23
5. Discussion and Conclusions 31
References 34
Appendix 36

List of Tables

Page

Table 1. Percentage of Not English First Language (NEFL) Examinees in the Different Examinee Groups 4
Table 2. Statistics for Saturday Critical Reading 6
Table 3. Statistics for Wednesday Critical Reading 7
Table 4. Statistics for Saturday Mathematics 8
Table 5. Statistics for Wednesday Mathematics 9
Table 6. Statistics for Saturday Writing 10
Table 7. Statistics for Wednesday Writing 11
Table 8. Cross-Classification of Differential Item Functioning (DIF) Categorizations for the PSAT/NMSQT Data 14
Table 9. Statistics for English First Language (EFL) and English Not First Language (NEFL) for the PSAT/NMSQT Test Forms 22

List of Figures

Page

Figure 1. Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday critical reading. 16

Figure 2. Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday critical reading. 17

Figure 3. Plot of the Mantel-Haenszel differential diffculty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday mathematics. 18

Figure 4. Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday mathematics..... 19

Figure 5. Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday writing. 20

Figure 6. Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday writing 21

Figure 7. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, female/male..... 25

Figure 8. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Asian American/White. 26
Figure 9. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Black/White. 27
Figure 10. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Hispanic/White. 28
Figure 11. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, Asian American/White. 29
Figure 12. Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, Asian American/White. 30

1. Research Question

Insufficient language proficiency might interfere with the measurement process. If an examinee does not possess the degree of language proficiency needed to understand what a question is asking, fair and valid measurement of the construct of interest may be adversely affected for that examinee, provided the construct of interest is not language proficiency itself. 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. Let us denote this population as sufficiently proficient in English (SPE). However, very few large-scale testing programs collect information on the examinees' English proficiency. Instead, most testing programs ask examinees if English is one of their first or one of their best languages. That information is inadequate to determine if an examinee is SPE.

Hence, in this paper, we are limited to studying English first language (EFL) and not English first language (NEFL) populations instead of SPE and not sufficiently proficient in English (NSPE) populations that are of real interest.

1.1 Differential Item Functioning (DIF) Procedures

Differential item functioning (DIF) procedures have been in place for more than two decades. Equating procedures have been in place for an even longer period. When these procedures were initially implemented, the testing populations were mostly homogeneous with respect to their native language, English. As a consequence, equatings were not likely to be affected by the inclusion of nonnative speakers in the testing population. Hence, equatings were usually performed using the full examinee sample.

In contrast, DIF appears to have been sensitive to language proficiency. Dorans and Kulick (1986) cited a study that used the self-reported English best language (EBL) categorizations, which are highly correlated with English proficiency, to explain what appeared to be many DIF items on the SAT® mathematics section for Asian Americans. As a result of that study, DIF analyses for several tests were limited to examinees who reported that English was their first language to prevent many items from being flagged for DIF because of an examinee language issue, not a content issue. 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, ETS personal communication, May 22, 2007).

However, the composition of the U.S. population has changed since the above-mentioned practices were adopted—most notably, there has been an increase in the proportion of nonnative examinees—and it will keep changing. For example, *America's Perfect Storm* by Kirsch, Braun, Yamamoto, and Sum (2007), among other facts about a changing U.S. population, noted that immigration has accounted for an increasingly large fraction of U.S. population growth over the past few decades and that 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 and more nonnative examinees (most of whom do not have English as their first language) take tests in English, the potential effects on DIF analyses are likely to grow in magnitude. For these reasons, revisiting the issue of choice of the examinee sample in DIF is needed.

1.2 The Impact of Language Status and Examinee Groups

A primary goal of this paper is to understand whether exclusion or inclusion of examinees from the analysis sample, based on whether English was their first language, affects the results of DIF analyses. In order to achieve that goal, we employed data from a large sample of examinees obtained from a PSAT/NMSQT® (Preliminary SAT/National Merit Scholarship Qualifying Test) administration. During the PSAT/NMSQT examination, examinees were asked about the language they first learned to speak. They could answer (a) English, (b) English and another, or (c) another. However, they did not have to answer the question. The PSAT/NMSQT operational DIF analysis was performed on sophomores and juniors who chose either the first option or second option. This subpopulation is henceforth referred to as EFL. The sophomores and juniors who chose the third option to the question are henceforth referred to as NEFL.

Using PSAT/NMSQT data, we performed female/male, Black/White, Asian American/White, and Hispanic[1]/White DIF analysis on three groups: (a) NEFL, (b) EFL, and (c) total (the combination of EFL and NEFL). Those who did respond to the question were excluded from our analysis. We then compared the three sets of results to find out if the first language status of the examinees had an effect on the results of the DIF analyses for PSAT/NMSQT.

It was also important to consider how the DIF results would be affected if the proportion of NEFL examinees increased in the examinee population, especially in light of the above-mentioned findings of Kirsch et al. (2007). Hence, we studied the sensitivity of the DIF results to

the proportion of NEFL examinees by creating several *synthetic subpopulations* from the available data set and running DIF analyses on these synthetic subpopulations. A synthetic subpopulation was created by combining all the available NEFL examinees with an appropriately sized simple random sample of the EFL examinees.

The following sections provide some background material. Section 2 describes the PSAT/NMSQT data. Section 3 describes our methods. Section 4 describes the results for the PSAT/NMSQT data. Section 5 provides a discussion and conclusions. In a related paper, Liang, Dorans, and Sinharay (2009) examined the effects of exclusion of NEFL examinees on score equating.

2. Description of the Data

The PSAT/NMSQT has three sections: critical reading, mathematics, and writing. We analyzed data from all three sections of two recent PSAT/NMSQT test administrations: one Wednesday administration and one Saturday administration. The Saturday group scored higher than the Wednesday group on all three measures. The main reasons for employing PSAT/ NMSQT data were the large size of the examinee population and the presence of a significant proportion of NEFL examinees taking the test. A potential problem with PSAT/NMSQT data was that they included little DIF because PSAT/NMSQT items were taken from old operational SAT items, which rarely exhibit DIF.

Table 1 contains the percentage of NEFL examinees in each of the six groups for both the Saturday and Wednesday forms. The table also includes the total number for all examinee groups. Note that the Hispanic and Asian American groups contained the largest proportion of NEFL examinees; about one in three members of these groups were NEFL and hence were excluded from the PSAT/NMSQT operational DIF analysis.

The sample sizes, means, and standard deviations for total, EFL, and NEFL examinees within each of the seven groups (all, male, female, White, Black, Hispanic, and Asian American examinees) as well as the ratios and standardized mean differences of sample sizes are presented in Tables 2 to 7. Each table shows one of the six combinations of test and administration. Note that the sample sizes for critical reading, mathematics, and writing sections were the same for the Saturday form; the Wednesday form sample sizes were identical as well. The numbers listed as F/R ratio are the sample size ratios of focal group to reference group. The numbers listed as std mean diff (F/R) are standardized mean differences. Each set of numbers contains row entries for

focal/reference group pairs (female/male, Black/White, Hispanic/White, Asian American/White) under the column entries for total, EFL, and NEFL examinees.

Table 1

Percentage of Not English First Language (NEFL) Examinees in the Different Examinee Groups

Examinee group	Saturday form		Wednesday form	
	Sample size	Percentage of NEFL	Sample size	Percentage of NEFL
Total	514,895	7.80	2,323,309	9.46
Male	230,218	7.66	1,081,726	9.46
Female	284,103	7.91	1,236,900	9.45
White	357,020	1.63	1,260,698	1.45
Black	26,406	2.79	379,085	1.88
Asian American	64,613	29.97	145,798	31.89
Hispanic	38,814	27.71	399,203	33.39

The ratios of samples sizes for a given column were obtained by dividing the number of focal group members by the number of reference group members for the column group. For example, in Table 2, the ratio for female/male in the EFL column is 1.23, indicating that there were 1.23 females for every male in the EFL group.

The standardized mean raw score differences were obtained by subtracting the reference group mean from the focal group mean and then dividing the difference by the total group standard deviation. For example, in Table 2, the difference of -0.1 in the F/M row for EFL indicates that the mean difference between the female EFL group and the male EFL group divided by the standard deviation of the total group was -0.1.

Tables 2 to 7 show the following results:

- The sample sizes for the Wednesday form are much larger than those for the Saturday form. Comparisons should not be made across Wednesday and Saturday as the numbers in these tables are based on raw scores, not scaled scores. However, equated scaled scores generally tend to show that the Saturday group was more able than the Wednesday group.

- Rows 1 to 7 (all, male, female, White, Black, Hispanic, and Asian American) of Tables 2 to 7 show that the mean of the NEFL group was lower or about the same as the mean of the EFL group overall. This finding held true for all gender and ethnic/racial groups with a few exceptions. For example, the NEFL mean was 2.38 score points higher than the EFL mean for the Asian American examinees for Wednesday mathematics (see Table 5), and the NEFL mean was higher than the EFL mean for several groups for Saturday mathematics (see Table 4).
- Row 8 (F/R ratio) of Tables 2 to 7 shows that the ratios of female examinees to male examinees and of Black examinees to White examinees were close in the EFL and NEFL groups. In stark contrast, the ratios of Asian Americans to Whites and of Hispanics to Whites were much higher in the NEFL group than in the EFL group. For example, in Table 2, the ratios of Asian Americans to Whites were 0.13 in the EFL group and 3.33 in the NEFL group.
- Row 9 (std mean difference [F-R]) of Tables 2 to 7 shows that the mean standardized differences for the females and males were almost the same in both the EFL and NEFL groups. The same was not true for the Black/White, Asian American/White, and Hispanic/White differences, which indicated an interaction between mean score difference and the first language indicator. The interaction was strongest for the Hispanic/White differences, which were considerably wider in the NEFL group than in the EFL group (for example -0.80 versus -0.53 in Table 2). The Black/White differences were close for the EFL and NEFL groups for a few tests (such as Saturday critical reading and Wednesday mathematics; see Tables 2 and 5), but not so for a few others (such as Wednesday critical reading and Saturday writing; see Tables 3 and 6). The Asian American/White differences were close for the EFL and NEFL groups except for Wednesday mathematics and Saturday writing (see Tables 5 and 6). As such, the first language indicator may be a better proxy for actual limited proficiency in English in the Hispanic group than in the only other group, Asian Americans, which had a large percentage of examinees who indicated that English was not one of their first languages.

Table 2

Statistics for Saturday Critical Reading

		Total	EFL	NEFL
All	*N*	514,895	474,735	40,160
	Mean	21.818	22.053	19.044
	SD	10.017	9.905	11.252
Male	*N*	230,218	212,593	17,625
	Mean	22.355	22.585	19.583
	SD	10.178	10.037	11.385
Female	*N*	284,103	261,623	22,480
	Mean	21.39	21.626	18.637
	SD	9.92	9.774	11.122
White	*N*	357,020	351,200	5,820
	Mean	22.842	22.854	22.135
	SD	9.561	9.548	10.25
Black	*N*	26,406	25,668	738
	Mean	15.349	15.373	14.5
	SD	9.273	9.252	9.927
Hispanic	*N*	38,814	28,058	10,756
	Mean	16.585	17.522	14.142
	SD	9.716	9.776	9.116
Asian American	*N*	64,613	45,248	19,365
	Mean	22.466	22.924	21.394
	SD	10.845	10.428	11.693
F/R ratio	F/M	1.23	1.23	1.28
	B/W	0.07	0.07	0.13
	H/W	0.11	0.08	1.85
	A/W	0.18	0.13	3.33
Std mean diff (F-R)	F/M	-0.10	-0.10	-0.09
	B/W	-0.75	-0.75	-0.76
	H/W	-0.62	-0.53	-0.80
	A/W	-0.04	0.01	-0.07

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H), and Asian American (A). Reference groups are male (M) and White (W). EFL = English first language, NEFL = not English first language.

Table 3

Statistics for Wednesday Critical Reading

		Total	EFL	NEFL
All	*N*	2,323,309	2,103,450	219,859
	Mean	16.834	17.291	12.464
	SD	10.684	10.764	9.888
Male	*N*	1,081,726	979,365	102,361
	Mean	16.533	16.981	12.247
	SD	11.023	11.024	10.052
Female	*N*	1,236,900	1,120,033	116,867
	Mean	17.117	17.581	12.678
	SD	10.548	10.522	9.741
White	*N*	1,260,698	1,242,383	18,315
	Mean	20.264	20.294	18.223
	SD	10.388	10.381	10.674
Black	*N*	379,085	371,967	7,118
	Mean	10.796	10.808	10.208
	SD	8.504	8.493	9.059
Hispanic	*N*	399,203	265,922	133,281
	Mean	11.761	12.358	10.569
	SD	9.018	9.247	8.417
Asian American	*N*	145,798	99,296	46,502
	Mean	18.157	19.1	16.143
	SD	11.134	10.857	11.446
F/R ratio	F/M	1.14	1.14	1.14
	B/W	0.30	0.30	0.39
	H/W	0.32	0.21	7.28
	A/W	0.12	0.08	2.54
Std mean diff (F-R)	F/M	0.05	0.06	0.04
	B/W	-0.89	-0.89	-0.75
	H/W	-0.80	-0.74	-0.72
	A/W	-0.20	-0.11	-0.19

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H) and Asian American (A) examinees. Reference groups are male (M) and White (W) examinees. EFL = English first language; NEFL = not English first language.

Table 4

Statistics for Saturday Mathematics

		Total	EFL	NEFL
All	*N*	514,895	474,735	40,160
	Mean	19.737	19.639	20.899
	SD	8.694	8.571	10.033
Male	*N*	230,218	212,593	17,625
	Mean	21.3	21.184	22.706
	SD	8.779	8.669	9.903
Female	*N*	284,103	261,623	22,480
	Mean	18.476	18.389	19.494
	SD	8.426	8.281	9.902
White	*N*	357,020	351,200	5,820
	Mean	20.136	20.105	22.034
	SD	8.179	8.163	8.871
Black	*N*	26,406	25,668	738
	Mean	12.811	12.806	12.991
	SD	7.737	7.725	8.138
Hispanic	*N*	38,814	28,058	10,756
	Mean	14.904	15.37	13.69
	SD	8.035	8.171	7.537
Asian American	*N*	64,613	45,248	19,365
	Mean	23.84	23.142	25.469
	SD	9.11	9.068	8.998
F/R ratio	F/M	1.23	1.23	1.28
	B/W	0.07	0.07	0.13
	H/W	0.11	0.08	1.85
	A/W	0.18	0.13	3.33
Std mean diff (F-R)	F/M	-0.32	-0.32	-0.37
	B/W	-0.84	-0.84	-1.04
	H/W	-0.60	-0.54	-0.96
	A/W	0.43	0.35	0.40

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H), and Asian American (A). Reference groups are male (M) and White (W). EFL = English first language, NEFL = not English first language.

Table 5

Statistics for Wednesday Mathematics

		Total	EFL	NEFL
All	*N*	2,323,309	2,103,450	219,859
	Mean	14.188	14.365	12.499
	SD	9.483	9.433	9.953
Male	*N*	1,081,726	979,365	102,361
	Mean	14.984	15.154	13.361
	SD	9.984	9.922	10.423
Female	*N*	1,236,900	1,120,033	116,867
	Mean	13.509	13.691	11.767
	SD	8.994	8.925	9.461
White	*N*	1,260,698	1,242,383	18,315
	Mean	17.023	17.023	17.047
	SD	8.869	8.864	9.25
Black	*N*	379,085	371,967	7,118
	Mean	8.007	8.003	8.211
	SD	7.398	7.389	7.868
Hispanic	*N*	399,203	265,922	133,281
	Mean	9.849	10.15	9.248
	SD	8.057	8.177	7.779
Asian American	*N*	145,798	99,296	46,502
	Mean	19.341	18.581	20.961
	SD	10.238	10.032	10.483
F/R ratio	F/M	1.14	1.14	1.14
	B/W	0.30	0.30	0.39
	H/W	0.32	0.21	7.28
	A/W	0.12	0.08	2.54
Std mean diff (F-R)	F/M	-0.16	-0.15	-0.17
	B/W	-0.95	-0.95	-0.93
	H/W	-0.76	-0.72	-0.82
	A/W	0.24	0.16	0.41

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H) and Asian American (A). Reference groups are male (M) and White (W). EFL = English first language, NEFL = not English first language.

Table 6

Statistics for Saturday Writing

		Total	EFL	NEFL
All	*N*	514,895	474,735	40,160
	Mean	17.937	17.978	17.456
	SD	8.879	8.786	9.915
Male	*N*	230,218	212,593	17,625
	Mean	17.31	17.331	17.056
	SD	8.921	8.832	9.928
Female	*N*	284,103	261,623	22,480
	Mean	18.452	18.51	17.783
	SD	8.813	8.712	9.886
White	*N*	357,020	351,200	5,820
	Mean	18.477	18.452	19.982
	SD	8.535	8.524	9.049
Black	*N*	26,406	25,668	738
	Mean	12.506	12.515	12.199
	SD	8.085	8.074	8.432
Hispanic	*N*	38,814	28,058	10,756
	Mean	13.74	14.279	12.335
	SD	8.282	8.388	7.826
Asian American	*N*	64,613	45,248	19,365
	Mean	20.098	20.093	20.111
	SD	9.531	9.3	10.051
F/R ratio	F/M	1.23	1.23	1.28
	B/W	0.07	0.07	0.13
	H/W	0.11	0.08	1.85
	A/W	0.18	0.13	3.33
Std mean diff (F-R)	F/M	0.13	0.13	0.08
	B/W	-0.67	-0.67	-0.88
	H/W	-0.53	-0.47	-0.86
	A/W	0.18	0.18	0.01

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H), and Asian American (A). Reference groups are male (M) and White (W). EFL = English first language, NEFL = not English first language.

Table 7

Statistics for Wednesday Writing

		Total	EFL	NEFL
All	*N*	2,323,309	2,103,450	219,859
	Mean	13.831	14.19	10.399
	SD	9.441	9.476	9.101
Male	*N*	1,081,726	979,365	102,361
	Mean	13.103	13.435	9.93
	SD	9.515	9.5	9.067
Female	*N*	1,236,900	1,120,033	116,867
	Mean	14.487	14.868	10.832
	SD	9.45	9.403	9.113
White	*N*	1,260,698	1,242,383	18,315
	Mean	16.868	16.88	16.019
	SD	9.022	9.015	9.464
Black	*N*	379,085	371,967	7,118
	Mean	8.274	8.28	7.987
	SD	7.608	7.601	7.963
Hispanic	*N*	399,203	265,922	133,281
	Mean	9.497	10.002	8.489
	SD	8.263	8.382	7.924
Asian American	*N*	145,798	99,296	46,502
	Mean	15.425	16.038	14.117
	SD	9.815	9.627	10.08
F/R ratio	F/M	1.14	1.14	1.14
	B/W	0.30	0.30	0.39
	H/W	0.32	0.21	7.28
	A/W	0.12	0.08	2.54
Std mean diff (F-R)	F/M	0.15	0.15	0.10
	B/W	-0.91	-0.91	-0.85
	H/W	-0.78	-0.73	-0.80
	A/W	-0.15	-0.09	-0.20

Note. F/R ratio is the ratio of the sample size for the focal group to the reference group. Focal groups are female (F), Black (B), Hispanic (H), and Asian American (A). Reference groups are male (M) and White (W). EFL = English first language, NEFL = not English first language.

3. Methods

3.1 Differential Item Functioning (DIF) Analyses on the Observed Population

The operational DIF analyses for the PSAT/NMSQT were performed on the EFL subpopulation with either White examinees or male examinees as a reference group for any focal group with sufficient sample size. We ran the same DIF analyses (female/male, Black/White, Asian American/White, and Hispanic/White) for the PSAT/NMSQT on the total population and the NEFL subpopulation as well. Then we compared the three sets of DIF statistics (from total, EFL, and NEFL groups).

The Mantel-Haenszel (MH D-DIF) statistic and the standardized P-difference (STD P-DIF) statistic (Dorans & Kulick, 1986; Holland & Wainer, 1993) were used as DIF statistics.

ETS has a system of categorizing the extent of DIF based on both the magnitude of the MH D-DIF statistic and the statistical significance of the results. An item has a DIF classification of C if the absolute value of MH D-DIF is at least 1.5 and is significantly greater than 1 at the 5% significance level. An item has a DIF classification of A if either the absolute value of MH D-DIF is less than 1 or the MH D-DIF value is not significantly different from 0. Items that cannot be classified as A or C belong to category B. Items in the C category are subjected to further inspection; they are typically eliminated from the item pool and sometimes dropped from previously administered tests if an explanation for the DIF as a source of construct irrelevant variance in the test scores can be found.

To compare two sets of DIF statistics (for example, a set for the EFL examinees and another set for the NEFL examinees), we used graphical plots and simple correlations. We also used mean difference, which is the difference between the arithmetic mean of the two sets of statistics and root mean squared differences. Suppose the first set of DIF statistics (which could be the MH D-DIF statistics for the EFLs) are $X_i, i = 1, 2, ...I,$ and the second set of DIF statistics (which could be the MH D-DIF statistics for the NEFLs) are $Y_i, i = 1, 2, ...I.$ The mean difference is defined as $MD = \frac{1}{I}\sum_i X_i - \frac{1}{I}\sum_i Y_i$, and the root mean squared difference is defined as $RMSD = \sqrt{\frac{1}{I}\sum_i (X_i - Y_i)^2}$.

3.2 Differential Item Functioning (DIF) Analyses on the Synthetic Populations

The above analyses, after being performed on all the available data sets, provided results for the currently observed values of the percentage of NEFL examinees (which were all slightly less than 10 percentage points) but did not reveal how DIF results would change if the percentage of NEFL examinees increased to a higher value, say 25%. Hence, to study the DIF results for percentages of NEFL examinees higher than those observed for the PSAT/NMSQT data (which is important given the above-mentioned findings of Kirsch et al., 2007), we created *synthetic populations* from the data by combining the NEFL group with a random sample from the EFL group. For example, consider the Saturday form of the PSAT/NMSQT that was taken by 474,735 EFL examinees and 40,160 NEFL examinees. The percentage of NEFL examinees among all examinees was 7.8%.[2] However, if we were to draw a random sample of 40,160 examinees from the 474,735 EFL examinees and combine them with the 40,160 NEFL examinees, it would yield a population with 50%[3] NEFL examinees. We created synthetic populations with proportions of NEFL examinees from 0.1 to 0.9 in steps of 0.1. For each synthetic population, we performed female/male, Black/White, Asian American/White, and Hispanic/White DIF analysis separately on the total, EFL, and NEFL groups and then compared the three sets of DIF statistics (from total, EFL, and NEFL groups) using graphical plots, correlation, mean difference, and root mean squared difference. Such analyses allowed us to study the sensitivity of the DIF results as the percentage of NEFL examinees varied. A goal here was to find for each test a specific value *p* so that the DIF results would be significantly affected if the proportion of NEFL examinees increased above *p*.

4. Results: Impact of First Language Status on PSAT/NMSQT Data

4.1 Differential Item Functioning (DIF) Results for the Observed Population

Table 8 shows the cross-classification of the DIF categorizations of the PSAT/NMSQT items for each of the six form/score combinations when the DIF analysis was run on the EFL population and on the NEFL population.

For example, for Wednesday critical reading, the number 43 denotes that 43 of 48 items had an A-DIF category when the DIF analysis was run on both the EFL population and the NEFL population. An item's DIF category in the table is A if it is an A-DIF item in all four DIF analyses (female/male, Black/White, and Asian American/White, and Hispanic/White). An

item's DIF category is C if it is a C-DIF item in any one of the four DIF analyses. An item's DIF category is B if its DIF category is neither A nor C.

Table 8

Cross-Classification of Differential Item Functioning (DIF) Categorizations for the PSAT/NMSQT Data

EFL	NEFL		
	A	B	C
Wednesday critical reading			
A	43	1	1
B	3	0	0
C	0	0	0
Saturday critical reading			
A	36	6	0
B	2	4	0
C	0	0	0
Wednesday mathematics			
A	35	1	0
B	1	1	0
C	0	0	0
Saturday mathematics			
A	32	1	0
B	2	3	0
C	0	0	0
Wednesday writing			
A	33	2	1
B	1	1	1
C	0	0	0
Saturday writing			
A	27	7	3
B	0	2	0
C	0	0	0

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

In Table 8, 217 of 250 items lie along the diagonals of the table, that is, they have the same DIF category when the analysis was run on the EFL and the NEFL. This positioning gives the impression that the DIF categorizations for EFL were in good agreement with those in NEFL. However, this agreement was mostly because the PSAT/NMSQT items had very little DIF, and hence, most items were classified as A-DIF items for EFL and A-DIF items for NEFL. Such a

lack of DIF was an outcome of pretesting and subsequent removal of items with extreme DIF for the SAT (the items in PSAT/NMSQT items were taken from old SAT items).

Categorization of DIF into A, B, and C categories often hides part of the information that is present in the DIF statistics themselves. Hence we proceeded to examine the actual DIF statistics.

Figures 1 to 6 plot the MH D-DIF statistics for the EFL population versus those for the NEFL population for the six form/score combinations. The four panels in each figure show results for female/male DIF, Black/White DIF, Hispanic/White DIF, and Asian American/White DIF. The values of the correlation, root mean squared difference, and mean difference are shown in the figures as well.

Table 9 shows the values of the correlation, root mean squared difference, and mean difference for the MH D-DIF statistic and for the STD P-DIF statistic for the six form/score combinations from EFL and NEFL. Note that the numerical root mean squared difference (RMSD) and mean difference (MD) values for MH D-DIF are not comparable to those for STD P-DIF because these two statistics are expressed in different metrics.

Figures 1 to 6 and Table 9 demonstrate that the correlations for the female/male DIF analysis were always quite high across form/score combinations, which means that the results of the female/male DIF analysis were the same regardless of whether the DIF analysis was performed only with EFL examinees or only with NEFL examinees. This invariance result did not occur for the other DIF analyses. With a few exceptions, the correlation for the Hispanic/White DIF was the lowest among the four types of DIF analyses. It was interesting to note that all items in the Hispanic/White DIF analysis were often of the A-DIF category (as the corresponding MH D-DIF statistics for all items were less than 1 in absolute value for all the items); an investigator examining only the corresponding DIF categorizations will observe only good agreement between the EFL DIF categorizations and NEFL DIF categorizations and will miss the low correlations observed for the Hispanic/White DIF analysis.

For Black/White DIF, Hispanic/White DIF, and Asian American/White DIF, the correlations were lower and the root mean squared differences were higher for critical reading and writing compared to mathematics.

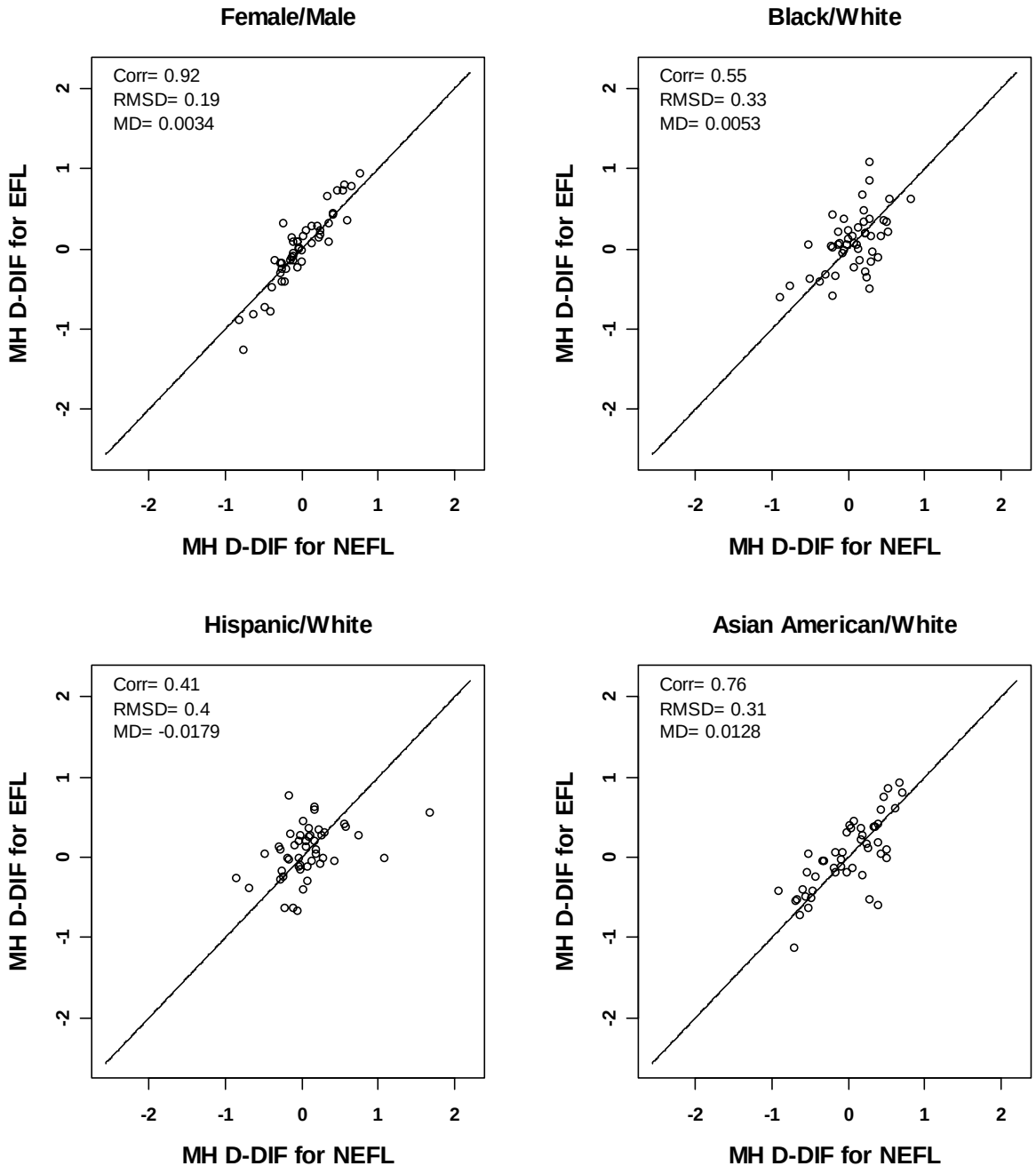

Figure 1. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday critical reading.**

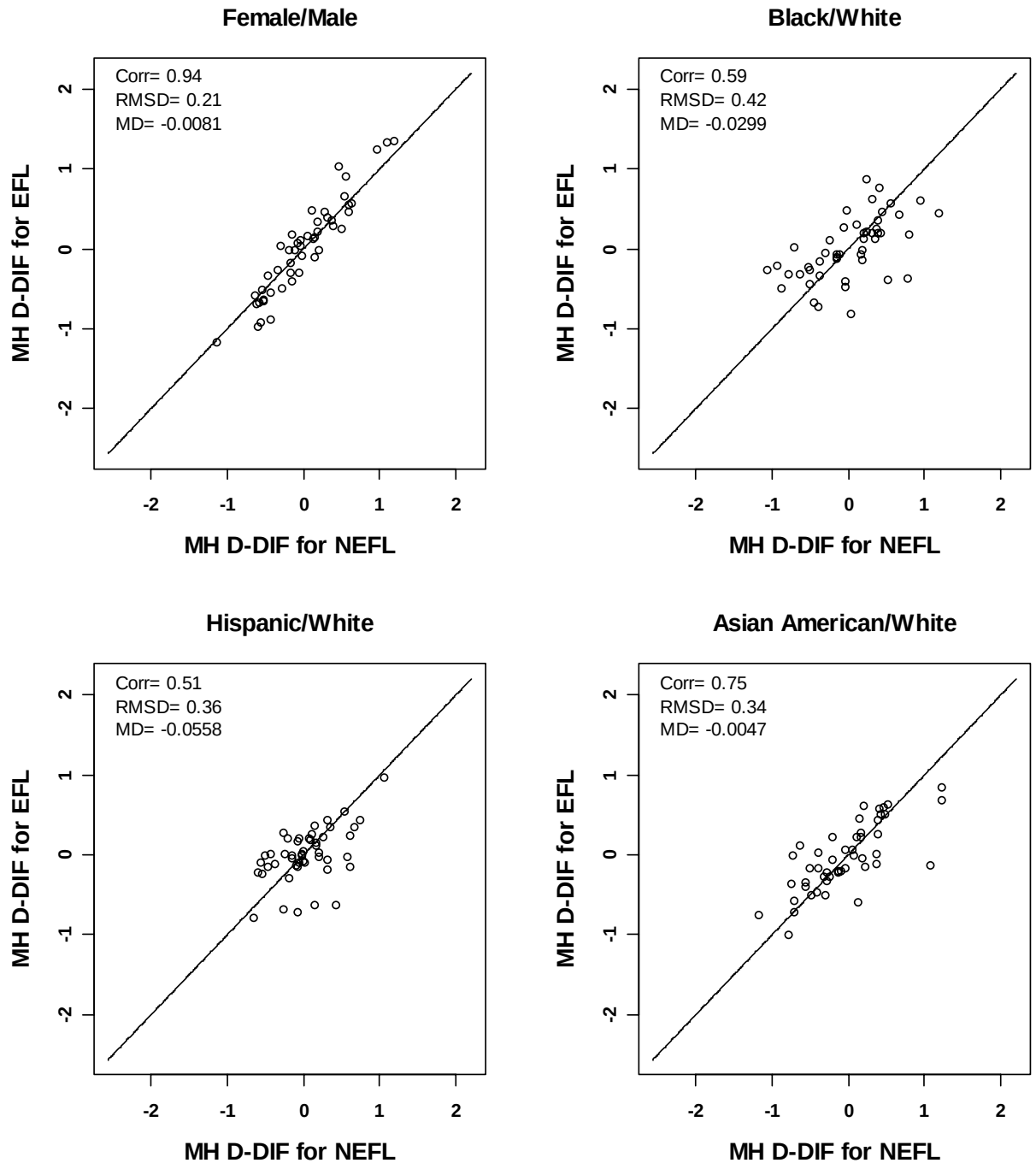

Figure 2. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday critical reading.**

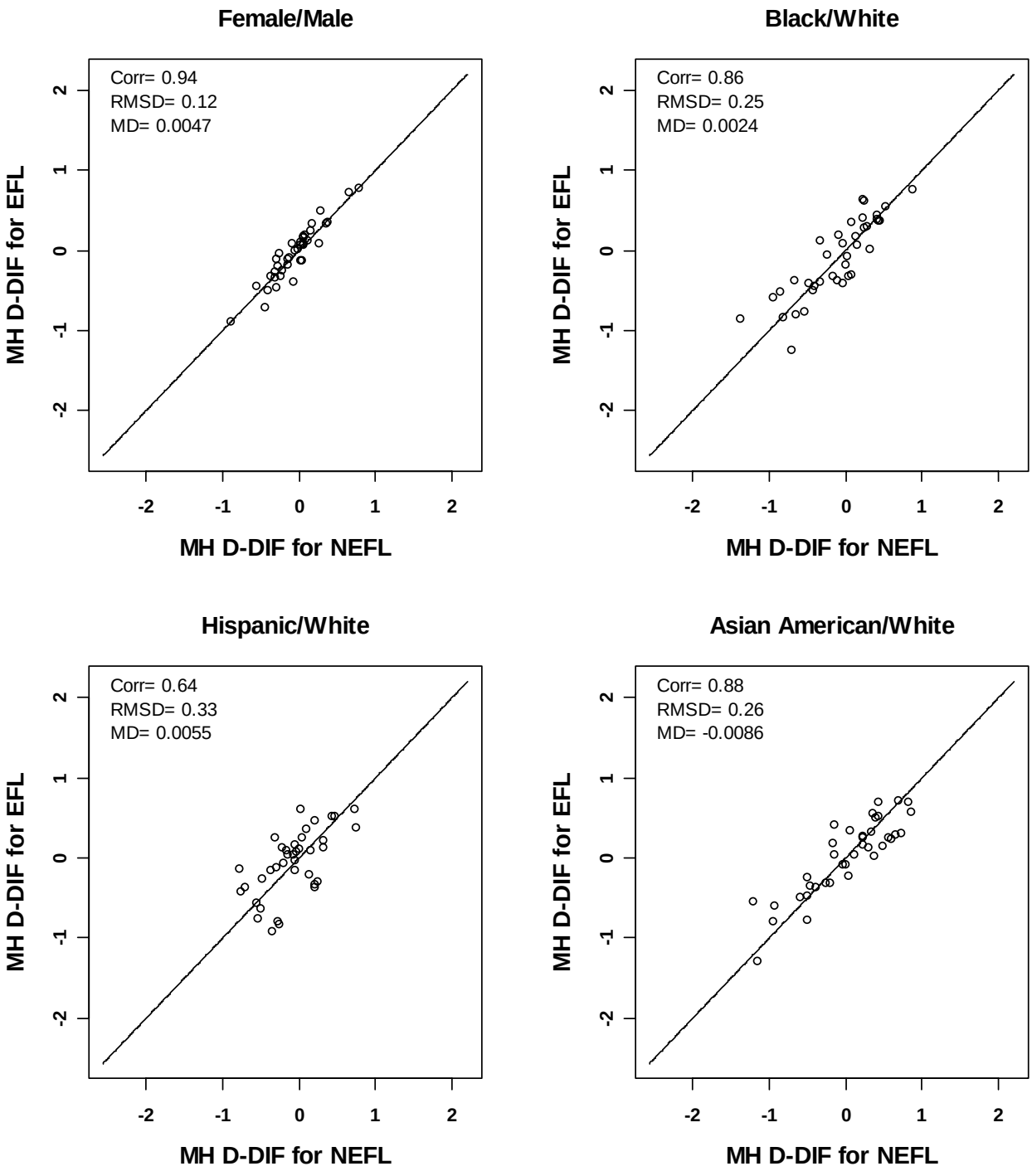

Figure 3. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday mathematics.**

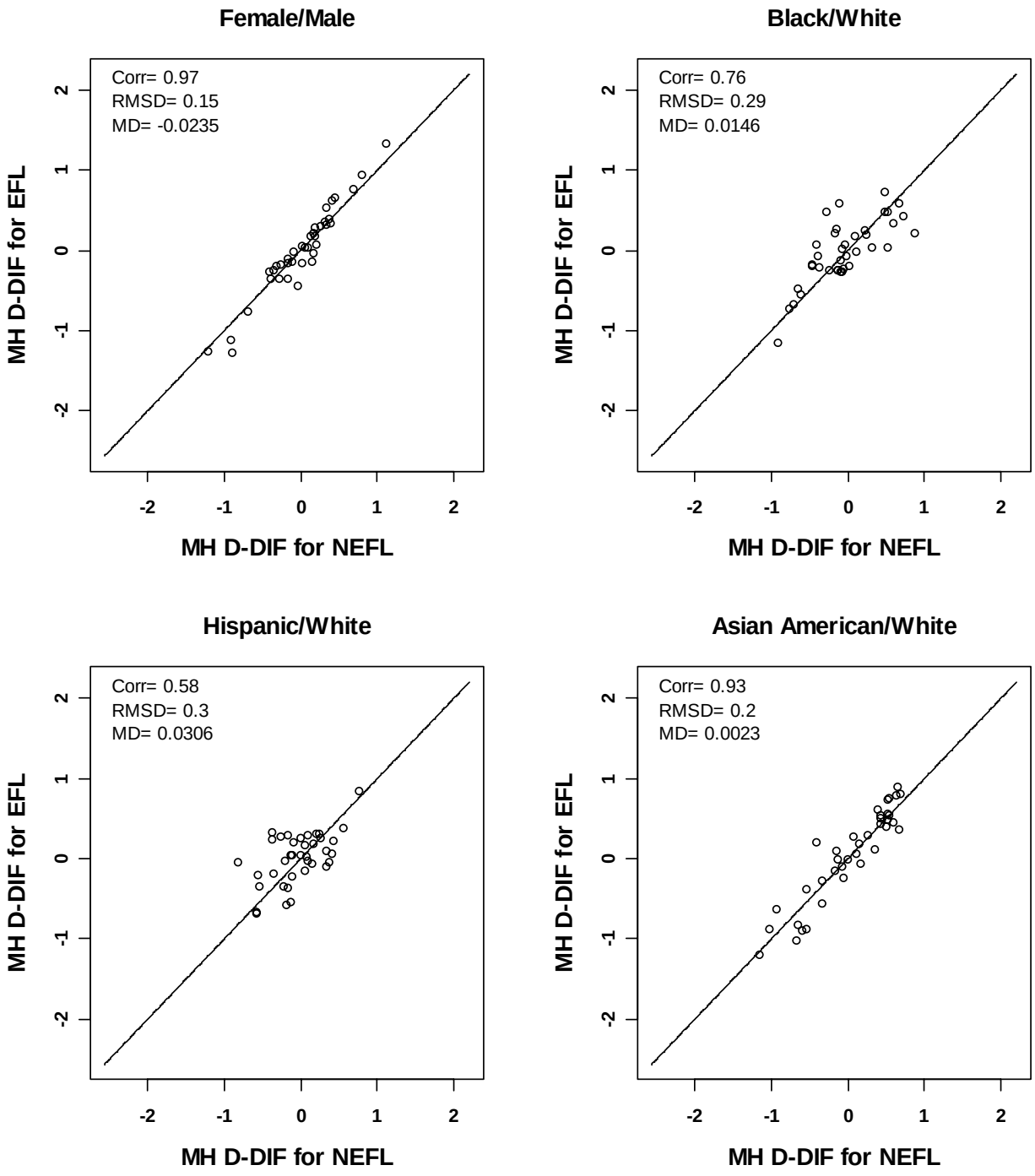

Figure 4. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday mathematics.**

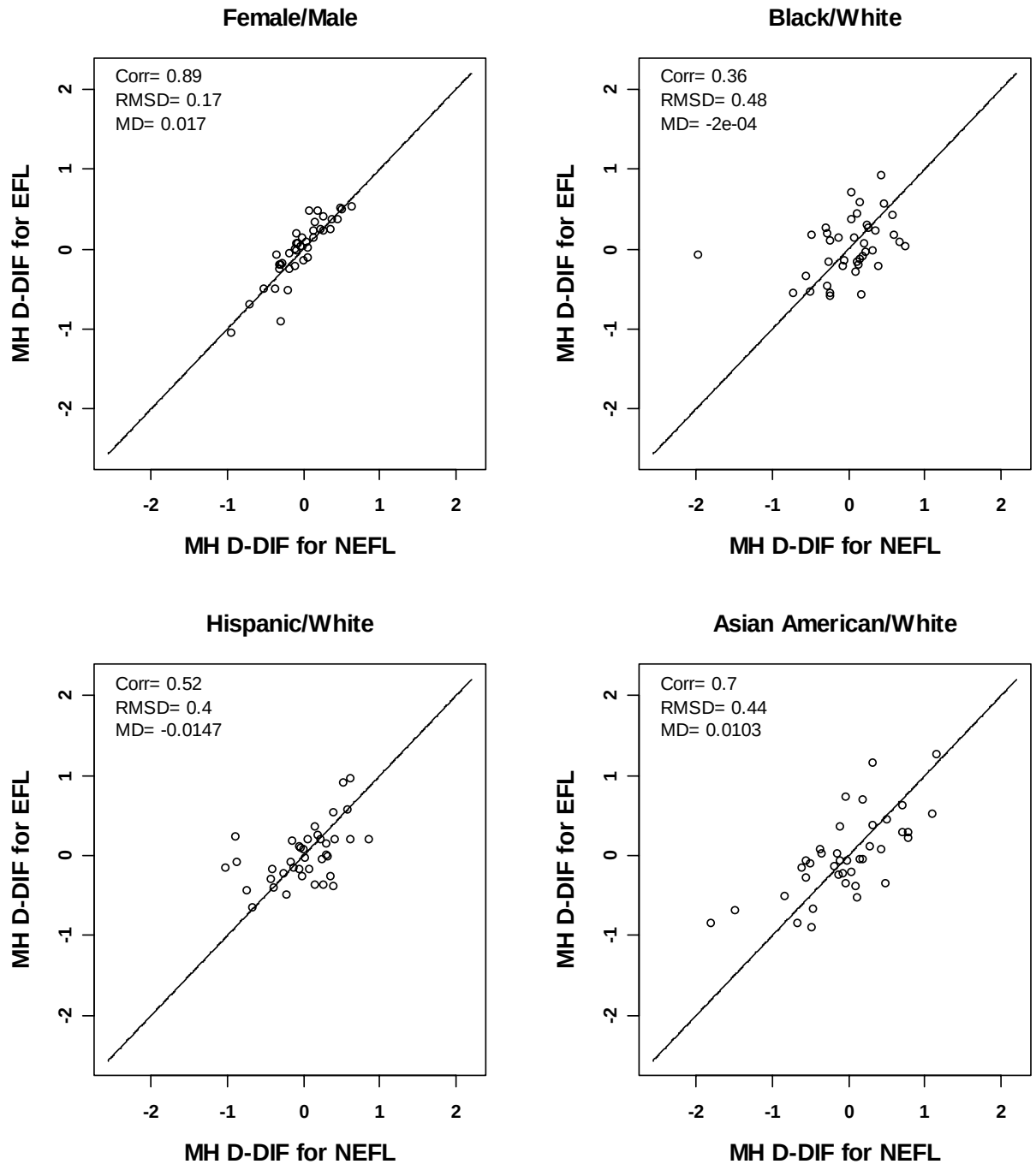

Figure 5. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday writing.**

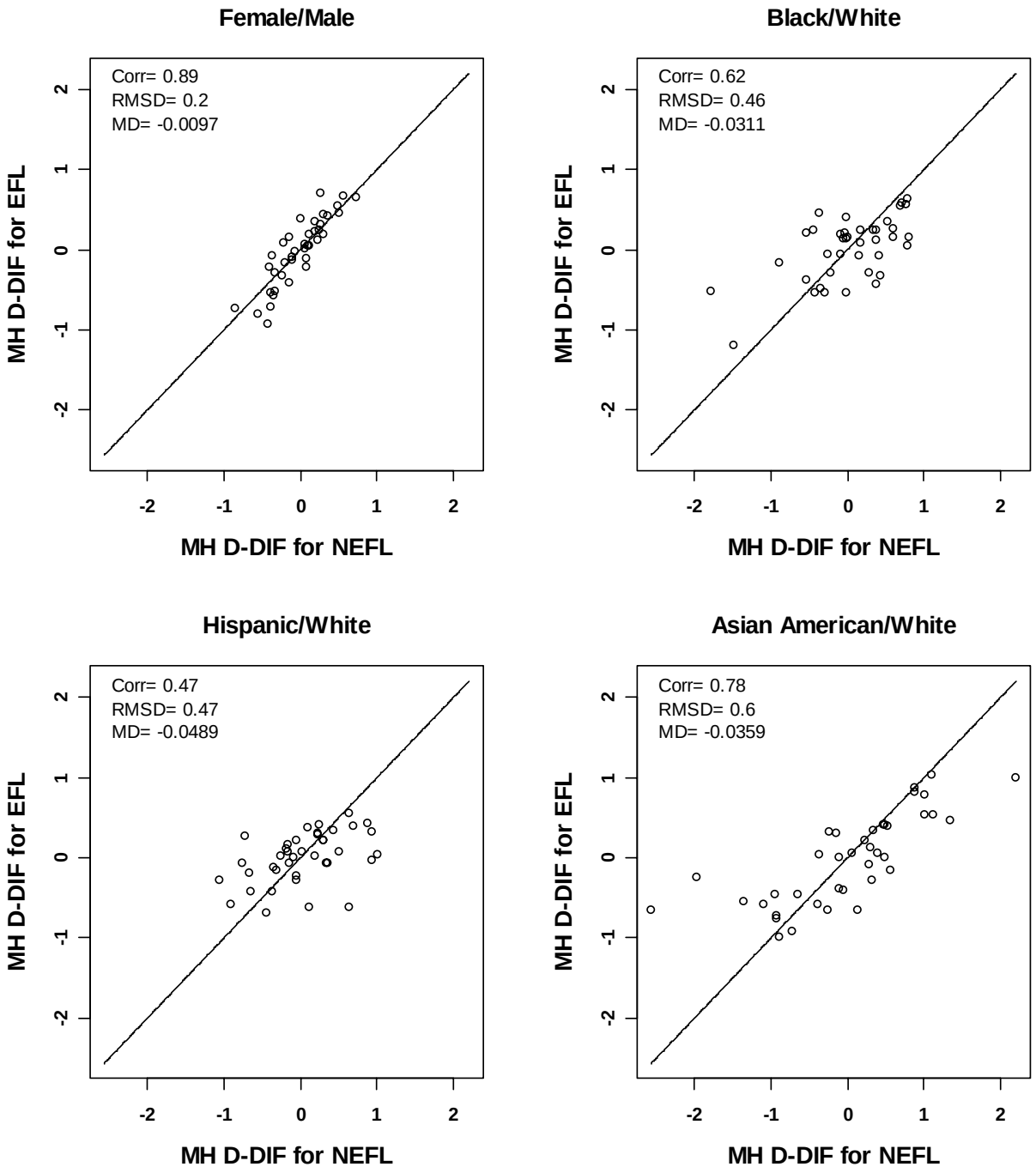

Figure 6. **Plot of the Mantel-Haenszel differential difficulty (MH D-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday writing**

Table 9

Statistics for English First Language (EFL) and English Not First Language (NEFL) for the PSAT/NMSQT Test Forms

Test form	DIF analysis	MH D-DIF			STD P-DIF		
		Corr.	RMSD	MD	Corr.	RMSD	MD
Wed. CR	Female/Male	0.92	0.19	.003	0.93	0.01	-.001
	Black/White	0.55	0.33	.005	0.53	0.02	.000
	Hispanic/White	0.41	0.40	-.018	0.39	0.03	-.002
	Asian American/ White	0.76	0.31	.013	0.72	0.02	.001
Sat. CR	Female/Male	0.94	0.21	-.008	0.94	0.01	-.001
	Black/White	0.59	0.42	-.030	0.61	0.03	.000
	Hispanic/White	0.51	0.36	-.056	0.62	0.02	-.002
	Asian American/ White	0.75	0.34	-.005	0.77	0.02	.000
Wed. M	Female/Male	0.94	0.12	.005	0.93	0.01	.000
	Black/White	0.86	0.25	.002	0.90	0.01	.001
	Hispanic/White	0.64	0.33	.006	0.78	0.02	-.001
	Asian American/White	0.88	0.26	-.009	0.91	0.01	-.001
Sat. M	Female/Male	0.97	0.15	-.024	0.97	0.01	-.001
	Black/White	0.76	0.29	.015	0.76	0.02	.002
	Hispanic/White	0.58	0.30	.031	0.67	0.02	.000
	Asian American/ White	0.93	0.20	.002	0.94	0.01	.000
Wed. W	Female/Male	0.89	0.17	.017	0.86	0.01	.000
	Black/White	0.36	0.48	.000	0.34	0.04	.000
	Hispanic/White	0.52	0.40	-.015	0.56	0.03	-.001
	Asian American/ White	0.70	0.44	.010	0.70	0.03	.001
Sat. W	Female/Male	0.89	0.20	-.010	0.90	0.01	.000
	Black/White	0.62	0.46	-.031	0.64	0.04	.001
	Hispanic/White	0.47	0.47	-.049	0.57	0.03	-.001
	Asian American/ White	0.78	0.60	-.036	0.78	0.04	0.000

Note. Corr. = correlation, CR = critical reading, M = mathematics, W = writing, DIF = differential item functioning, MD = mean difference, MH D-DIF = Mantel-Haenszel differential difficulty, STD P-DIF = standardized P-difference, RMSD = root mean squared difference.

Though the mean differences were mostly close to 0 (indicating that positive and negative differences cancel out), they were larger, for example, for Hispanic/White DIF for Saturday critical reading and Saturday writing (see Tables 2 and 6). This observation was clearer from the mean differences for the MH D-DIF statistics.

A few outlier items can be observed in the figures. For example, for Hispanic/White DIF analysis in Wednesday critical reading (see Figure 1), an item lies to the far right of the plot; it has an MH D-DIF value of 1.68 for the NEFL population (i.e., it is a C-DIF item) and an MH D-DIF value of only 0.55 for the EFL population (i.e., it is an A-DIF item). For Black/White DIF analysis in Wednesday writing (see Figure 5), an item lies to the far left of the plot; it has an MH D-DIF value of about -2 for the NEFL population (i.e., it is a C-DIF item) and just below 0 for the EFL population (i.e., it is an A-DIF item). These outliers remind us that individual items can be affected substantially by whether the DIF analysis is performed on the EFL group or not, even when most of the items are not affected. DIF involves an item-by-item level of analysis; these outliers can not be ignored.

4.2 Differential Item Functioning (DIF) Results for Synthetic Populations

Figures 7 to 12 show how a change in the proportion of NEFL examinees affected the PSAT/NMSQT DIF results. Figures 7 to 10 show results for Saturday critical reading. Figures 11 and 12 show the results for the Asian American/White DIF analyses for Saturday writing and Saturday mathematics respectively. These selected six plots show the typical patterns that we observed in all the plots for all form/score combinations. (More such plots are shown in the appendix.) Each figure has six panels—the three left panels show the results for the MH D-DIF statistic while the three right panels show the results for the STD P-DIF statistic. The two topmost panels show the results for correlation, the two middle panels show the results for root mean squared difference, and the two bottommost panels show the results for mean difference. Any panel shows, for specific proportions of NEFL examinees (0.1, 0.2, …0.9), the values of the corresponding quantity (correlation, root mean squared difference, or mean difference) measuring association between the following statistics:

1. DIF statistics for the total population versus those for the EFL population (using an oval).
2. DIF statistics for the total population versus those for the NEFL population (using a triangle).

3. DIF statistics for the EFL population versus those for the NEFL population (using a plus sign).

For example, Figure 7 shows results for female/male DIF for Saturday critical reading. The top left panel of the figure shows results for correlation for the MH D-DIF statistics. The panel plots three sets of points. The ovals in the panel denote the correlations between the MH D-DIF statistic for the total population and the MH D-DIF statistic for the EFL population for NEFL proportions (0.1, 0.2, …, 0.9).

To make the results for the MH D-DIF statistic comparable to those for the STD P-DIF statistic, we made the range of the vertical scale for any plot for STD P-DIF about 0.08 times[4] that of the corresponding plot for MH D-DIF. For example, the range of the vertical scale for root mean squared difference for STD P-DIF was 0 to 0.055, about 0.08 times that of the MH D-DIF (0 to 0.693).

As viewed from left to right in any panel, the measures, based on fewer numbers of examinees, are more variable.

The figures lead to the following conclusions:

- The association between DIF statistics for Total-EFL mostly becomes weaker (correlation decreases and root mean squared difference increases) as the NEFL proportion increases. This association appears because an increase in the NEFL proportion causes the total population to differ more from the EFL population. In contrast, the association for Total-NEFL mostly becomes stronger (correlation increases and root mean squared difference decreases) as the NEFL proportion increases.

- Among Total-EFL, Total-NEFL, and EFL-NEFL, for lower proportions of NEFL examinees, the association among DIF statistics is strongest for Total-EFL and weakest for EFL-NEFL. This result is also expected as the extent of overlap of examinees is largest for Total-EFL, followed by Total-NEFL—there is no overlap for EFL-NEFL. For a high proportion of NEFL examinees, the association is strongest for Total-NEFL and weakest for EFL-NEFL. It often takes a high proportion of NEFLs (0.7 or higher) for the correlation for Total-NEFL to be larger than that for Total-EFL.

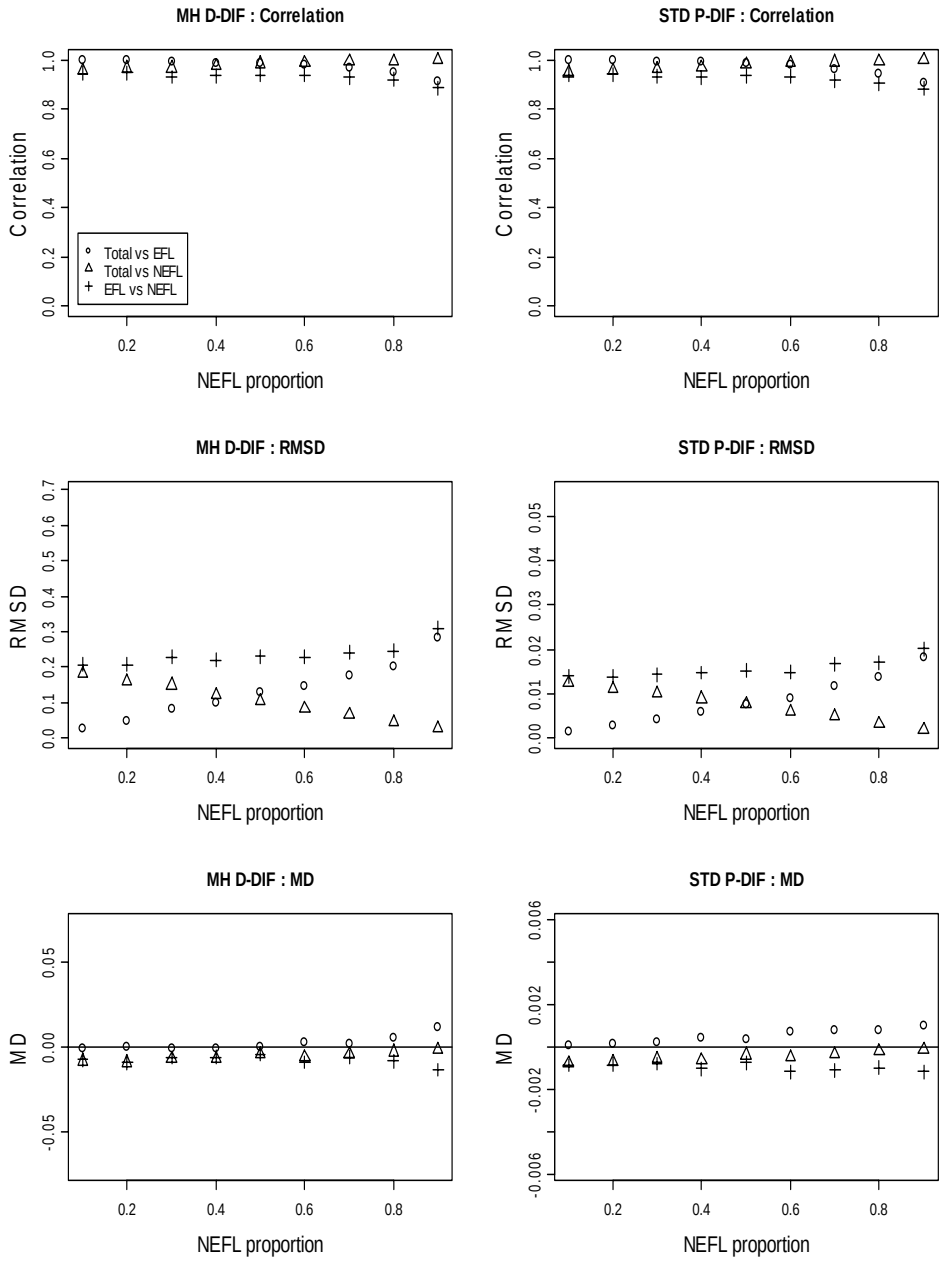

Figure 7. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

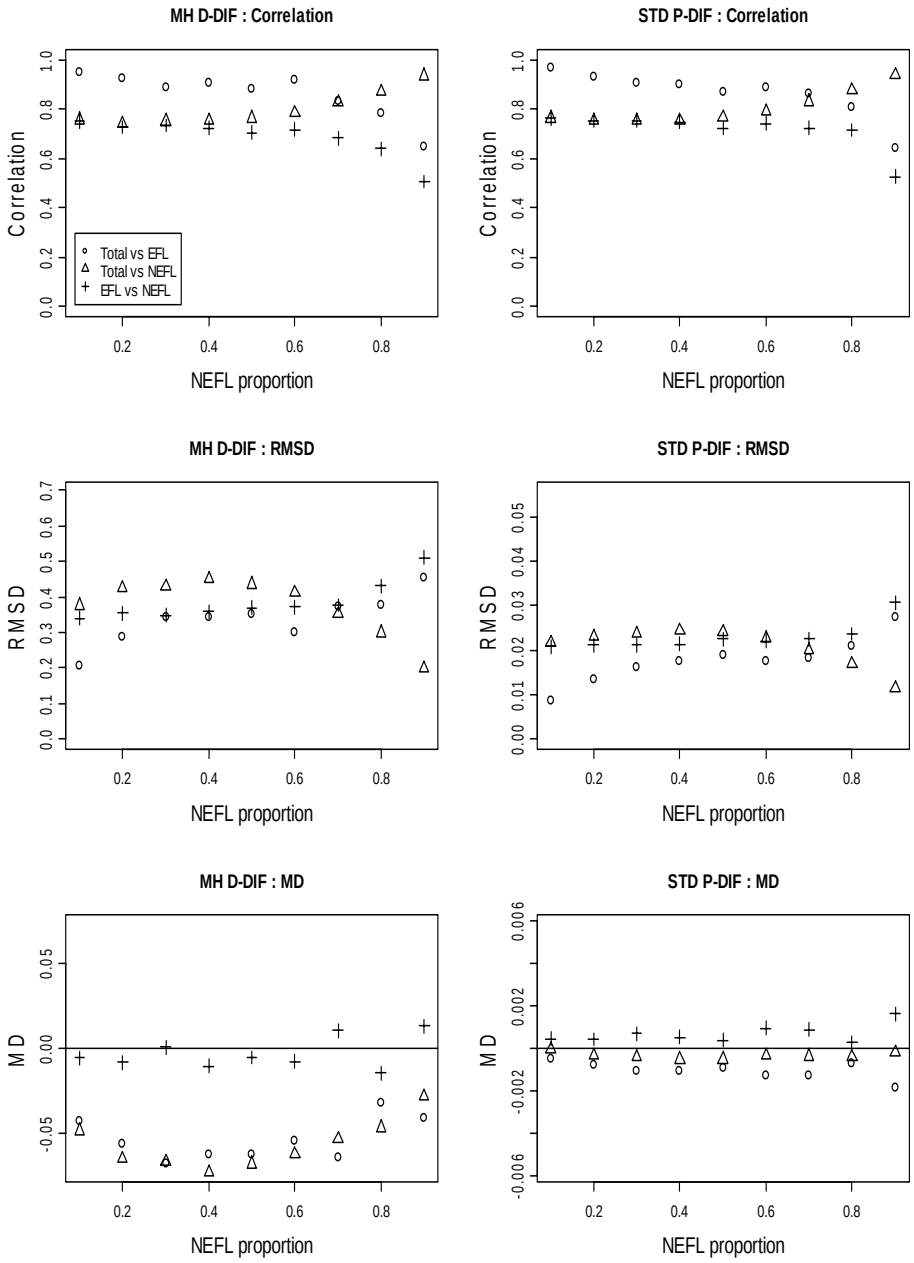

Figure 8. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference RMSD = root mean squared difference STD P-DIF = standardized P-difference.

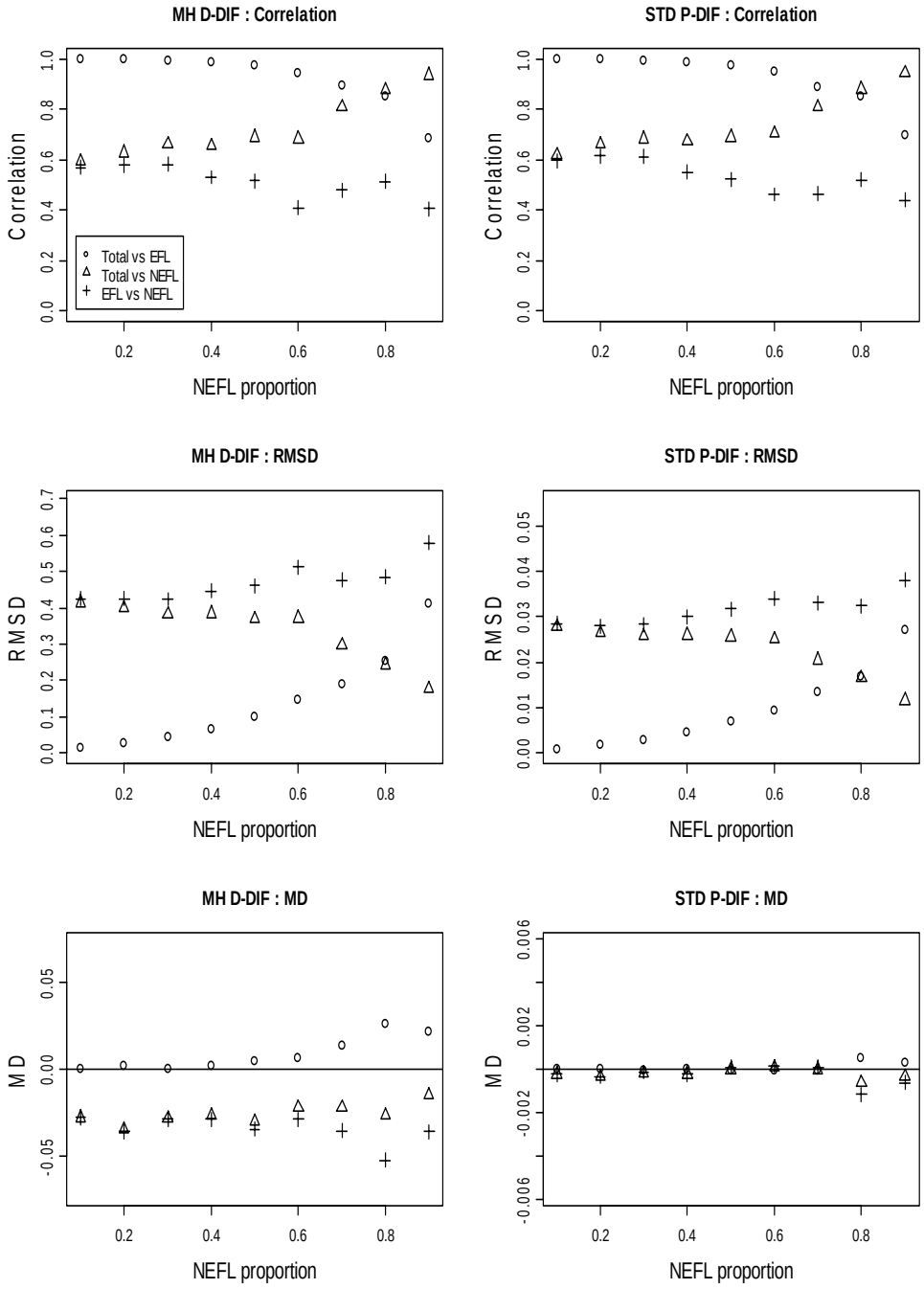

Figure 9. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

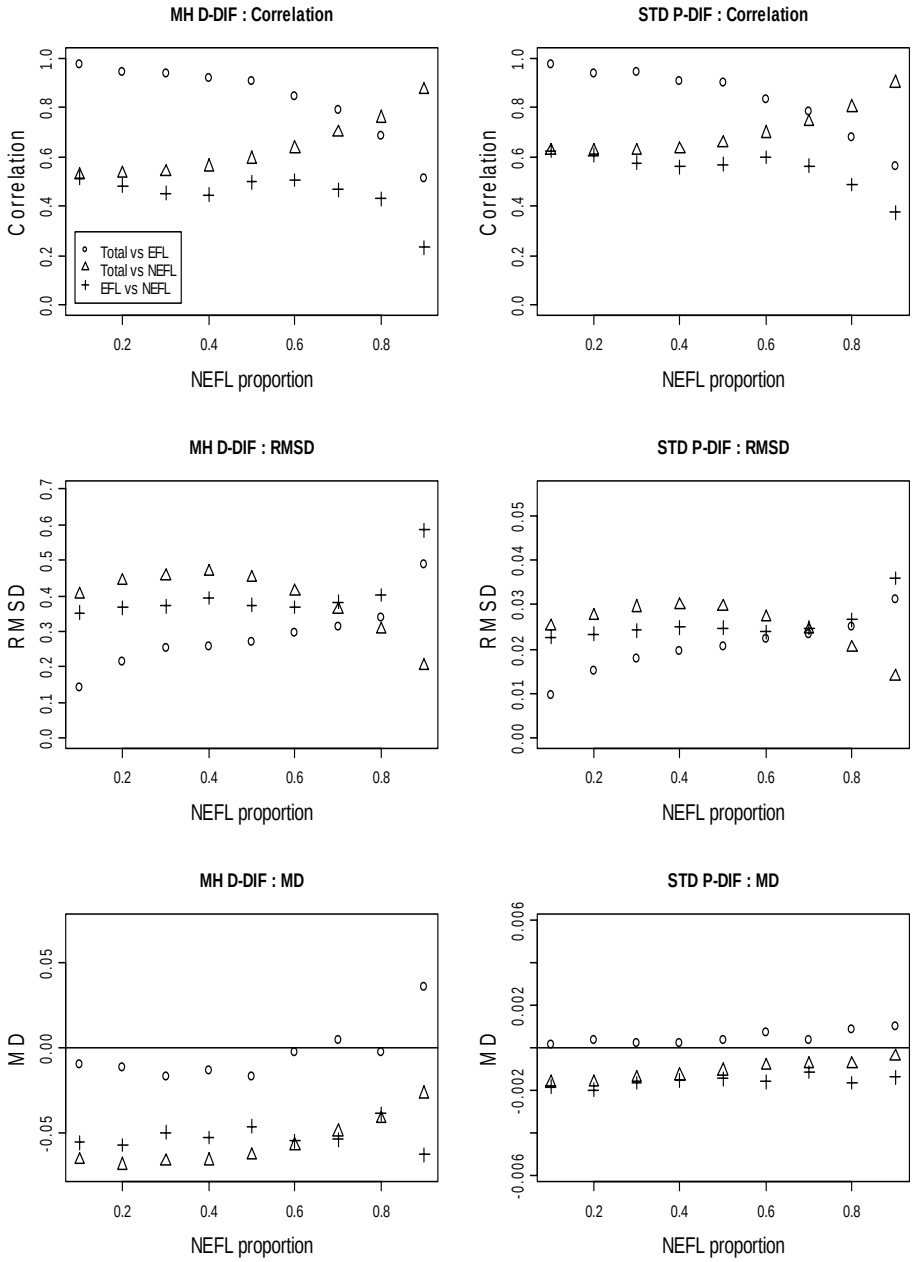

Figure 10. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

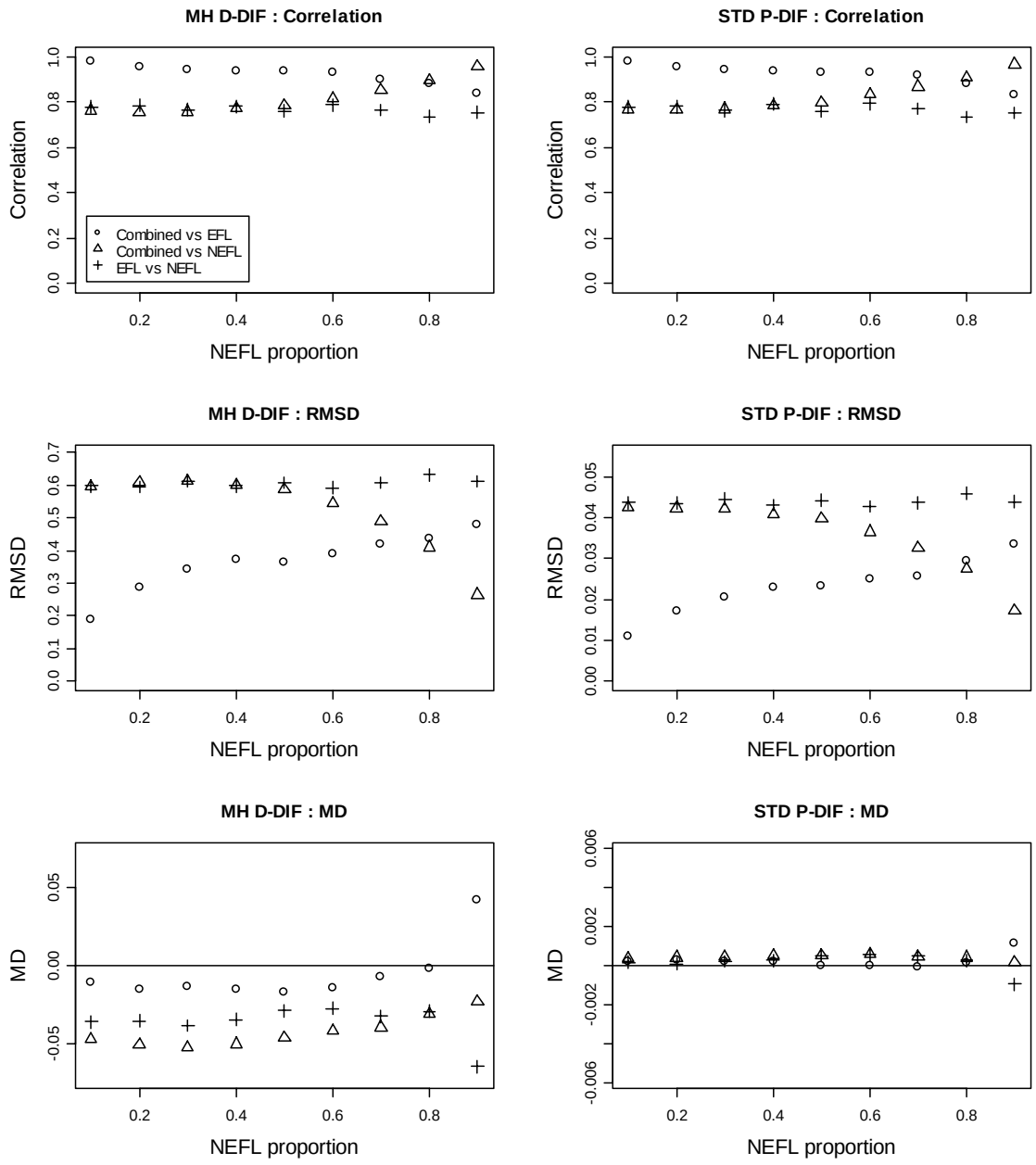

Figure 11. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

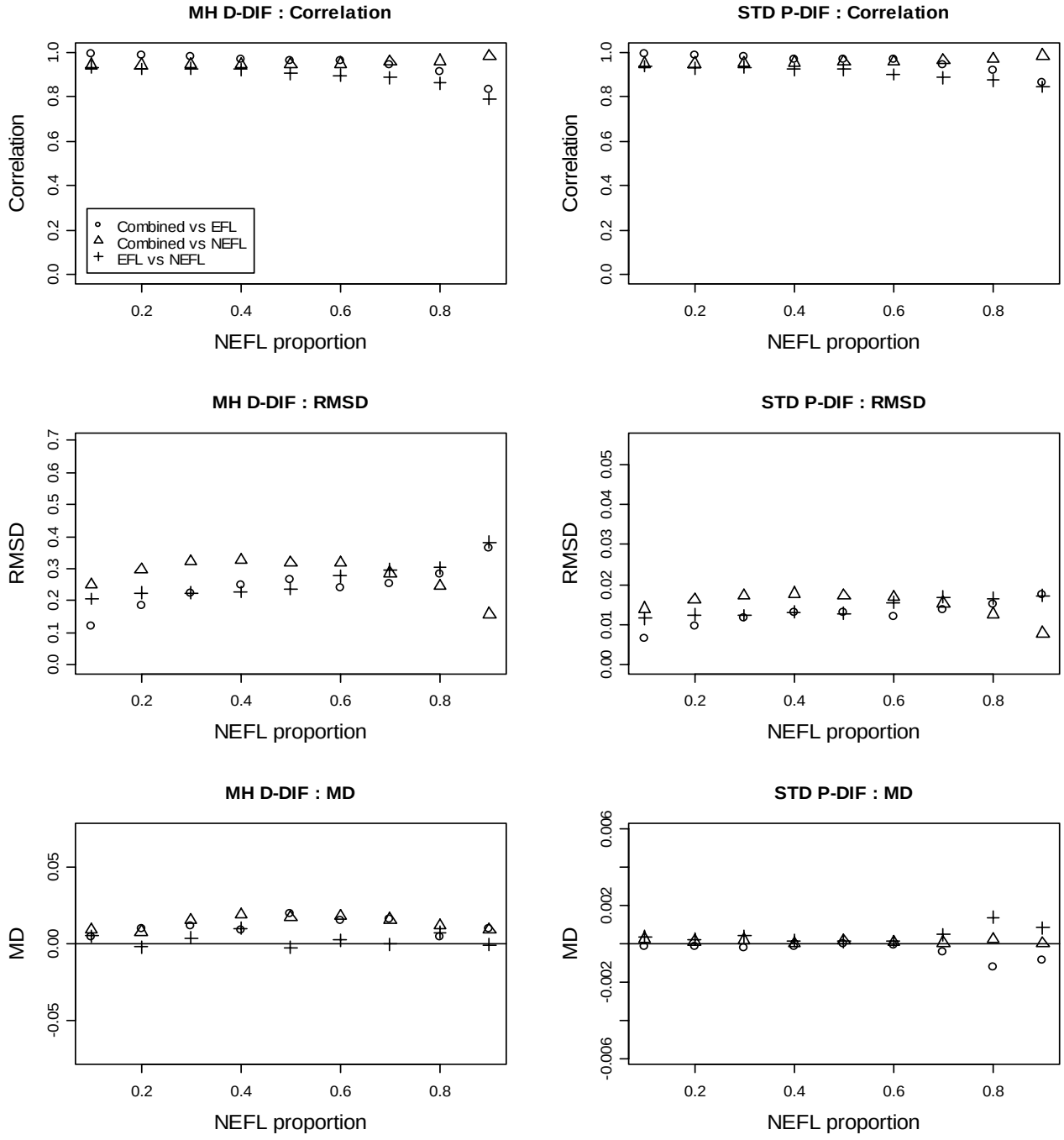

Figure 12. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

- The association between the DIF statistics for EFL-NEFL mostly becomes weaker (correlation decreases and root mean squared difference increases) as the NEFL proportion increases. This result may be due to the reduction in sample size with an increase in the NEFL proportion.
- Comparisons of Figures 7 to 10 (and of similar figures in the appendix) show that among the four types of DIF analysis, the association among DIF statistics for Total-EFL, Total-NEFL, and EFL-NEFL is strongest for female/male DIF, followed by Asian American/White, Black/White and Hispanic/White. For example, the correlations for Total-EFL, Total-NEFL, and EFL-NEFL are all close to 1 for female/male DIF for all the tests, while the correlations are as low as 0.2 for Hispanic/White DIF. (See Saturday critical reading in Figure 10.) This finding is consistent with that from Figures 1 to 6 and once again indicates that female/male DIF is affected the least by EFL status and that Hispanic/White DIF is affected the most.
- Focusing on the sections, a comparison of Figures 8, 11, and 12 (and of other similar figures in the appendix) shows that the association between DIF statistics is stronger for mathematics compared to critical reading or writing. This finding is surprising because we expected the DIF analysis for mathematics to be more affected by EFL status than critical reading or writing. In mathematics, English proficiency is a construct-irrelevant measure, while it is not so for critical reading or writing. We will investigate this issue further, but we expect that it has something to do with the fact that mathematics measures tend to be unidimensional, as evidenced by little DIF on SAT mathematics items over the decades, and the fact that our surrogate (EFL) for English proficiency may cause many SPEs to be wrongly classified as NEFLs, especially in the high-scoring Asian American NEFL group.
- The results for Total-EFL will be most pertinent to the practical question of whether the DIF analysis should be performed on the total group or on the EFL group only (remember that PSAT/NMSQT currently performs its operational DIF analysis on the EFL group). Figures 7 to 12 show that the DIF results for Total and EFL are essentially the same across subgroups for the proportion of NEFL examinees seen in

actual PSAT/NMSQT data. However, as the proportion of NEFL examinees increases, for example, beyond 0.4, the DIF results become somewhat affected for ethnic/racial DIF and not for female/male DIF. In addition, there may be outliers among individual items as suggested by selected panels in Figures 1 to 6 that would be more prominent as the NEFL proportion increases.

- The results for the MH D-DIF statistic are very similar to those for the STD P-DIF statistic. Some exceptions exist, mostly for the Saturday administration. For example, Figures 8, 9, and 10 show that for Saturday critical reading for Asian American/White, Black/White, and Hispanic/White , the mean differences for STD P-DIF are close to 0, while they are not so for MH D-DIF. To a certain extent, similar results are found for Saturday mathematics and Saturday writing.

5. Discussion and Conclusions

We employed data from the PSAT/NMSQT to study the relationship between the proportion of NEFL examinees in the analysis sample and standard DIF procedures. For these data, the DIF results across subgroups do not seem to be affected by the NEFL proportion seen in actual data (about 7.8% for PSAT/NMSQT). However, as the proportion of NEFL examinees increases, for example, beyond 0.4, the DIF results become somewhat affected for ethnic/racial DIF. These results suggest that for PSAT/NMSQT, it does not matter much now whether the DIF analysis is performed on the EFL examinees or on the total population, but it may matter in the future when the proportion of NEFL examinees taking PSAT/NMSQT increases. Hence, we recommend monitoring the proportion of NEFL examinees taking the PSAT/NMSQT.

Our study has several limitations. First, the criterion used in this study to categorize examinees as EFL or NEFL was the examinees' self report on a question asking if English was their first language. This question is printed on the test form and provides ready-to-use information. However, it is not an accurate measure of examinees' true English proficiency status, which is the construct that is likely to affect DIF analyses. For example, those who were born in another country and came to the United States at a very young age may indeed be proficient in English even though it is not their first language.

Second, while we used the PSAT/NMSQT data because of the large sample sizes, little DIF is included in these data. Though this situation demonstrates the high quality of the

PSAT/NMSQT items, this deficit is not beneficial for a study in which DIF statistics are the dependent variables. We did not observe much DIF, even for a synthetic subsample created under extreme conditions (such as an NEFL proportion of 0.9).

Third, we looked at only one test—PSAT/NMSQT—and at only two administrations of that test. Few tests have the volumes needed to provide the large numbers of NEFL examinees that are essential for a study like ours.

Finally, forecasts about immigration trends are uncertain, which complicates projecting their potential effects on procedures for assessing fairness.

In the future, we hope to remedy the first limitation by studying a testing program that includes a more direct measure of English proficiency. In all likelihood, however, the test will not exhibit much DIF and we will be restricted to searching for patterns among A and B items. It is also unlikely that we will find enough test titles with sufficient numbers of examinees who are not proficient in English to obtain results that generalize to other forms of that test.

References

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.

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

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

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (ETS Research Rep. No. RR-09-05). Princeton, NJ: ETS.

Notes

[1] The Hispanic group is an aggregate group of individuals who indicated they were either Mexican or Mexican American; Puerto Rican; or Other Hispanic, Latino, or Latin American.

[2] 100 x 40,160 / (40, 160 + 474, 735) = 7.8%

[3] 100 x 40,160 / (40, 160 + 40, 160) = 50%

[4] The STD P-DIF value of .08 was determined from as the mean/sigma equivalent associated with an MH D-DIF statistic of 1.0 based on a mean/sigma linking of the two DIF statistics using these data. Dorans and Kulick (1986) used the same value in their early DIF work with the SAT.

Appendix

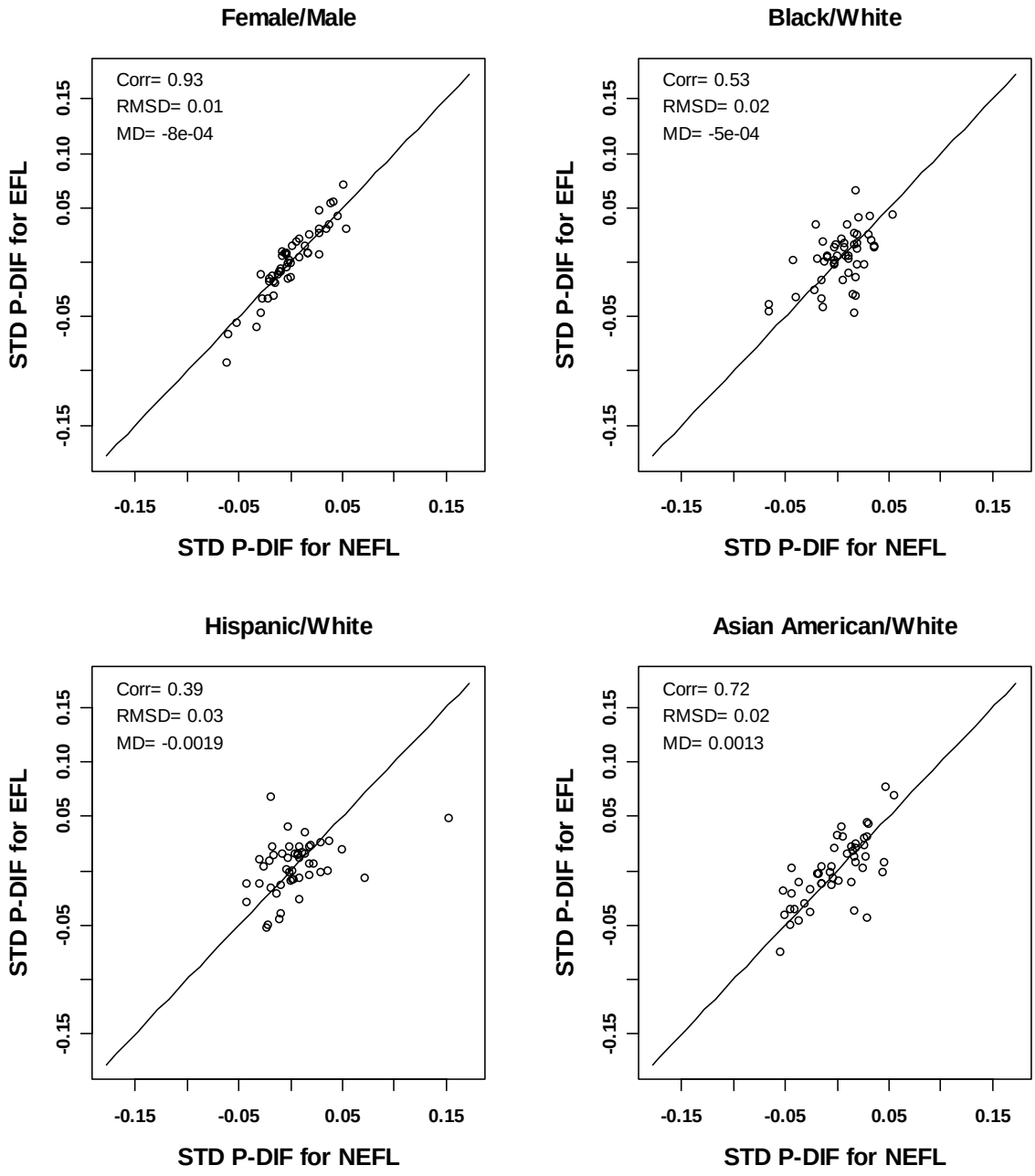

Figure A1. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday critical reading.**

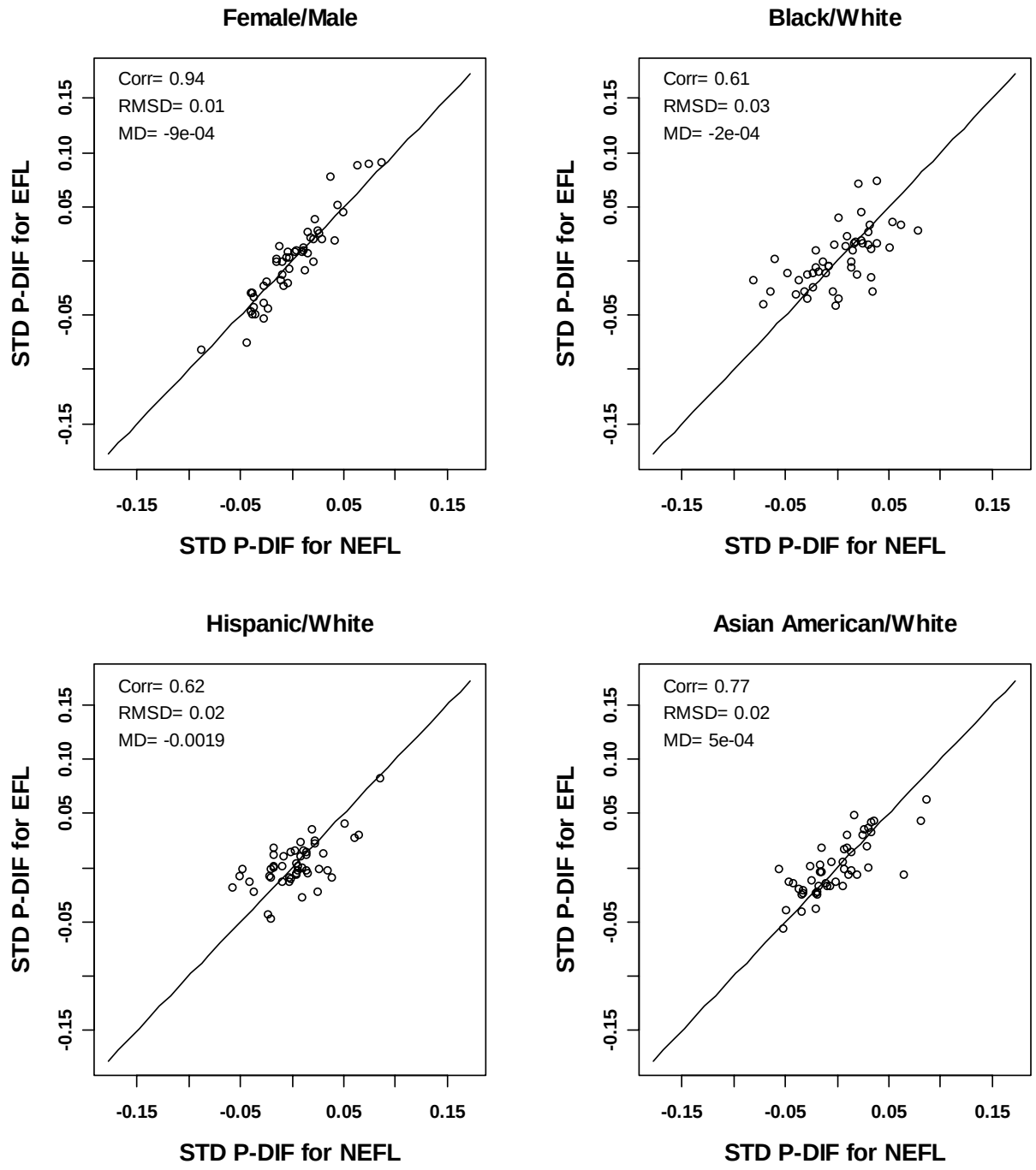

Figure A2. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday critical reading.**

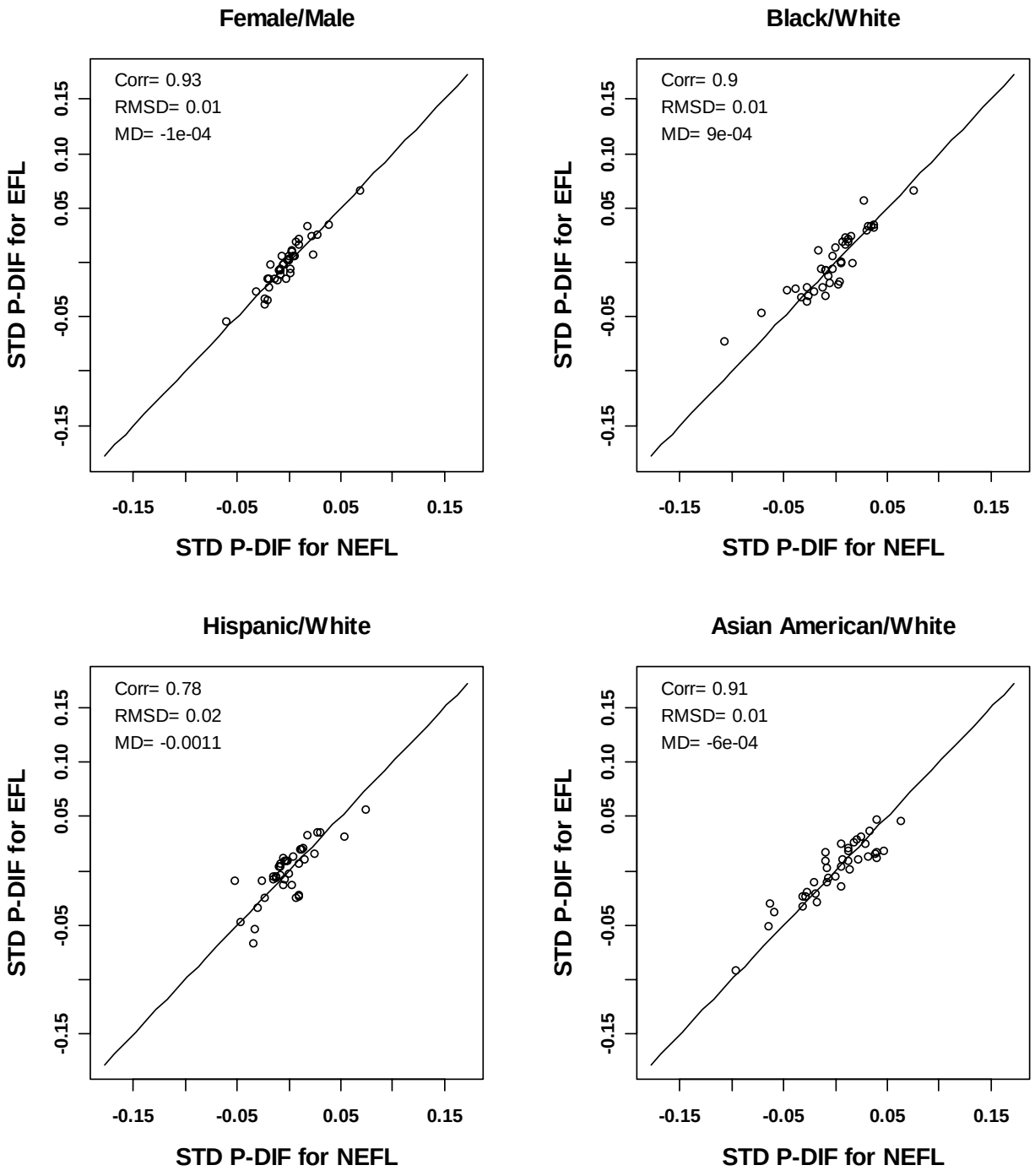

Figure A3. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday mathematics.**

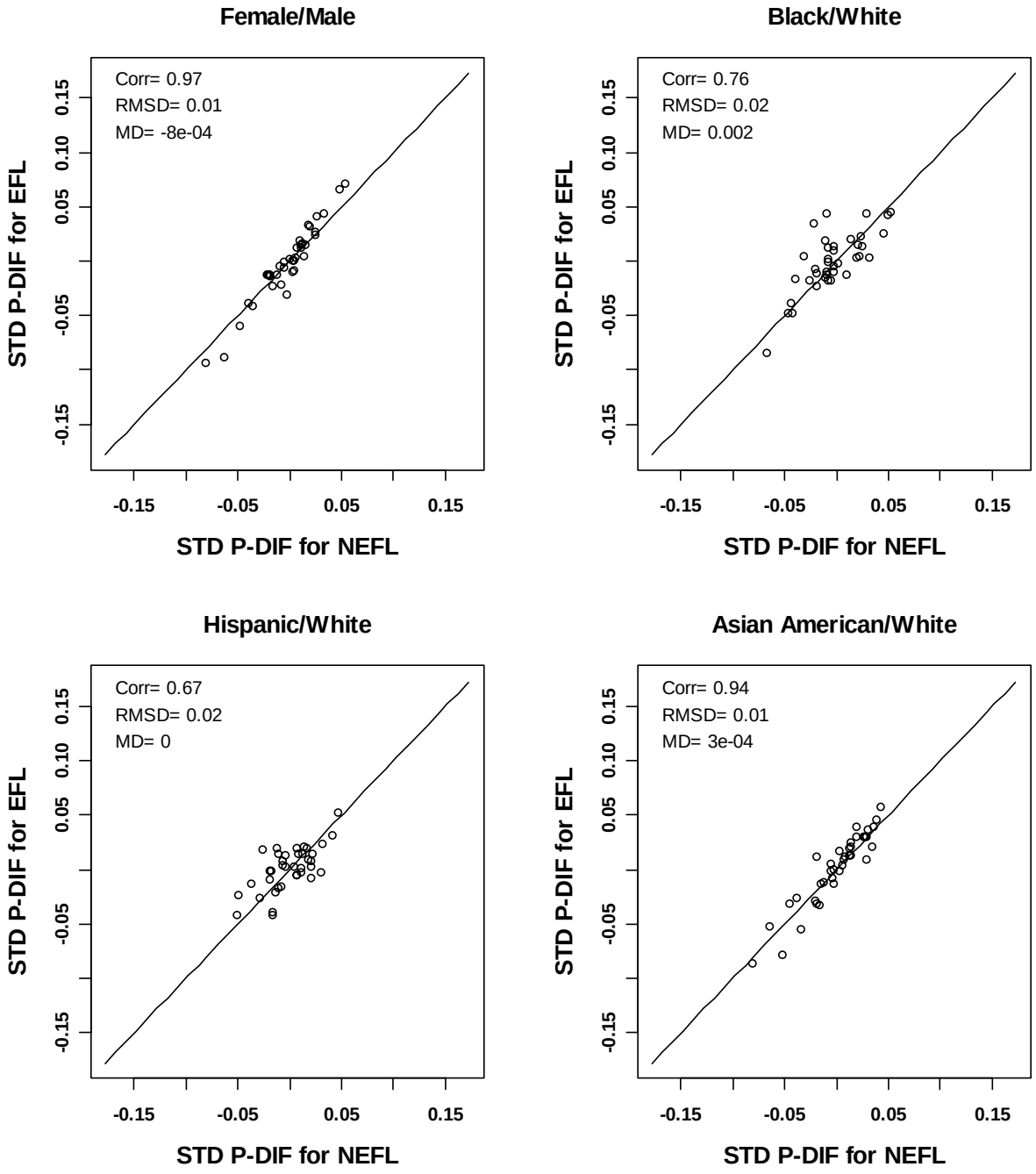

Figure A4. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday mathematics.**

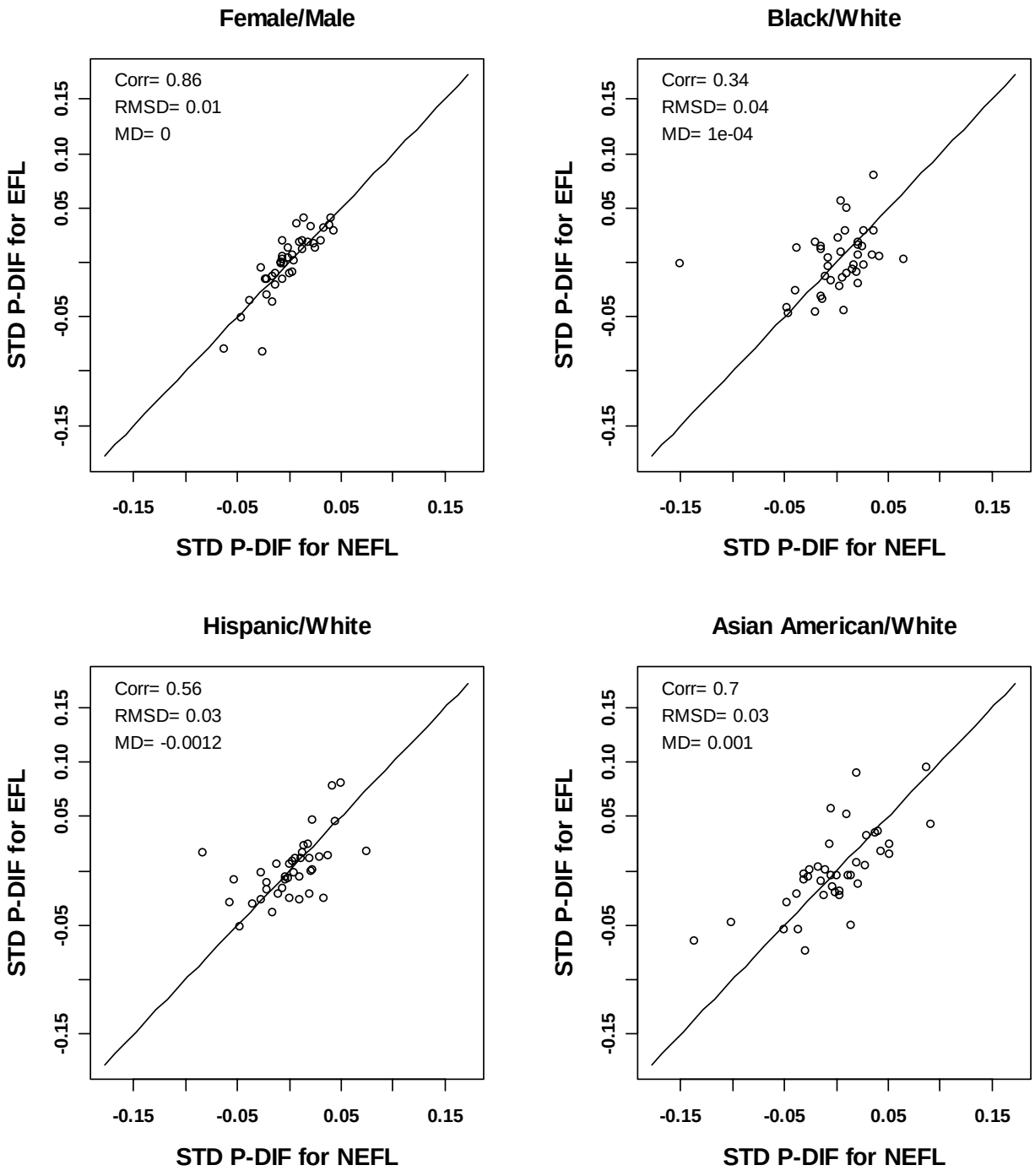

Figure A5. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference *(RMSD)* and mean difference *(MD)* for the English first language (EFL) population versus those for the not English first language (NEFL) population for Wednesday writing.**

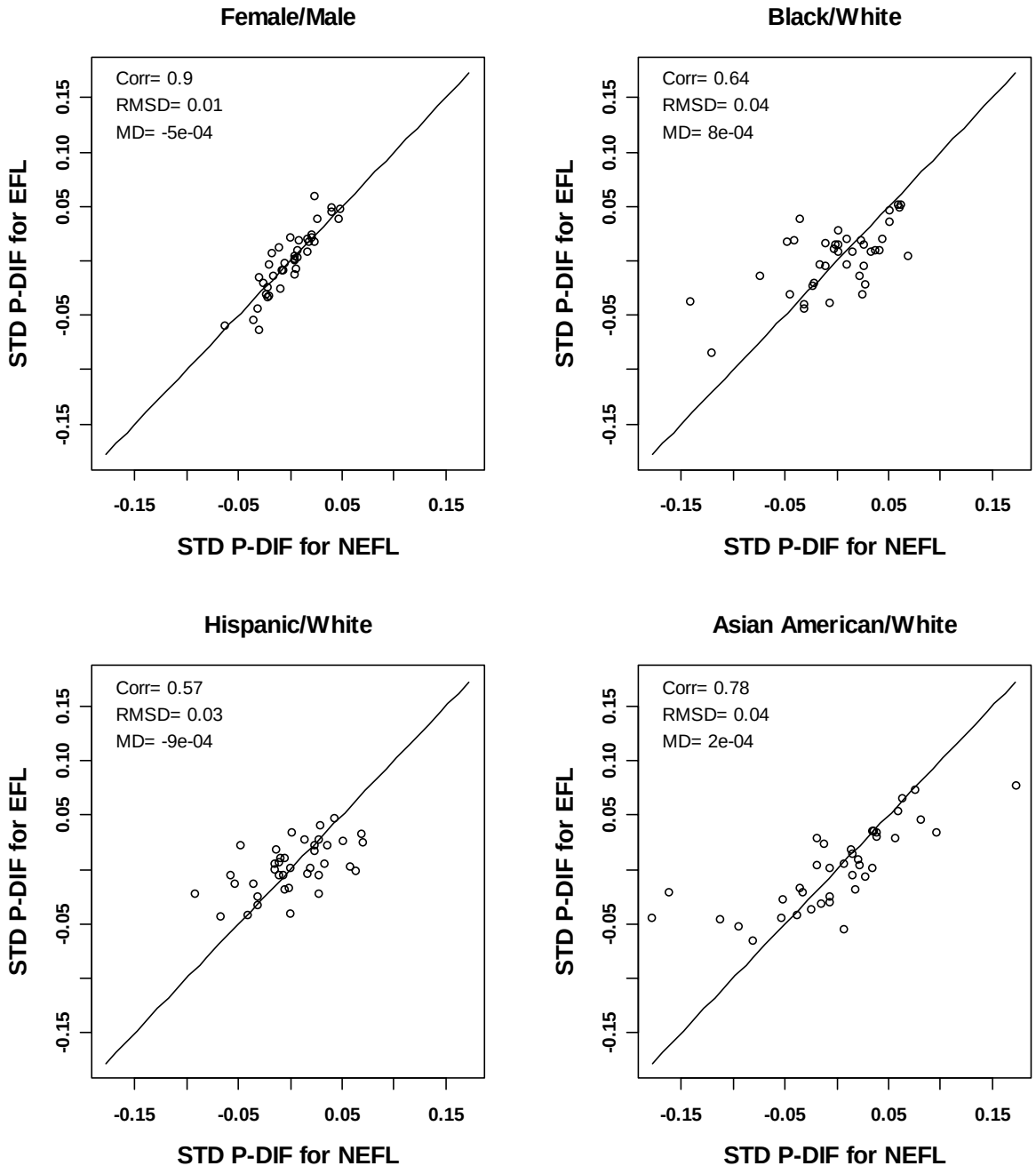

Figure A6. **Plot of the standardized P-difference (STD P-DIF) statistics with correlation, root mean squared difference (RMSD) and mean difference (MD) for the English first language (EFL) population versus those for the not English first language (NEFL) population for Saturday writing.**

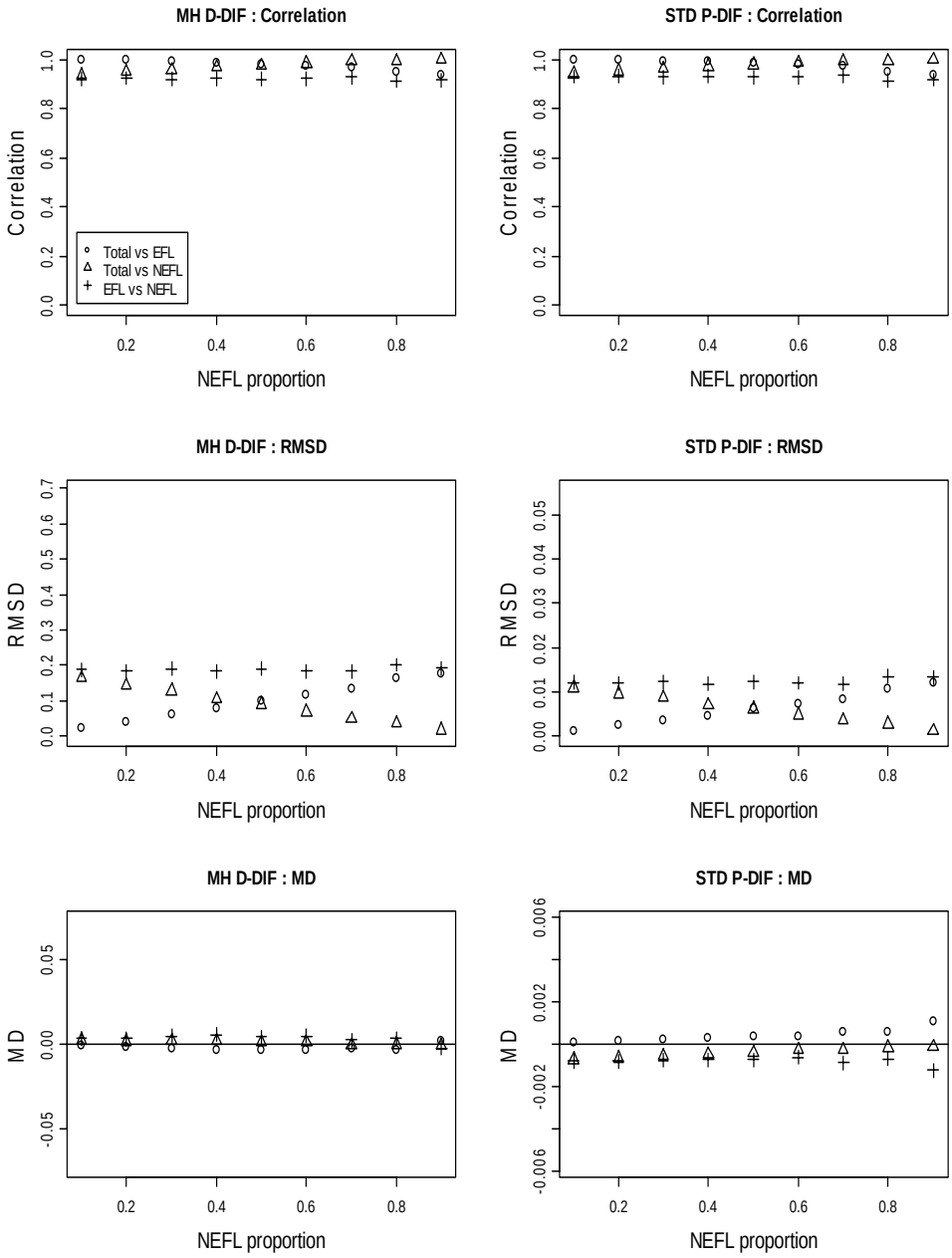

Figure A7. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday critical reading, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

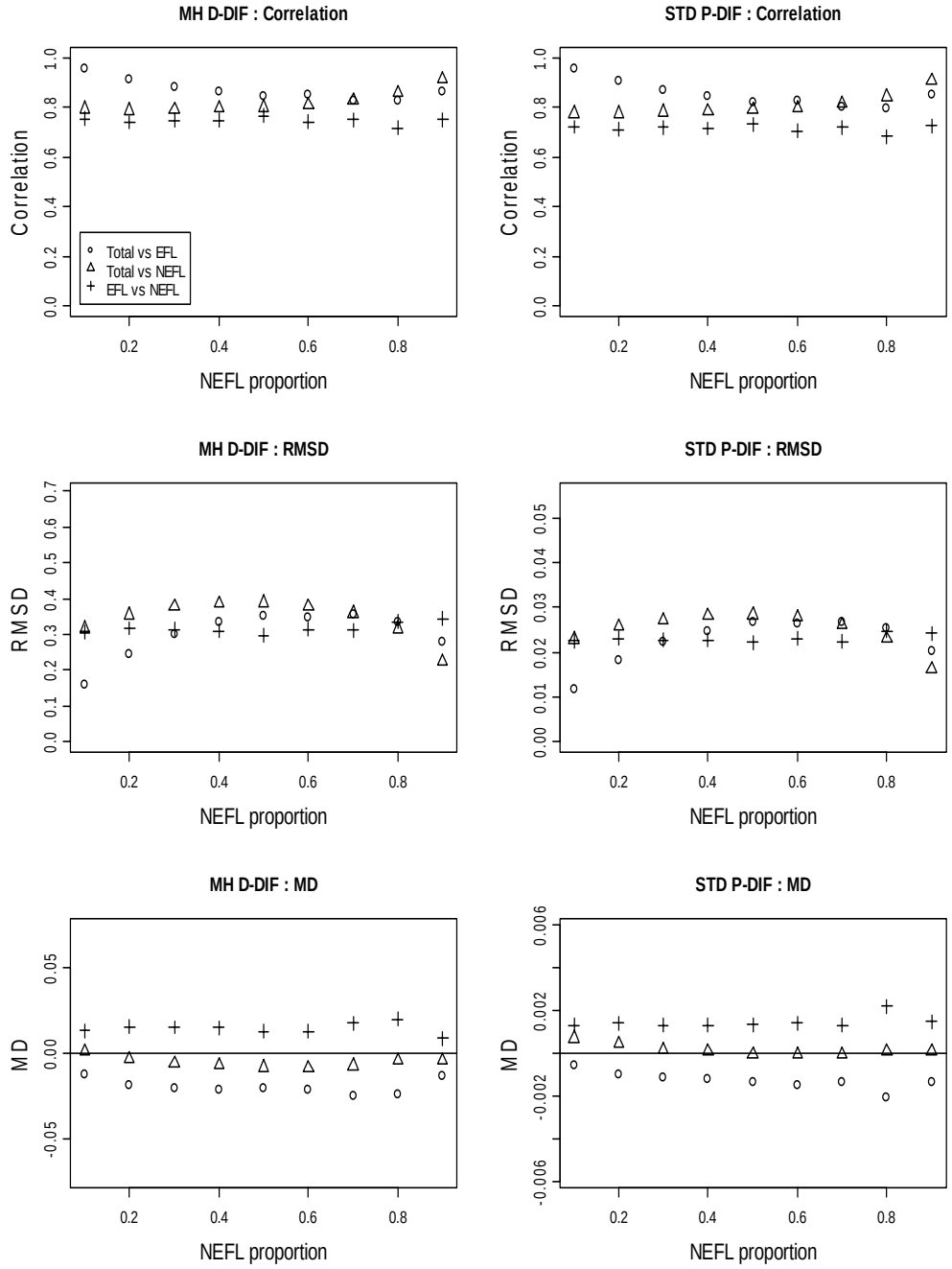

Figure A8. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday critical reading, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

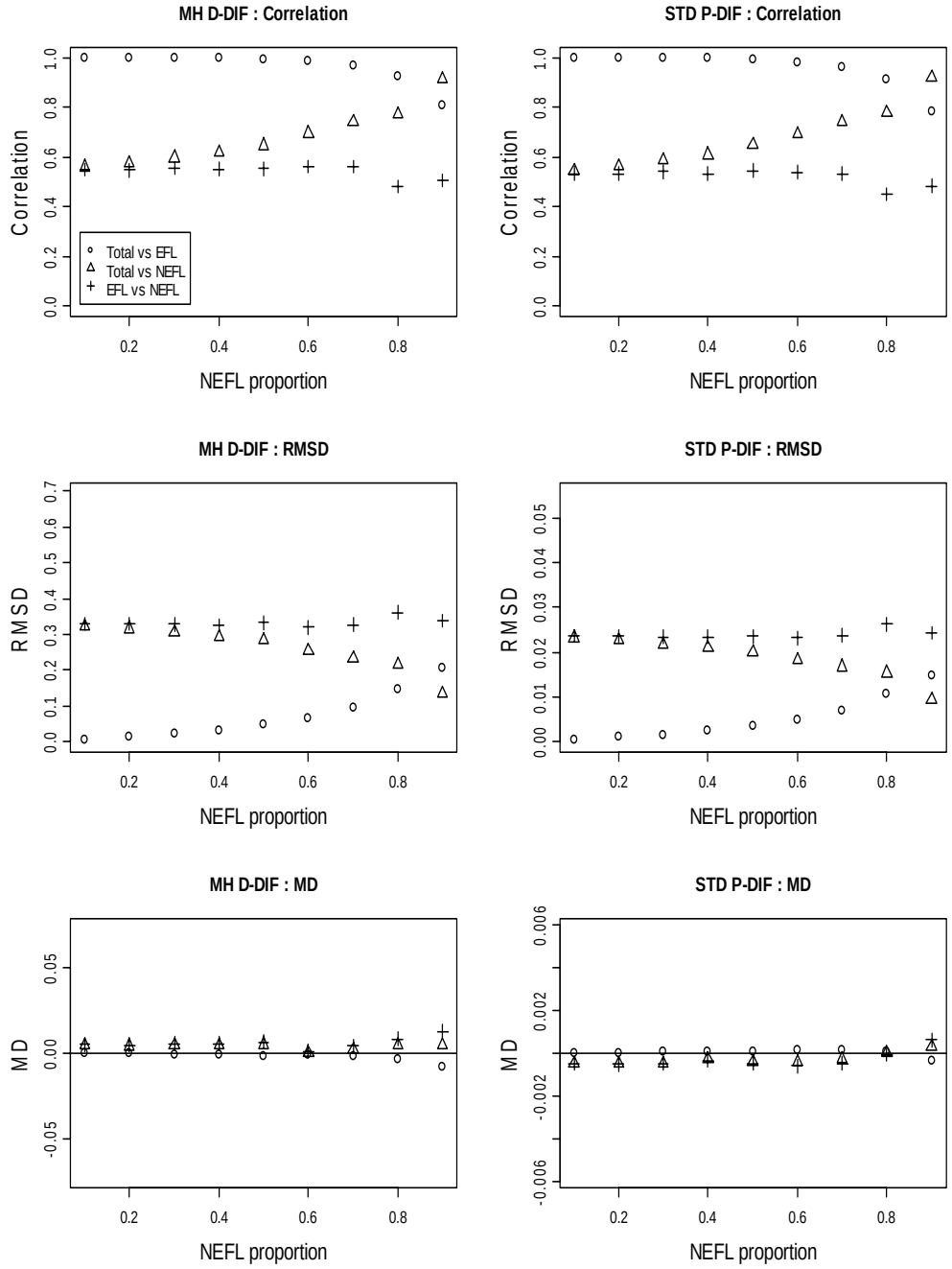

Figure A9. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday critical reading, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

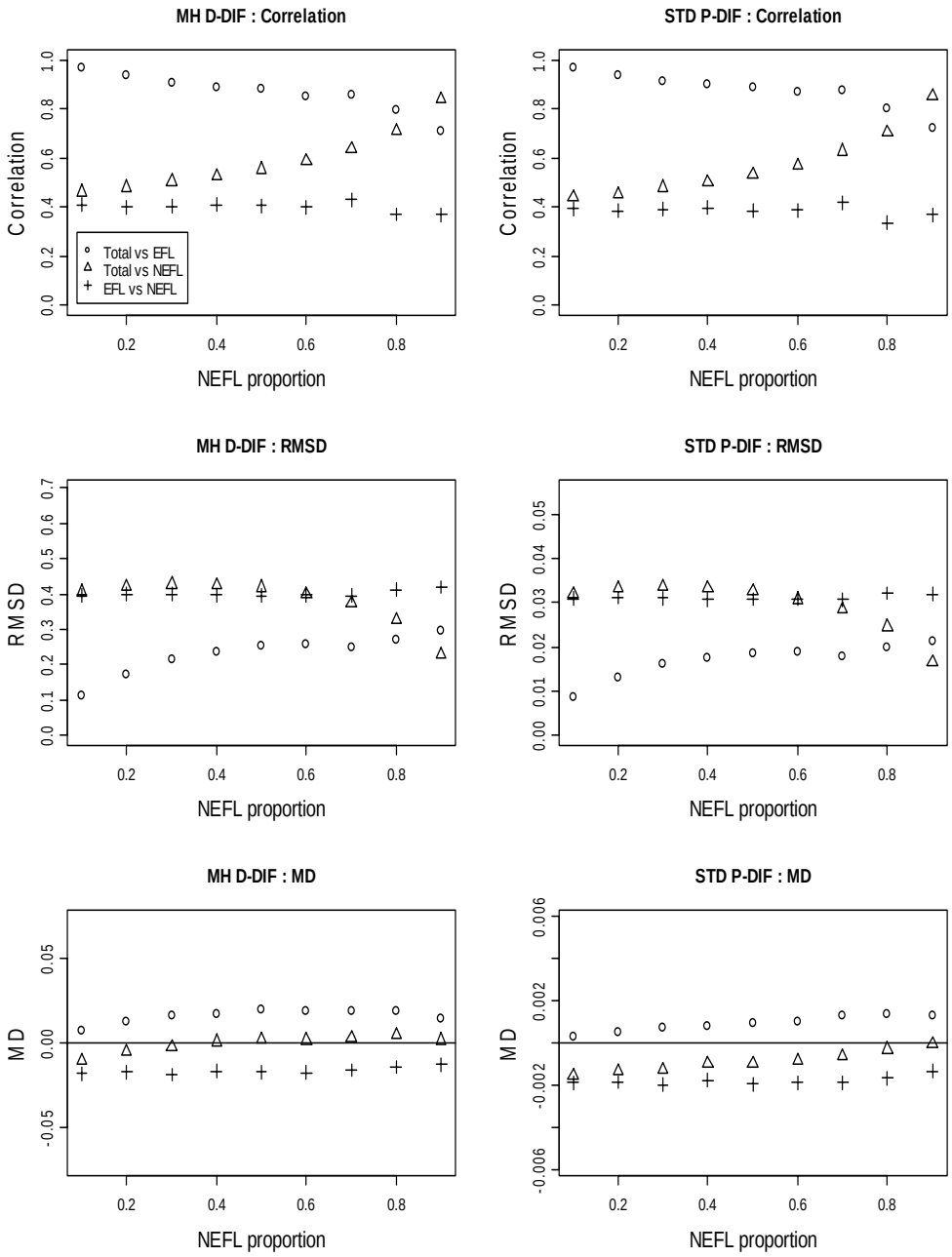

Figure A10. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday critical reading, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

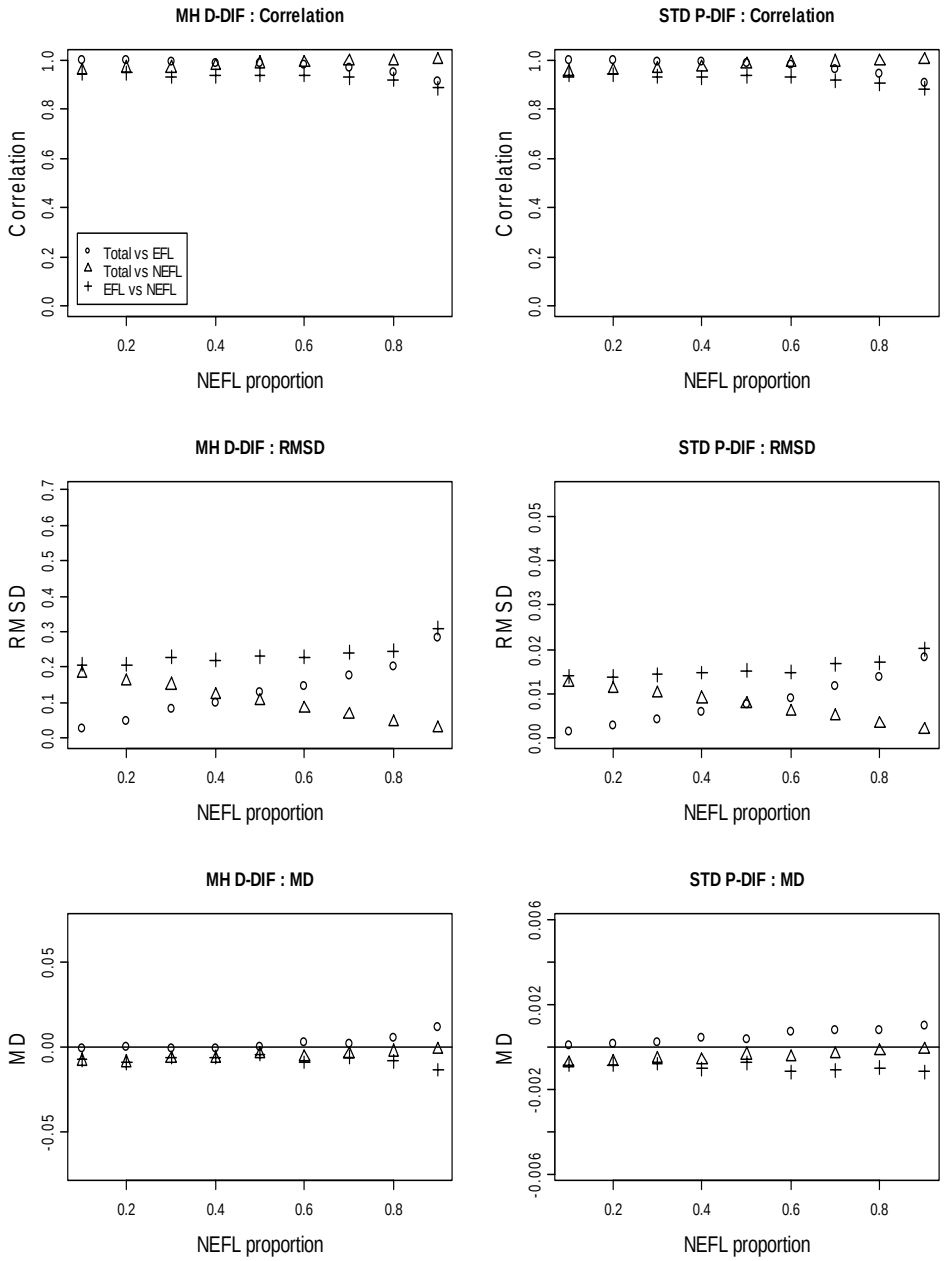

Figure A11. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

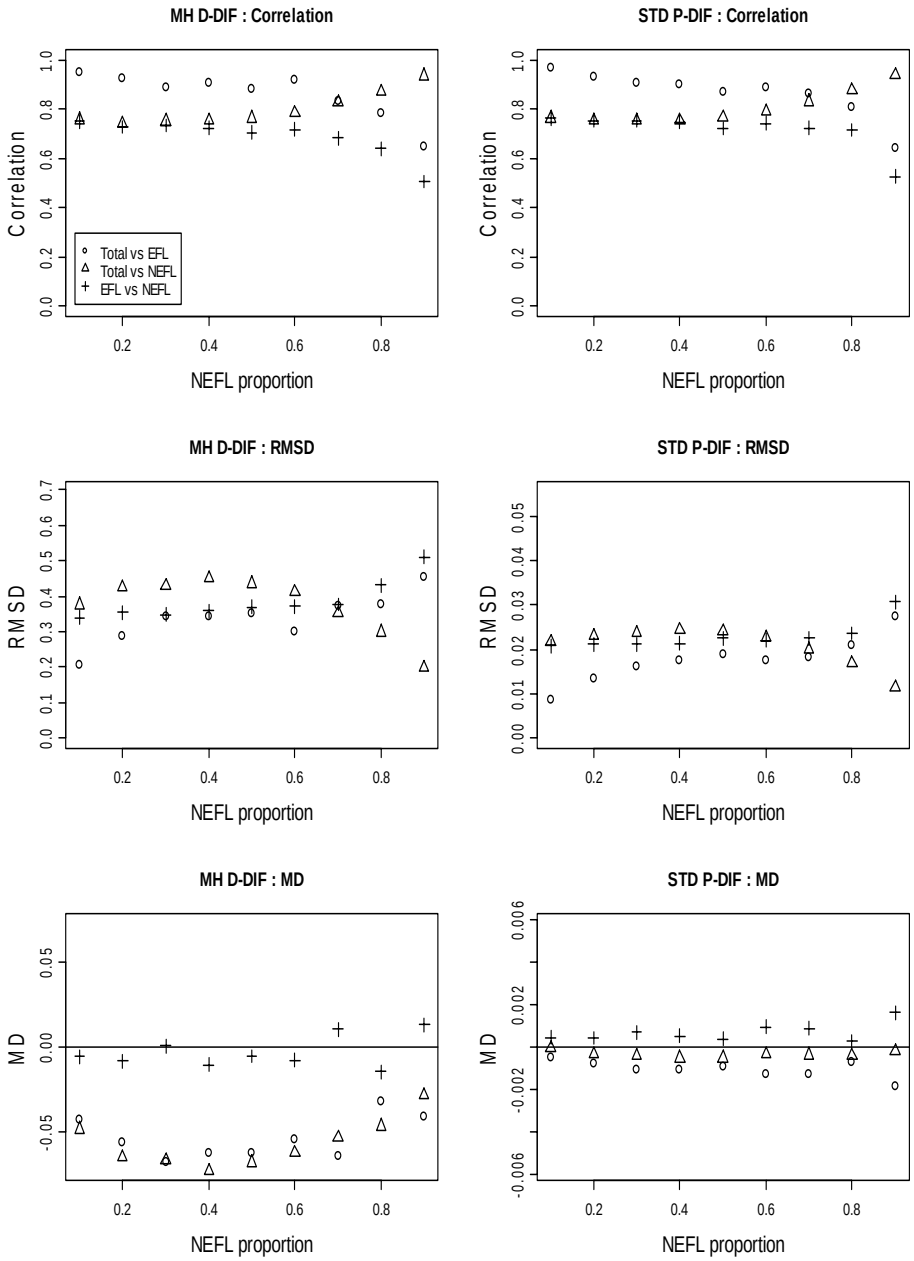

Figure A12. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

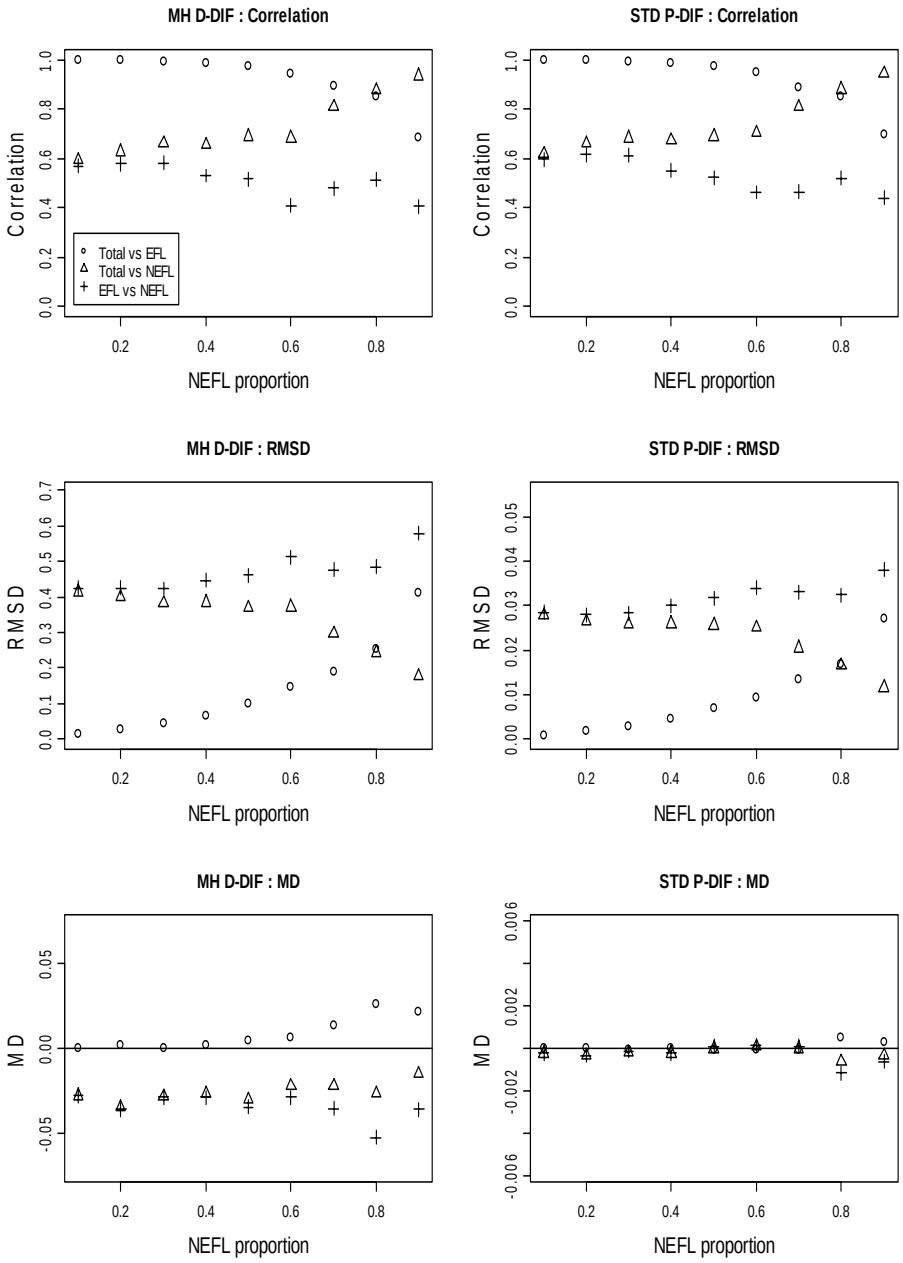

Figure A13. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

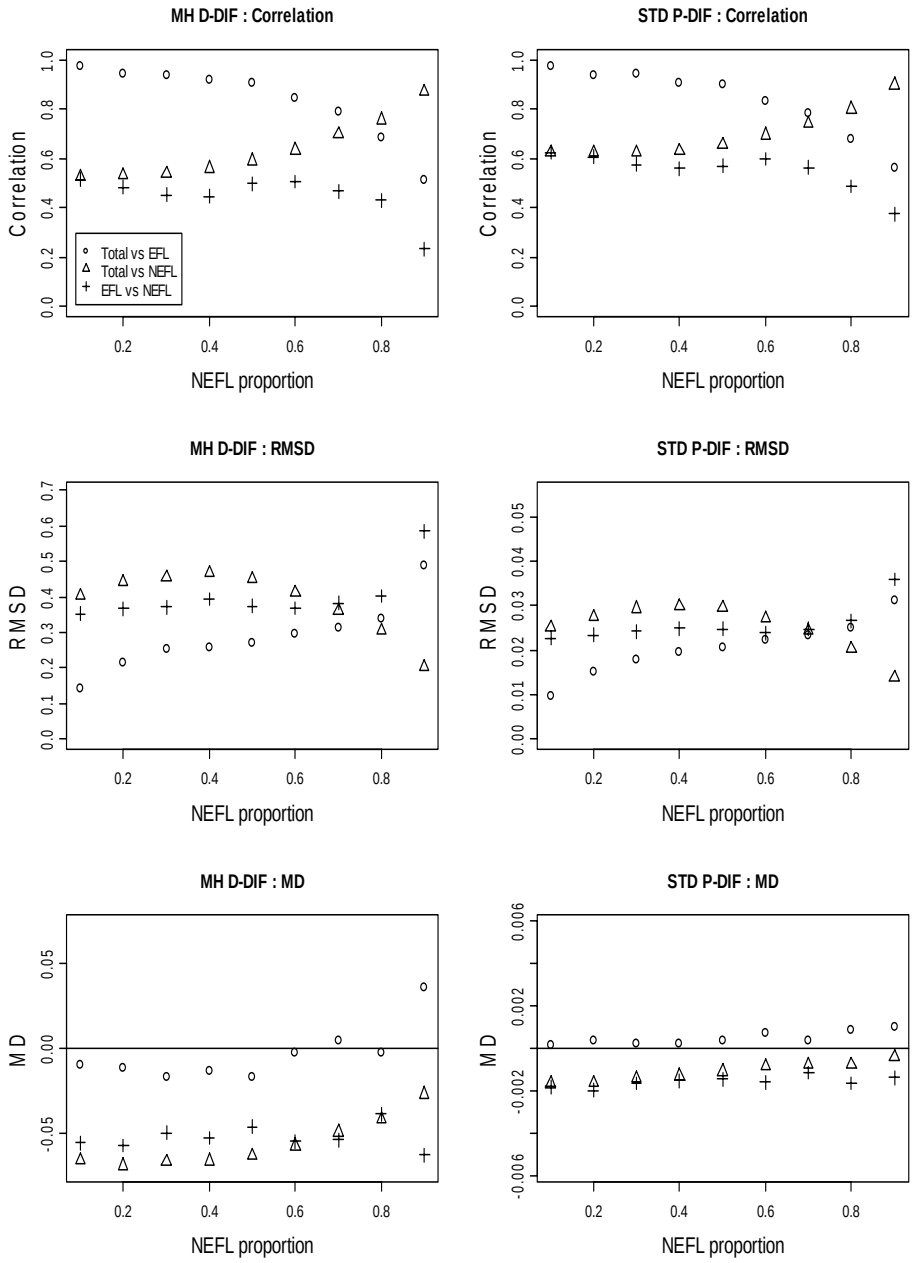

Figure A14. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday critical reading, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

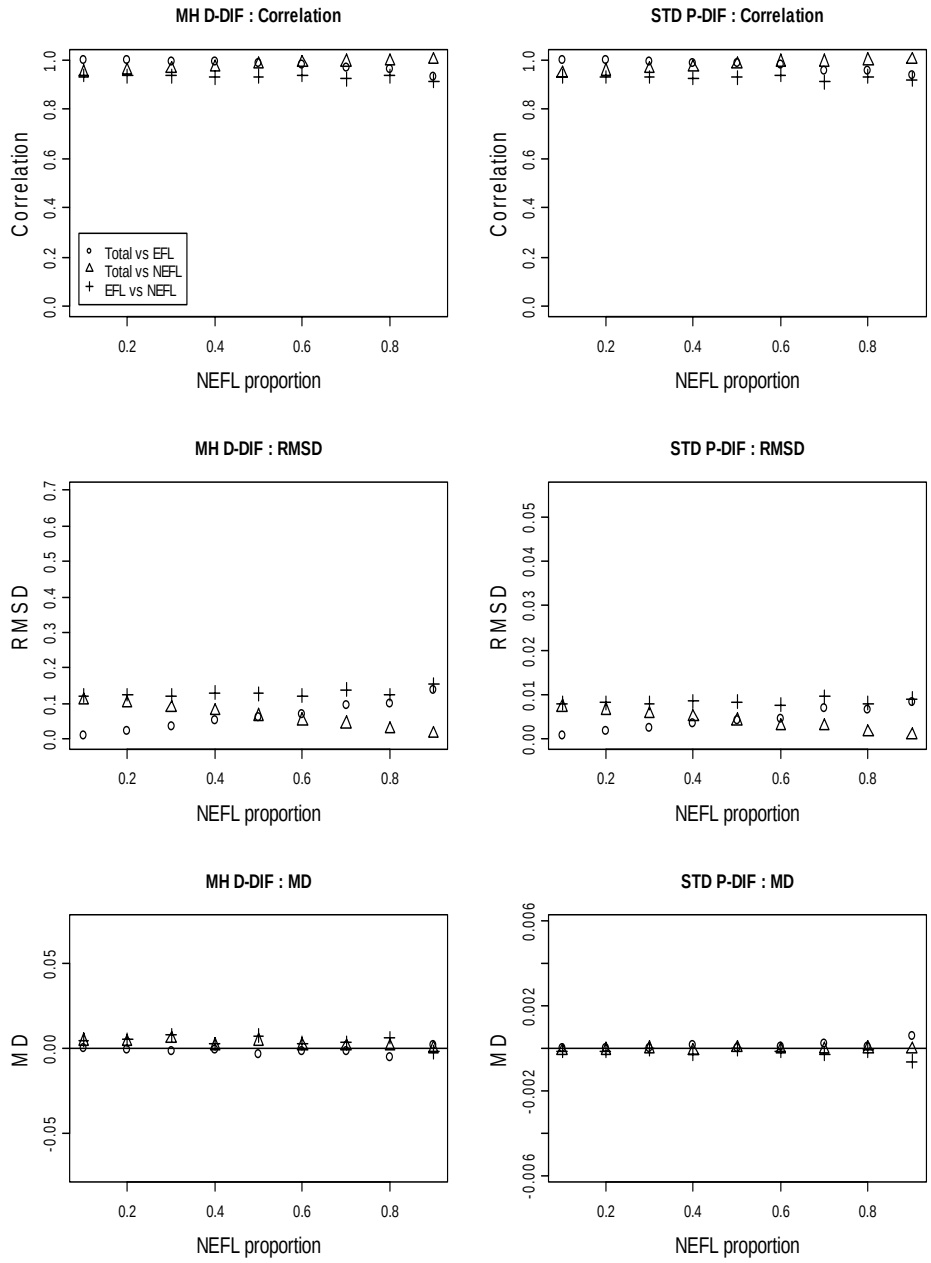

Figure A15. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday mathematics, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

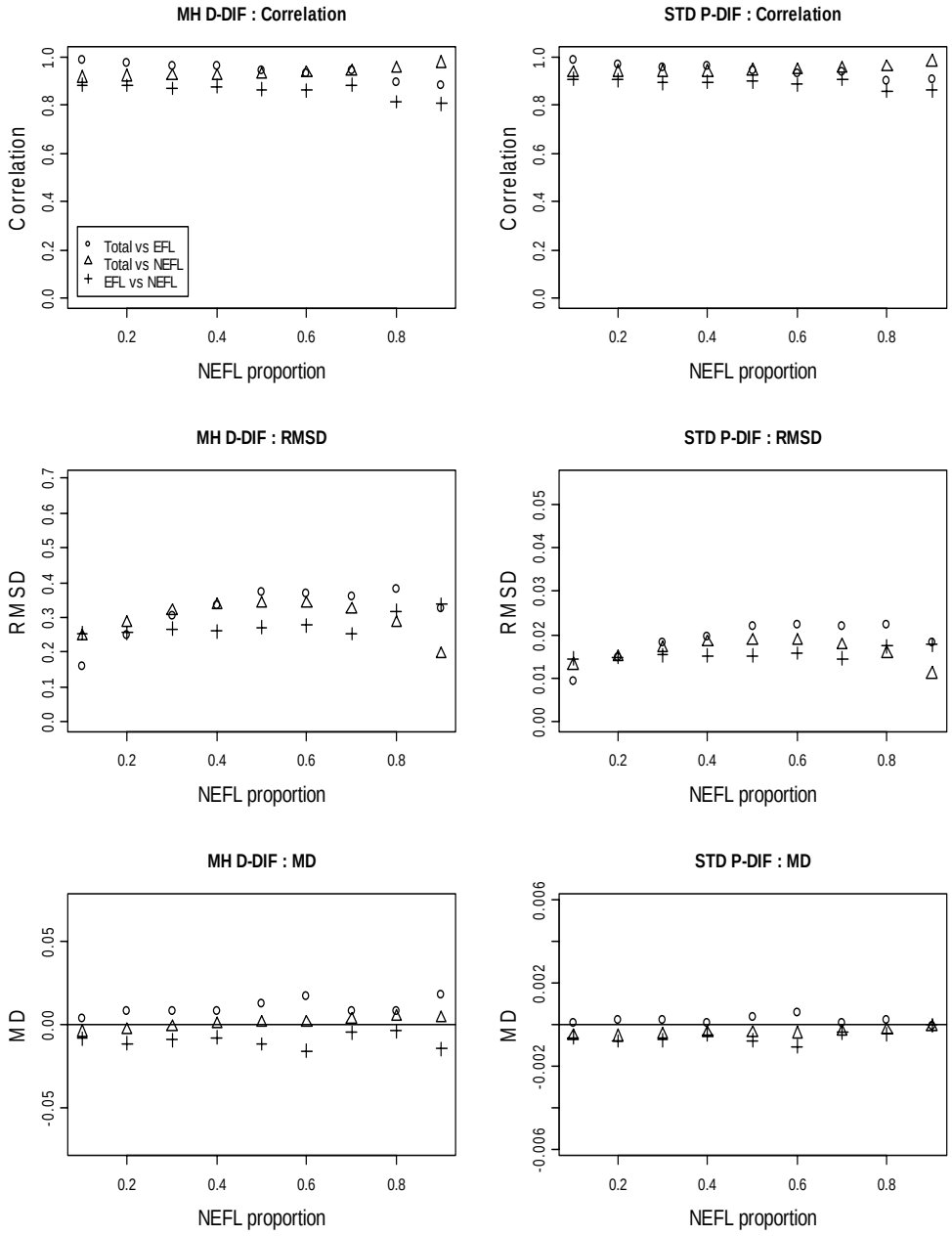

Figure A16. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday mathematics, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

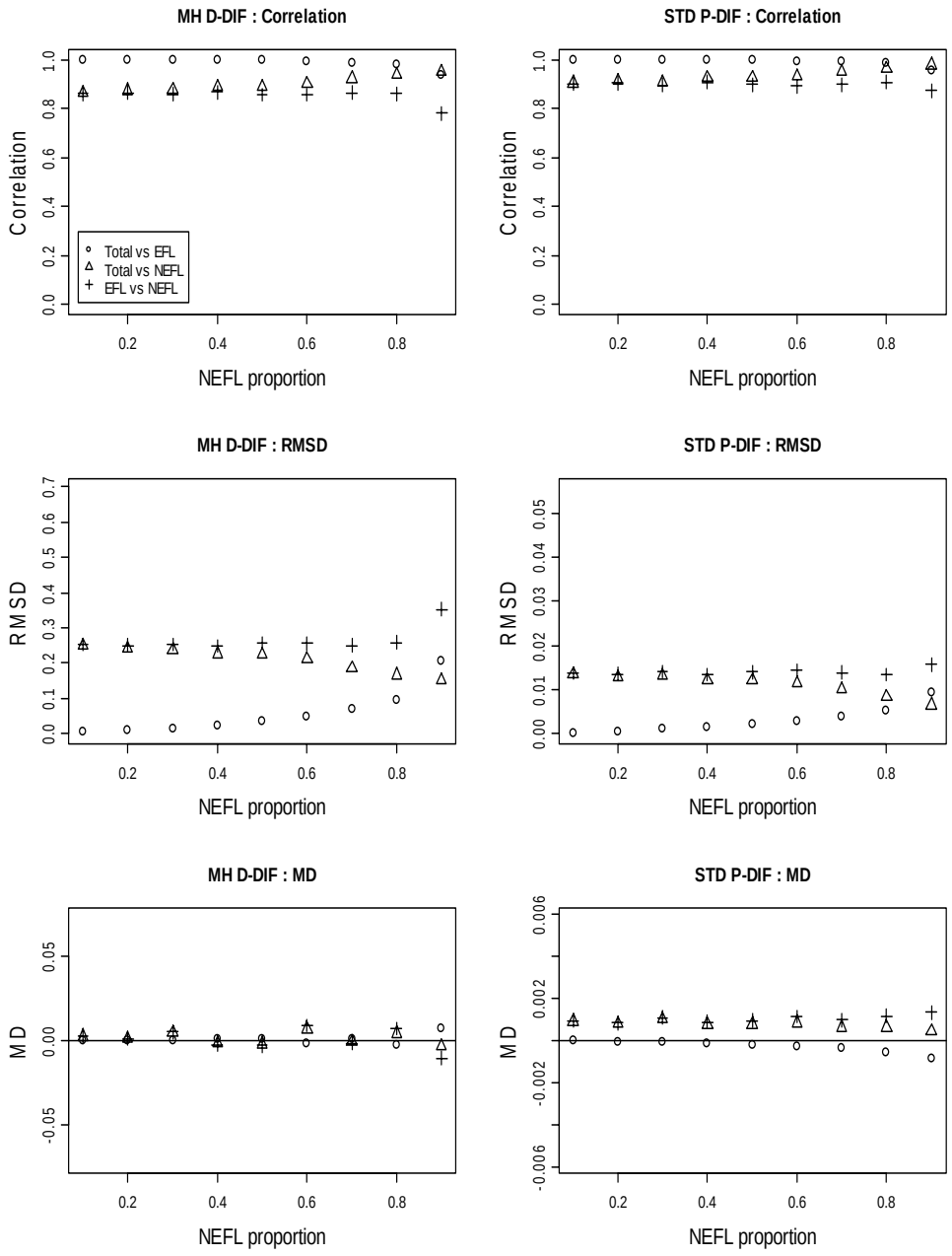

Figure A17. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday mathematics, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

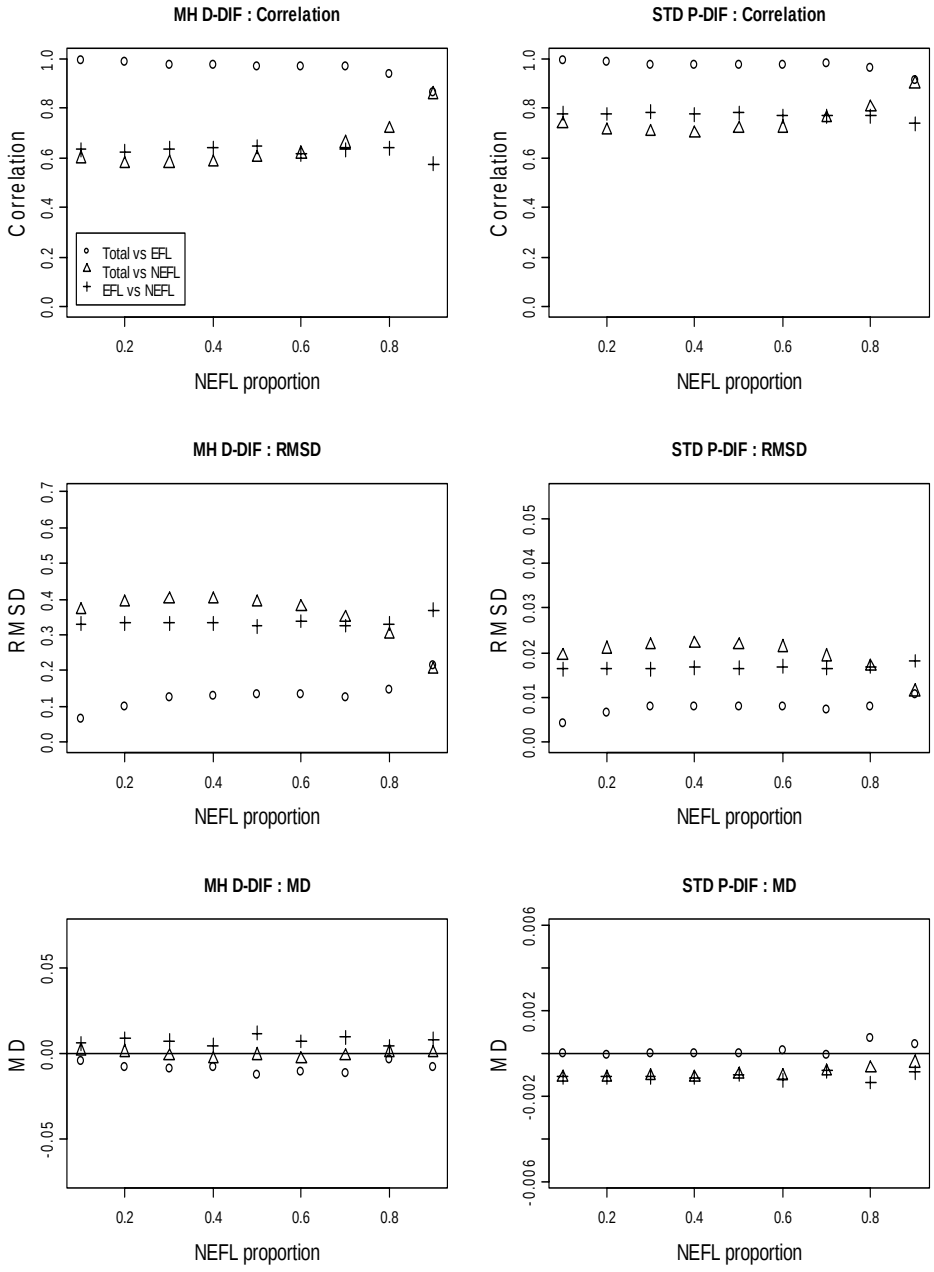

Figure A18. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday mathematics, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

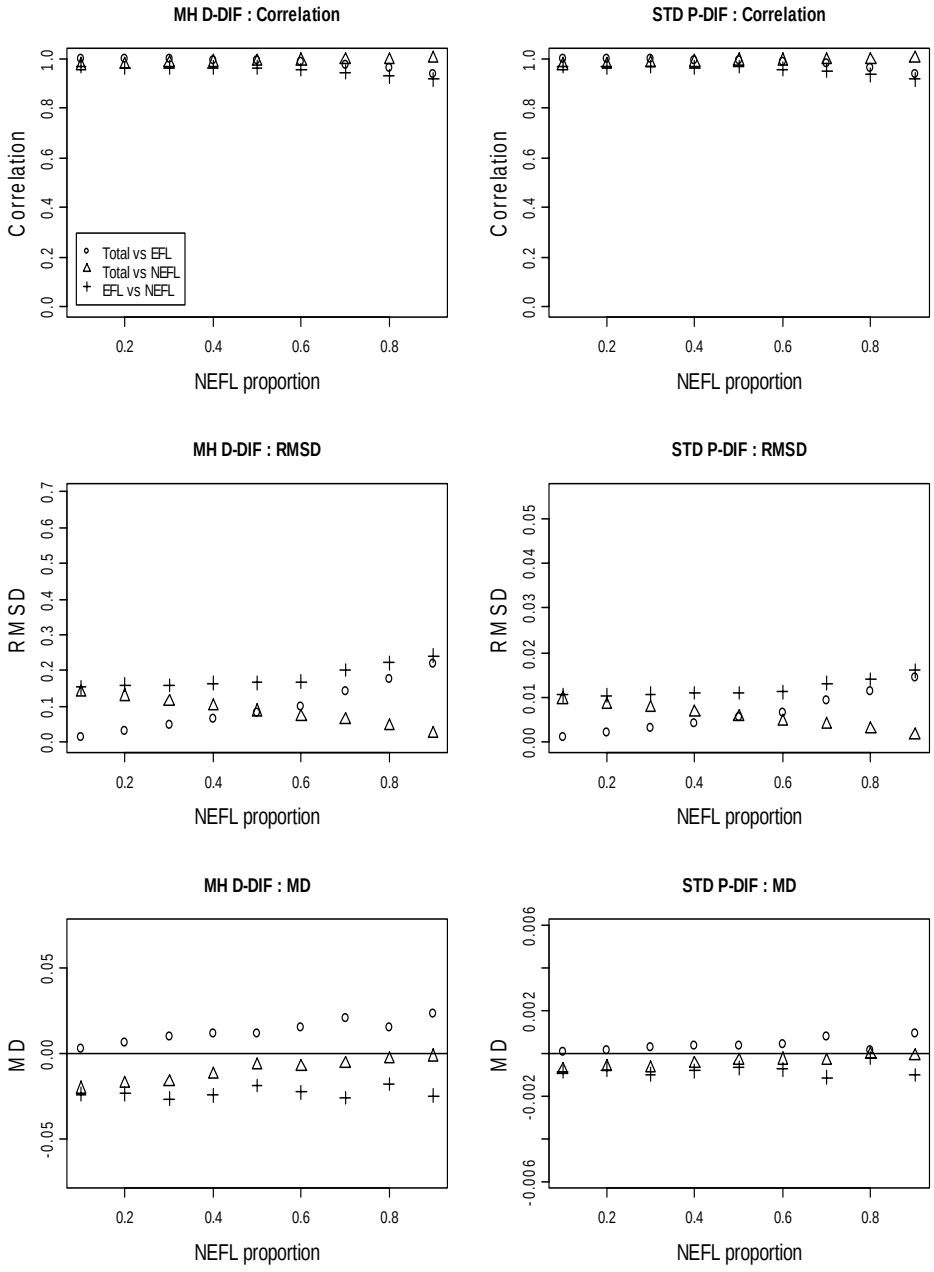

Figure A19. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

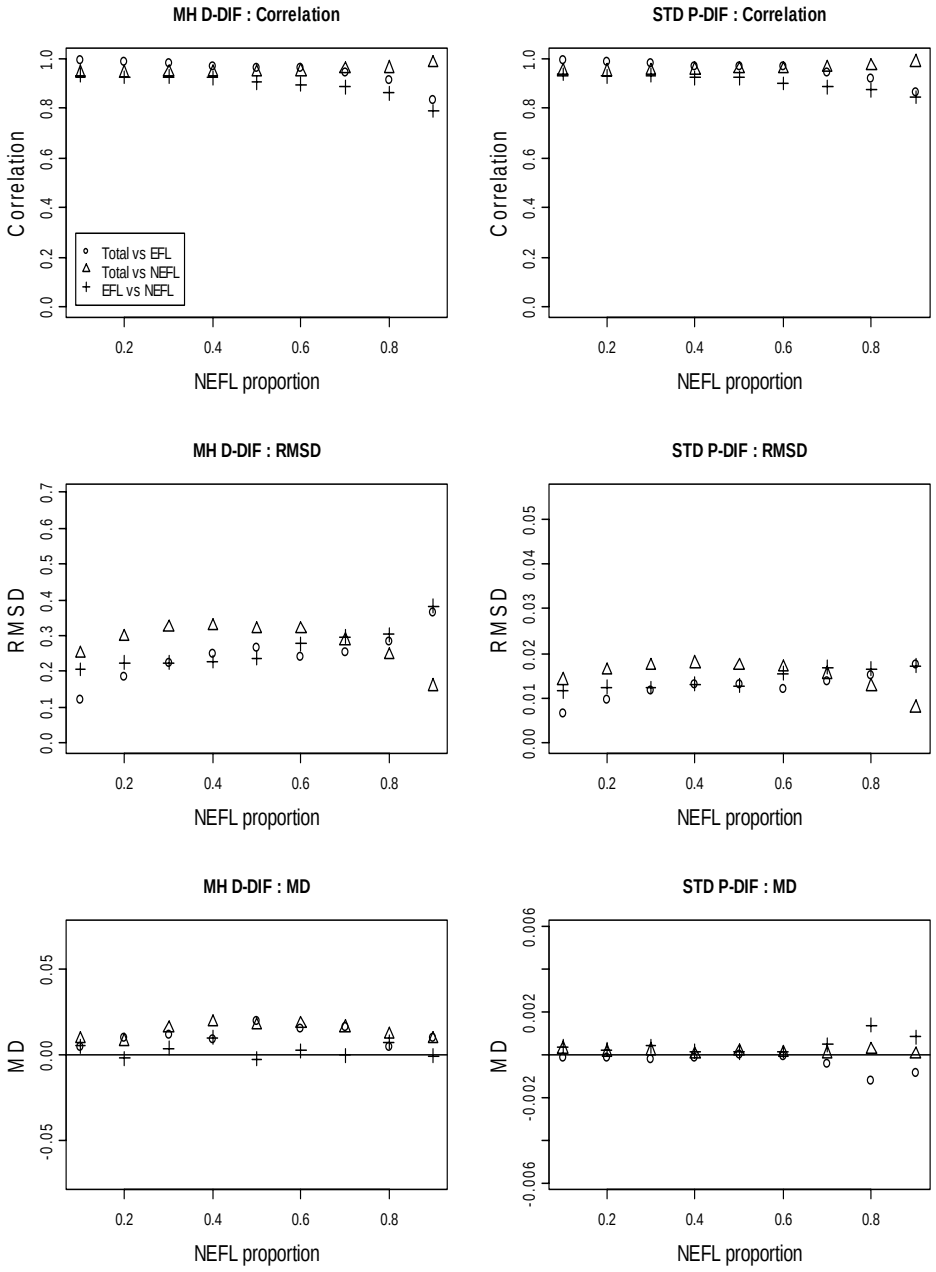

Figure A20. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

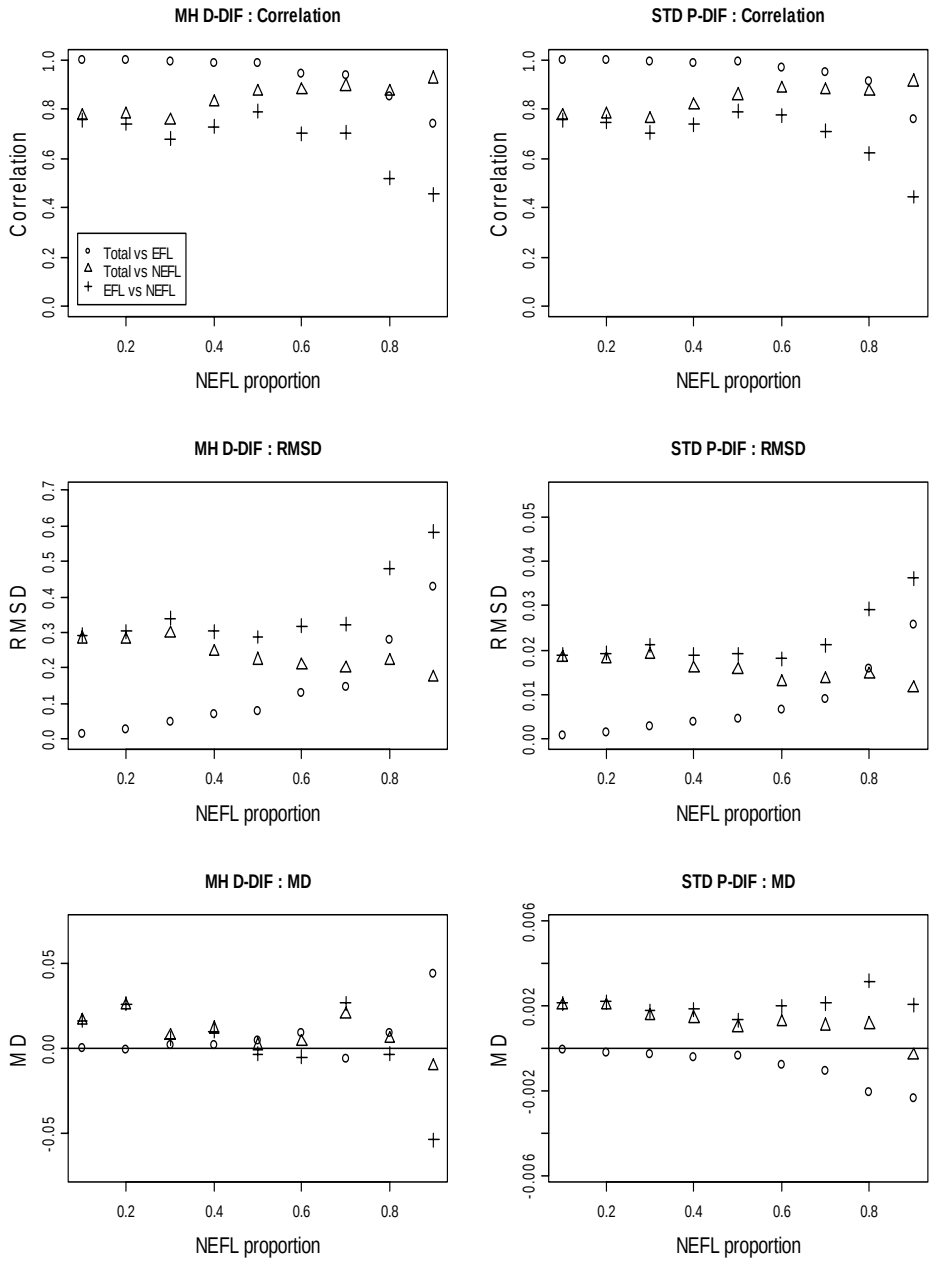

Figure A21. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

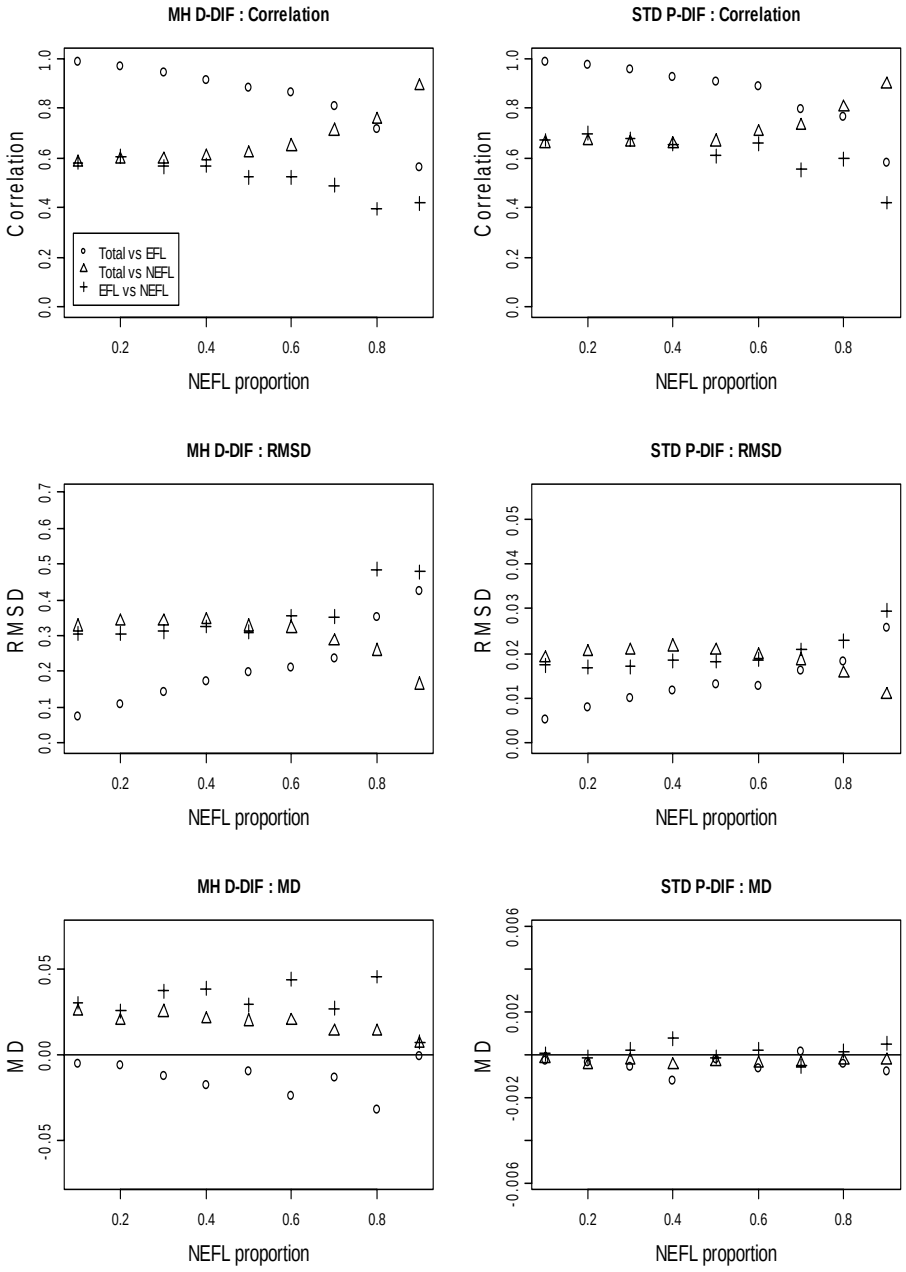

Figure A22. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday mathematics, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

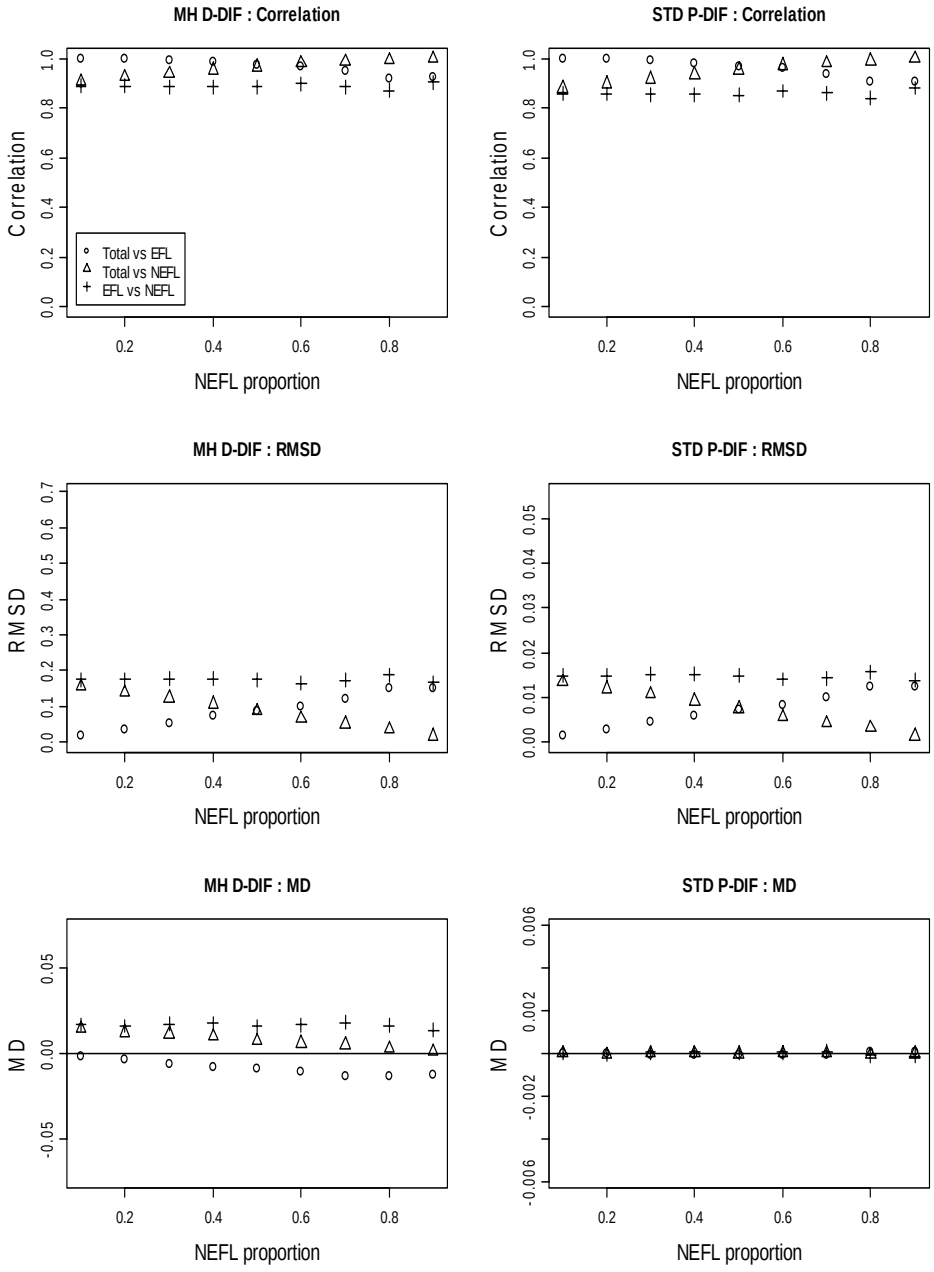

Figure A23. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday writing, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

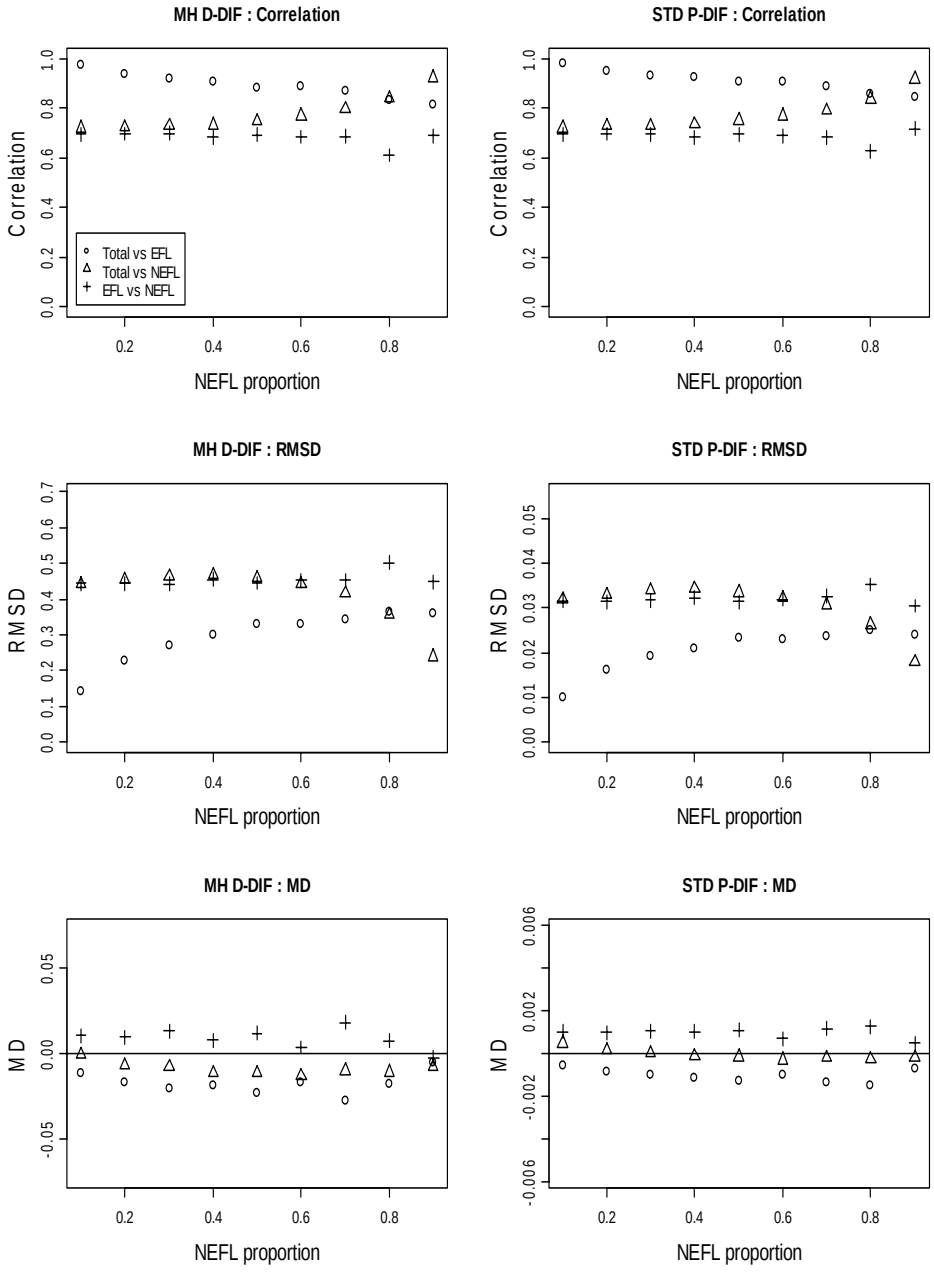

Figure A24. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday writing, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

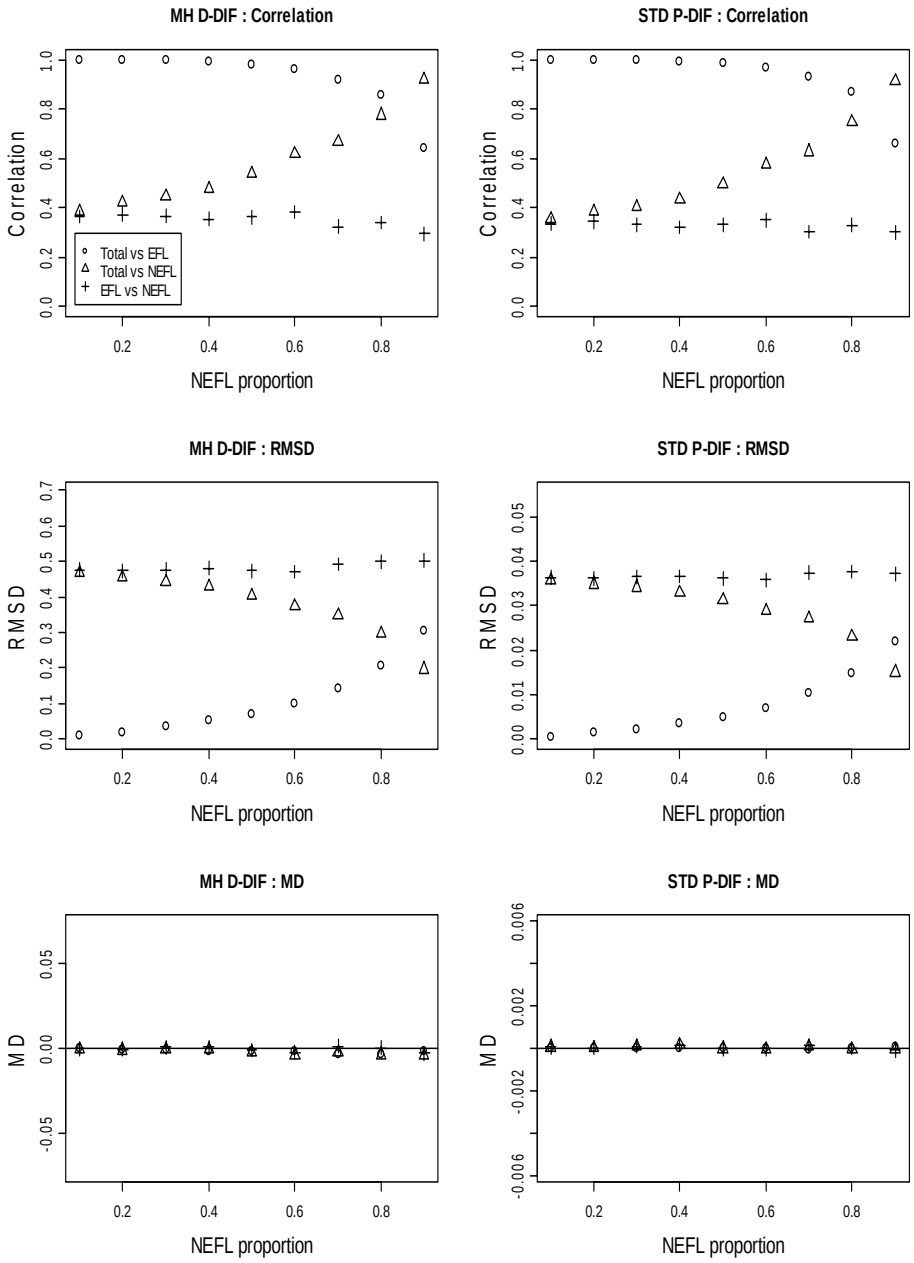

Figure A25. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday writing, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

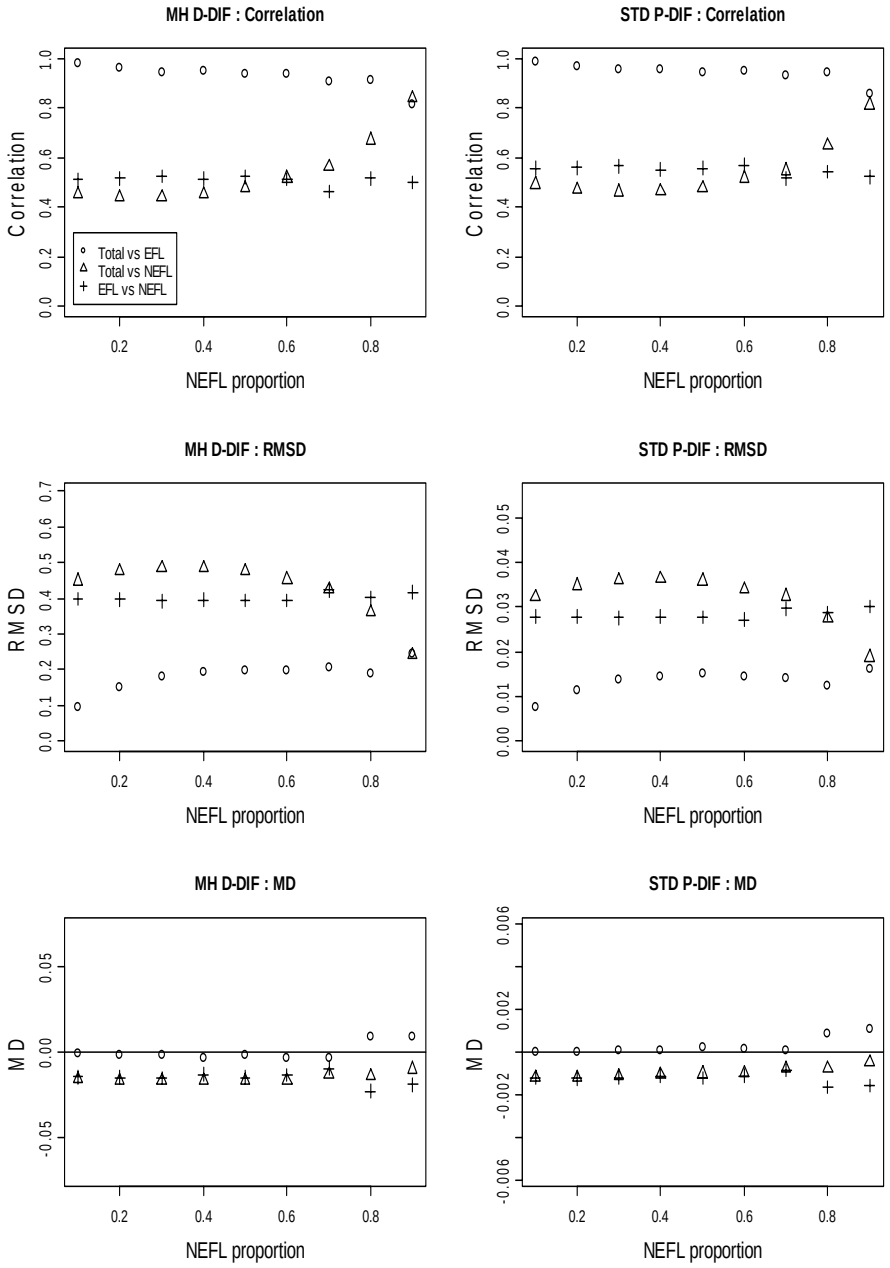

Figure A26. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Wednesday writing, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

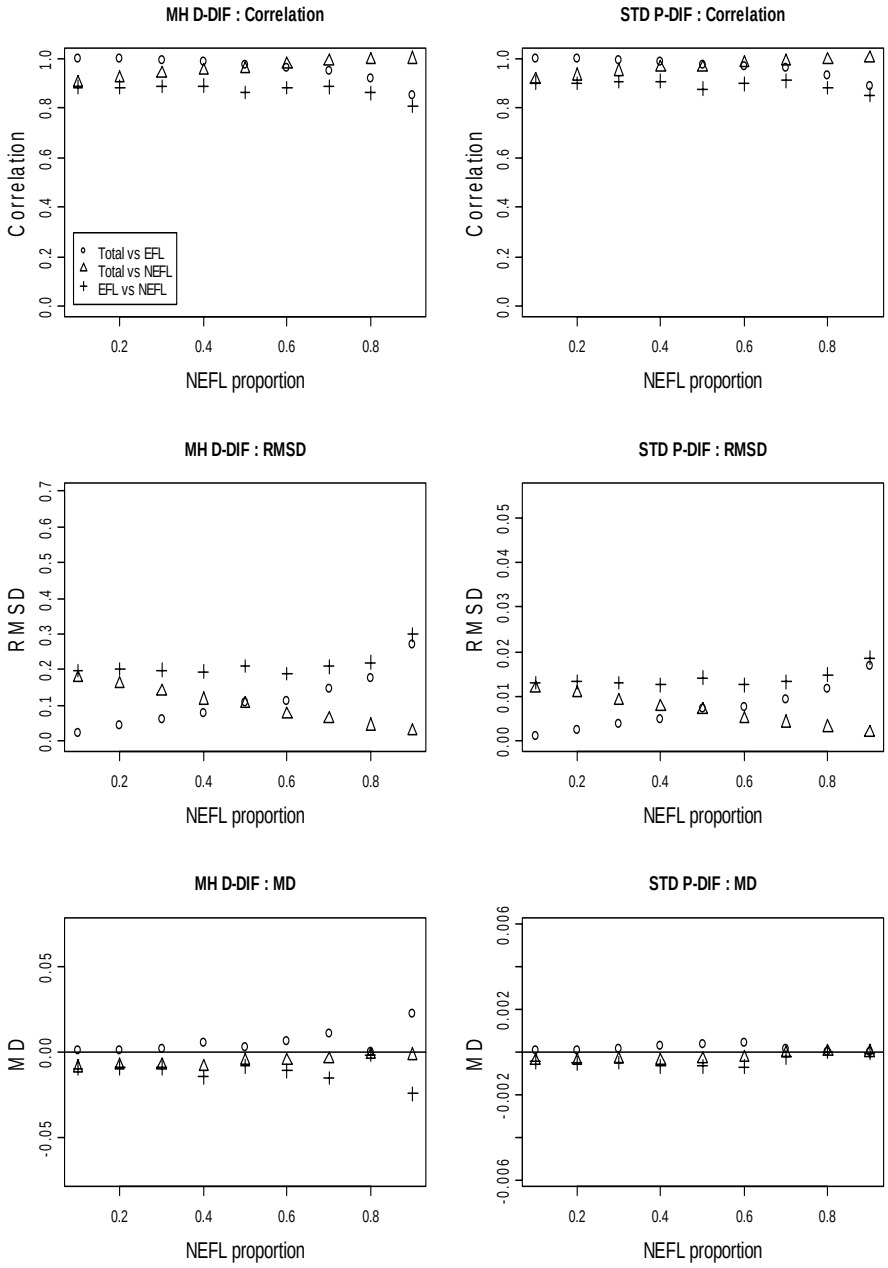

Figure A27. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, female/male.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

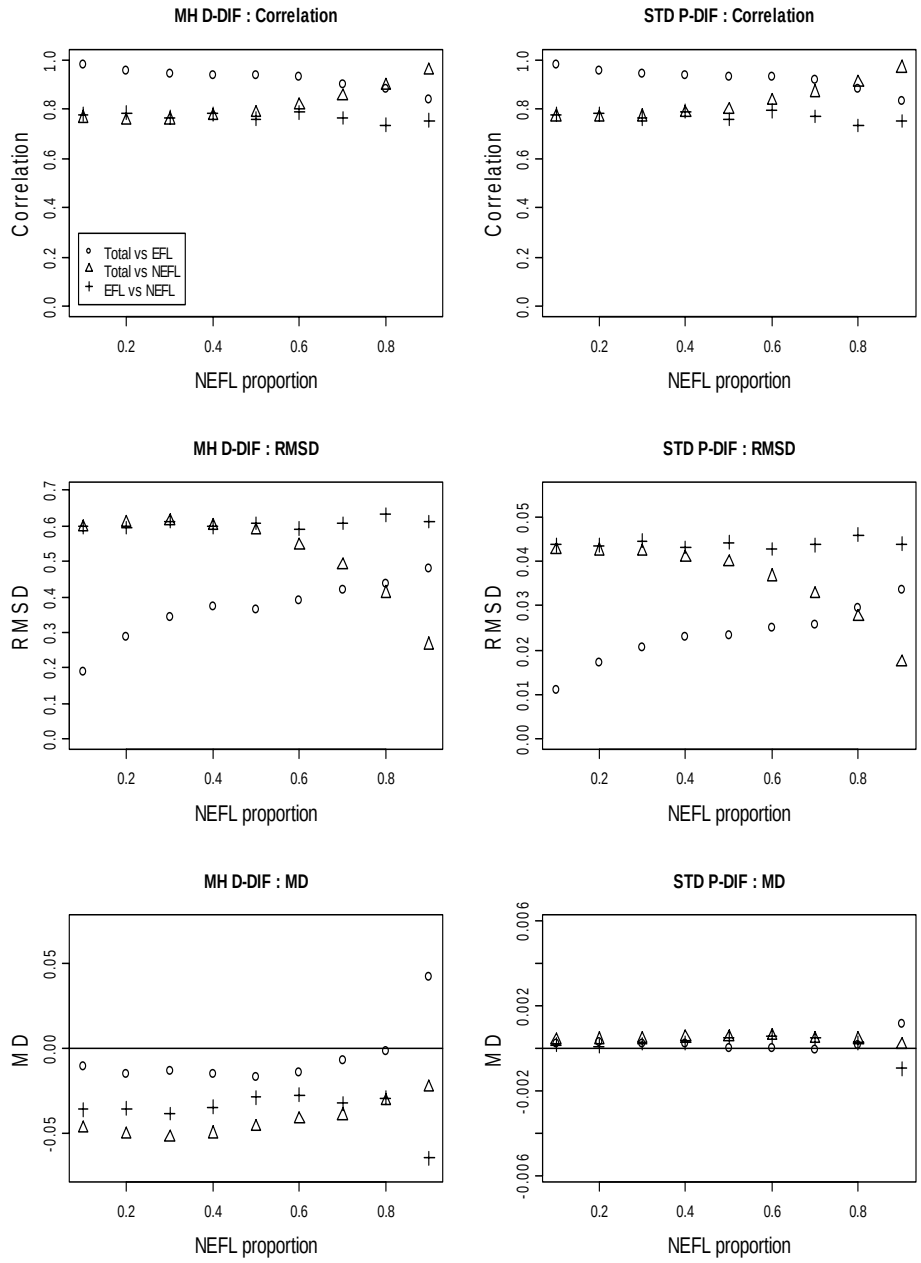

Figure A28. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, Asian American/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

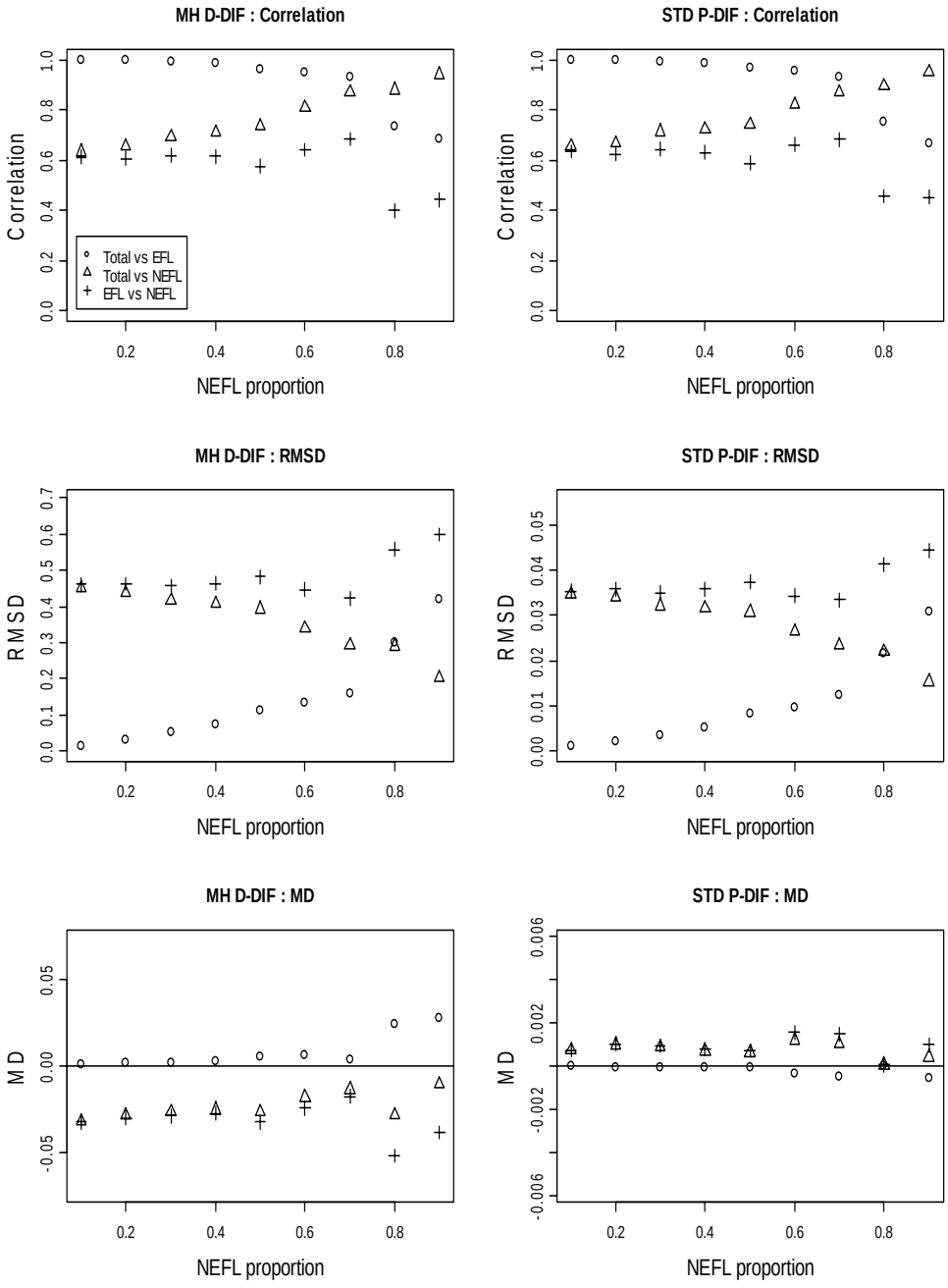

Figure A29. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, Black/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.

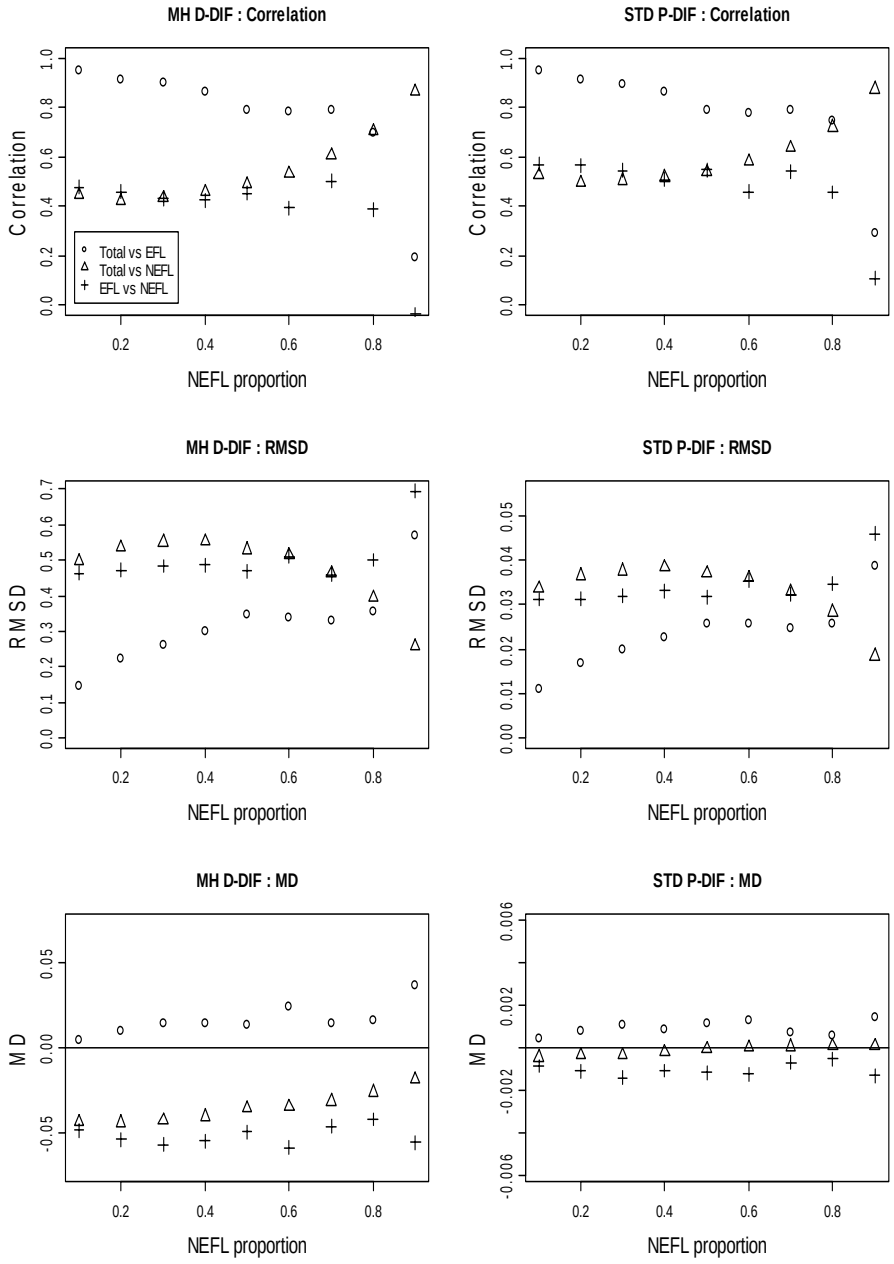

Figure A30. **Effect of change in not English first language (NEFL) proportion on differential item functioning (DIF) results for the synthetic population: Saturday writing, Hispanic/White.**

Note. MH D-DIF = Mantel-Haenszel differential difficulty, MD = mean difference, RMSD = root mean squared difference, STD P-DIF = standardized P-difference.