Listening. Learning. Leading.®

Research Report
ETS RR–11-43

How Does the Knowledge of Subgroup Membership of Examinees Affect the Prediction of True Subscores?

Shelby J. Haberman

Sandip Sinharay

November 2011

How Does the Knowledge of Subgroup Membership of Examinees Affect the Prediction of True Subscores?

Shelby J. Haberman and Sandip Sinharay, ETS, Princeton, New Jersey

November 22, 2011

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

Technical Review Editor: Matthias von Davier and James E. Carlson

Technical Reviewers: Samuel A. Livingston and Frank Rijmen

Abstract

Subscores are reported for several operational assessments. Haberman (2008) suggested a method based on classical test theory to determine if the true subscore is predicted better by the corresponding subscore or the total score. Researchers are often interested in learning how different subgroups perform on subtests. Stricker (1993) and Livingston and Rupp (2004) found that the mean difference between the subgroups was not the same for the different subscores. We suggest new methods to investigate whether the quality of prediction of the true subscore improves if the investigator knows about the subgroup membership. We applied our suggested method to data from 4 operational testing programs. We found that the quality of prediction of the true subscore does not improve if the investigator knows the subgroup membership. We also found that whether the subscores have value added does not depend on the subgroups.

Key words: augmented subscore, classical test theory, mean squared error, proportional reduction in mean squared error, reliability

Acknowledgments

The author thanks Neil J. Dorans, Skip Livingston, Frank Rijmen, and Matthias von Davier, and the anonymous reviewers for their advice. The author gratefully acknowledges the help of Ruth Greenwood with copy editing. Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

Table of Contents

Subscores are reported by several operational assessments. It is possible to determine if there is justification for reporting subscores using an approach based on classical test theory (CTT, Haberman, 2008). The approach is based on the regression of a true subscore on the corresponding observed subscore and the other observed subscores. A subscore is declared to have added value if it predicts the true subscore better than does the observed total score. Several researchers (Haberman, Sinharay, & Puhan, 2009; Lyren, 2009; Puhan, Sinharay, Haberman, & Larkin, 2010) applied the approach of Sinharay and Haberman (2008) to data from a several operational assessments. The analysis of Haberman (2008) does not take into account any subgroup information. The analysis includes all the examinees in the sample. In addition, no one has explored whether the added value of the subscores varies between subgroups.

Performances of subgroups (for example, those based on gender or ethnicity) on subtests have been studied in the context of educational testing. Stricker (1993) and Livingston and Rupp (2004) found that the performance gap between the subgroups was variable over the different subscores. For example, Livingston and Rupp found that men tend to score worse, relative to women, on constructed-response (CR) tests than on multiple-choice (MC) tests in Praxis™ Principles of Learning and Teaching assessments that are given to secondary school teachers.

Therefore, testing programs that report subscores may be interested in the following three important questions:

- Does information on subgroup membership of the examinees lead to a more accurate estimation of the true subscore? We do not intend to recommend the use of subgroup membership to predict true subscore if the answer to this question is found to be "yes." Rather, in that case, we intend to recommend a detailed investigation of the test content and the demographic composition of the examinees. The investigation might reveal why the subgroup membership leads to a better prediction of the true subscore. It is like an application of score equity assessment in which an investigator assesses if the equating function is variable over the subgroups. Variability would not lead to reporting of different equating conversions. However, it would lead to follow-up analyses.

- For a specific subgroup, does an observed subscore lead to a better prediction of the corresponding true subscore than the total score does?

- Is it possible that subscores have added value for some subgroups and no added value for some

other subgroups? If the answer to this question is "yes," then the approach of Haberman (2008) could theoretically be applied to each subgroup and subscores could theoretically be reported only for subgroups for which the subscores have added value.

Methods

In this section, we discuss methods that can be used to answer the three questions asked above.

We extended the CTT-based method of Haberman (2008) to answer the first question, whether information on subgroup membership of the examinees leads to a more accurate estimation of the true subscore. In this extension, we consider the effect on estimation of true subscores that results from the use of subgroup membership in addition to the use of the test scores.

Consider examinee groups (or subgroups) g from 1 to $n_g \geq 2$. For the g-th group, let:

- $p_g > 0$ be the fraction of the whole sample in that group,
- $\hat{\rho}^2_{sg}$ be an estimate of ρ^2_{sg}, the reliability of observed subscore s,
- $\bar{s}_g$ be the sample mean for observed subscore s,
- $\bar{x}_g$ be the sample mean for observed total score x,
- V_{sg} be the estimate of the variance of s,
- V_{xg} be the estimate of the variance of the observed total score x,
- $\mathrm{Cov}_g(s, x)$ be the estimate of the covariance of the observed subscore s and x, and
- $\mathrm{Cov}_g(s_t, x)$ be the estimate of the covariance of the true subscore s_t and x.

For the g-th group, it is possible to consider the following estimates of s_t:

- $\bar{s}_g$.
- $s_{sg} = \bar{s}_g + \hat{\rho}^2_{sg}(s - \bar{s}_g)$, which is based on s.
- $s_{xg} = \bar{s}_g + c_g(x - \bar{x}_g)$, which is based on x, where $c_g = \mathrm{Cov}_g(s_t, x)/V_{xg}$.
- $s_{sxg} = \bar{s}_g + a_g(s - \bar{s}_g) + b_g(x - \bar{x}_g)$, which is a weighted average of s and x.

For the g-th group, the group-specific estimate $\bar{s}_g$ leads to the group-specific mean squared error MSE_g, which is an estimate of

$$E\left([s_t - \mu_{sg}]^2 \,|G = g\right) = \text{Variance of } s_t \text{ for examinees in group } g,$$

where $\mu_{sg} = E(s|G = g)$. In addition, the group-specific estimates s_{sg}, s_{xg}, and s_{sxg} lead to the respective group-specific mean squared errors MSE_{sg}, MSE_{xg}, and MSE_{sxg}. The quantities $\bar{s}_g$, s_{sg}, s_{xg}, and s_{sxg} have the respective proportional reductions in mean squared error

$$\text{PRMSE}_{sg} = \hat{\rho}^2_{sg},$$

PRMSE_{xg}, and PRMSE_{sxg}. A comparison of PRMSE_{sg}, PRMSE_{xg}, and PRMSE_{sxg} with the overall quantities PRMSE_s, PRMSE_x, and PRMSE_{sx} may reveal interesting information.[1] Let RMSE_g denote the square root of MSE_g, so that RMSE_g is the estimate of the standard deviation of s_t for the g-th group.

The quantities PRMSE_{sg}, PRMSE_{xg}, and PRMSE_{sxg} are based on subgroup information in addition to the scores from the other parts of the assessment. To investigate if inclusion of subgroup membership results in a better prediction of s_t, one has to compare them to PRMSEs of estimates of s_t that do not use subgroup membership. Comparison of the PRMSEs computed from the full sample, PRMSE_s, PRMSE_x, and PRMSE_{sx}, to PRMSE_{sg}, PRMSE_{xg}, and PRMSE_{sxg} does not tell us whether subgroup membership leads to a better prediction of s_t.

For examinees of the g-th group, it is possible to prove that

$$\text{MSE}_{sg*} = \text{MSE}_{sg} + (\hat{\rho}^2_{sg} - \hat{\rho}^2_s)^2 V_{sg} + B^2_{sg}, \tag{1}$$

is the estimated MSE of

$$s_s = \bar{s} + \hat{\rho}^2_s(s - \bar{s})\cdot$$

The bias of s_s for Group g is

$$B_{sg} = -(1 - \hat{\rho}^2_s)(\bar{s}_g - \bar{s}), \tag{2}$$

and its normalized version is

$$\beta_{sg} = B_{sg}/\,\text{RMSE}_g\,. \tag{3}$$

The quantity β_{sg} is small if either the group means for the subscore are close for different groups or if the subscore is highly reliable. The corresponding PRMSE from use of s_s instead of $\bar{s}_g$ is

$$\text{PRMSE}_{sg*} = \text{PRMSE}_{sg} - \frac{(\hat{\rho}_{sg}^2 - \hat{\rho}_s^2)^2}{\hat{\rho}_{sg}^2} - \beta_{sg}^2. \tag{4}$$

For the g-th group, the estimated MSE of

$$s_x = \bar{s} + c(x - \bar{x}),$$

which is based on x, where $c = \text{Cov}(s_t, x)/V_x$, is

$$\text{MSE}_{xg*} = \text{MSE}_{xg} + (c_g - c)^2 V_{xg} + B_{xg}^2.$$

The bias of s_x for the g-th group is

$$B_{xg} = c(\bar{x}_g - \bar{x}) - (\bar{s}_g - \bar{s}) \tag{5}$$

and the corresponding normalized value is

$$\beta_{xg} = B_{xg} / \text{RMSE}_g\,. \tag{6}$$

The quantity β_{xg} will be small if, for example, examinee groups with high means for the total score have high subscore means. The PRMSE from use of s_x instead of $\bar{s_g}$ is

$$\text{PRMSE}_{xg*} = \text{PRMSE}_{xg} - \frac{(c_g - c)^2 V_{xg}}{\text{MSE}_g} - \beta_{xg}^2. \tag{7}$$

For the g-th group and

$$s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x}),$$

where a and b are constants, the estimated mean square error is

$$\text{MSE}_{sxg*} = \text{MSE}_{sxg} + (a_g - a)^2 V_{sg} + 2(a_g - a)(b_g - b)\,\text{Cov}_g(s, x) + (b_g - b)^2 V_{xg} + B_{sxg}^2.$$

The bias of s_x for the g-th group is

$$B_{sxg} = a(\bar{s}_g - \bar{s}) + b(\bar{x}_g - \bar{x}) - (\bar{s}_g - \bar{s}) \tag{8}$$

and the corresponding normalized value is

$$\beta_{sxg} = B_{sxg} / \text{RMSE}_g\,. \tag{9}$$

The bias will be small if groups with high means of total scores also have high subscore means. The PRMSE from use of s_{sx} instead of $\bar{s_g}$ is

$$\mathrm{PRMSE}_{sxg*} = \mathrm{PRMSE}_{sxg} - \frac{(a_g - a)^2 V_{sg} + 2(a_g - a)(b_g - b)\,\mathrm{Cov}_g(s, x) + (b_g - b)^2 V_{xg}}{\mathrm{MSE}_g} - \beta_{sxg}^2. \quad (10)$$

To determine if use of subgroup information leads to a better prediction of s_t, it is possible to compare PRMSE_{sg*}, PRMSE_{xg*}, and PRMSE_{sxg*} to PRMSE_{sg}, PRMSE_{xg}, and PRMSE_{sxg}. For example, if PRMSE_{sg*} is much smaller than PRMSE_{sg}, that would imply that the use of the subgroup information leads to a better estimation of the true subscore.

One can also look at the magnitudes of the quantities β_{sg}, β_{xg}, and β_{sxg}. Some of the differences between PRMSE_{sg*} and PRMSE_{sg}, or between PRMSE_{xg*} and PRMSE_{xg} involve these quantities. Among these three, β_{sg} would most often be larger than β_{sxg} and β_{xg}. The subscore means of the subgroups need to be close to each other, or the subscore reliability needs to be high, in order for β_{sg} to be small. In contrast, β_{xg} and β_{sxg} would be small if the subscore profiles of the subgroups are parallel. This would most often happen for one of the many nearly one-dimensional assessments.

Overall measures of the influence of subgroups on estimation of the true subscores are provided by the respective overall MSEs

$$\mathrm{MSE}_{s+} = \sum_{g=1}^{n_g} p_g\,\mathrm{MSE}_{sg}, \quad (11)$$

$$\mathrm{MSE}_{x+} = \sum_{g=1}^{n_g} p_g\,\mathrm{MSE}_{xg},$$

and

$$\mathrm{MSE}_{sx+} = \sum_{g=1}^{n_g} p_g\,\mathrm{MSE}_{sxg}\,.$$

The resulting PRMSEs are

$$\mathrm{PRMSE}_{s+} = 1 - \mathrm{MSE}_{s+} \,/\, \mathrm{MSE}, \quad (12)$$

$$\mathrm{PRMSE}_{x+} = 1 - \mathrm{MSE}_{x+} \,/\, \mathrm{MSE},$$

and

$$\mathrm{PRMSE}_{sx+} = 1 - \mathrm{MSE}_{sx+} \,/\, \mathrm{MSE}\,.$$

If there are substantial discrepancies between the group-specific estimates and corresponding overall estimates, that would require a follow-up analysis. For example, if PRMSE_{s+} is substantially different from PRMSE_{s2}, that would imply that there is something special about the performance of Subgroup 2 on the subtest. Further, big discrepancies between, for example, PRMSE_s and PRMSE_{s+} might indicate that the subgroup information leads to better prediction of s_t and would require a follow-up analyses.

One can obtain a measure similar to PRMSE_{sg*} for the total score. Let $\bar{x}_g$ denote the average total score for the g-th group and x_x denote the regression of the true total score on x for all examinees. The proportional reduction PRMSE^x_{xg*} obtained for examinees of g-th group in predicting the true total score by x_x instead of by $\bar{x}_g$ can be expressed as

$$\text{PRMSE}^x_{xg*} = \text{PRMSE}^x_{xg} - \frac{(\hat{\rho}^2_{xg} - \hat{\rho}^2_x)^2}{\hat{\rho}^2_{xg}} - (\beta^x_{xg})^2, \tag{13}$$

where the total test reliability for Group g is PRMSE^x_{xg}, $\beta^x_{xg} = -(1-\hat{\rho}^2_x)(\bar{x}_g - \bar{x})/\sqrt{\text{MSE}^x_g}$, and

$$\text{MSE}^x_g = \text{Variance of the true total score computed only using examinees in group } g\cdot$$

Whether the group membership of the examinees helps in prediction of the true total score of the examinees can be assessed from a comparison of PRMSE^x_{xg} and PRMSE^x_{xg*}.

Similarly, it is possible to obtain expressions for the total score that are analogous to the overall expressions MSE_{s+} and PRMSE_{s+}. For example, the expression that is similar to PRMSE_{s+} is

$$\text{PRMSE}^x_{x+} = 1 - \text{MSE}^x_{x+} \,/\, \text{MSE}^x,$$

where the following hold:

$$\begin{aligned}
\text{MSE}^x_{x+} &= \sum_{g=1}^{n_g} p_g \,\text{MSE}^x_{xg}, \\
\text{MSE}^x_{xg} &= \text{True total score variance in Group } g \times (1 \text{ - Total score reliability in Group } g), \\
\text{MSE}^x &= \text{Variance of the true total score computed using the full sample.}
\end{aligned}$$

Whether the group membership of the examinees helps in prediction of the true total score of the examinees can be assessed from a comparison of PRMSE^x_{x+} with the total test reliability for the full sample.

It is possible to use the proportional reductions in mean square error PRMSE_{sg} and PRMSE_{xg} to answer the last two questions asked on Page 1. If, for example, PRMSE_{xg} is

smaller than PRMSE_{sg} for the g-th group, then the observed subscore leads to better prediction, compared to the observed total score, of s_t for that group. If, for example, PRMSE_{xg} is smaller than PRMSE_{sg} for some groups but is larger for some other groups, then the added value of the subscore differs over the groups. If PRMSE_x is found to be larger than PRMSE_s for all the subscores of an assessment, but PRMSE_{xg} is found to be smaller than PRMSE_{sg} for most of the subscores for one or two examinee groups, then subscores could theoretically be reported for only those subgroups for which the subscores have added value. This will, in theory, allow one to report subscores for some groups for an assessment for which subscores do not have added value for the full sample.

Data

Data were obtained from four assessments. Subscores are operationally reported for all these assessments. The assessments are described below.

Assessment A

We had data from four recent forms of an English proficiency assessment that has four parts. Parts 1 and 2 consist of mostly dichotomous items with scores 0 and 1. In typical cases, all 34 items in Part 1 are dichotomous and Part 2 contains three trichotomous items with possible scores of 1, 2, and 3 and 42 dichotomous items. Parts 3 and 4 consist of only constructed response items—six for Part 3 and two for Part 4. The sample sizes for the four forms were between 8,500 and 14,500. We show results for seven subgroups based on the self-reported first language of the examinees. One of the subgroups included the small language-based subgroups and those who did not provide their first language.

Assessment B

Assessment B is a battery of assessments and measures school and individual student progress. We considered two assessments from Assessment B, Assessments B1 and B2. Here, we report results from one recent form each of Assessments B1 and B2. The resulting sample sizes were 6,563 for Assessment B1, and 7,362 for Assessment B2. In the 50-item (both MC and constructed-response [CR] items) Assessment B1, there are four subscores. In the 55-item Assessment B2 (both MC and CR items), there are six subscores. These subscores are not

reported on the individual student score reports, but they are included in the student data files sent to the schools. We show results for two subgroups based on gender and five subgroups based on ethnicity. One of the ethnicity-based subgroups for each assessment included the small ethnicity-based subgroups (those with fewer than 100 examinees) and those who did not provide their ethnicity.

Assessment C

Assessment C is a battery of assessments that measures achievement in several fields. We considered two titles under Assessment C—we denote them here as Assessments C1 and C2. We analyzed data from two forms of each of these two assessments. In Assessment C1, which has roughly 205 multiple-choice (MC) items, two subscores are reported. Some questions (about 17%) contribute to the total reported score on Assessment C1 but are not part of a reported subscore. We assume that these items contribute to a third subscore for the assessment. Assessment C2 has roughly 200 MC items contributing to three subscores. A three-subscore analysis was performed for both these assessments. We show the results for two subgroups based on gender and five subgroups based on ethnicity for Assessment C1. We show the results for two subgroups based on gender and three subgroups based on ethnicity for Assessment C2. The last ethnicity-based subgroup for each form included the small ethnicity-based subgroups (those with fewer than 100 examinees) and those who did not indicate their ethnicity.

Assessment D

Assessment D is a battery of teacher-certification assessments. We considered two titles from Assessment D. These are denoted as Assessments D1 and D2. Assessment D1 is administered to potential teachers in schooles. It includes 120 MC items contributing equally to four subscores. Assessment D2 is administered to prospective and practicing paraprofessionals. The 75 MC items are distributed equally among three subtests. We analyzed data from one form each of Assessments D1 and D2. For both assessments, we show results for two subgroups based on gender and four subgroups based on ethnicity.

Results

Assessment A

Figures 1 and 2 show the standardized section score means of the language-based groups for the Forms 1 and 3 of Assessment A. To compute the standardized subscore mean for a subgroup, we first standardized each subscore by dividing the difference between the subscore and its mean by its standard deviation and then computed the mean of these standardized values.[2] The section score profiles of the subgroups are not parallel, primarily due to a difference of the third section score and the other section scores.

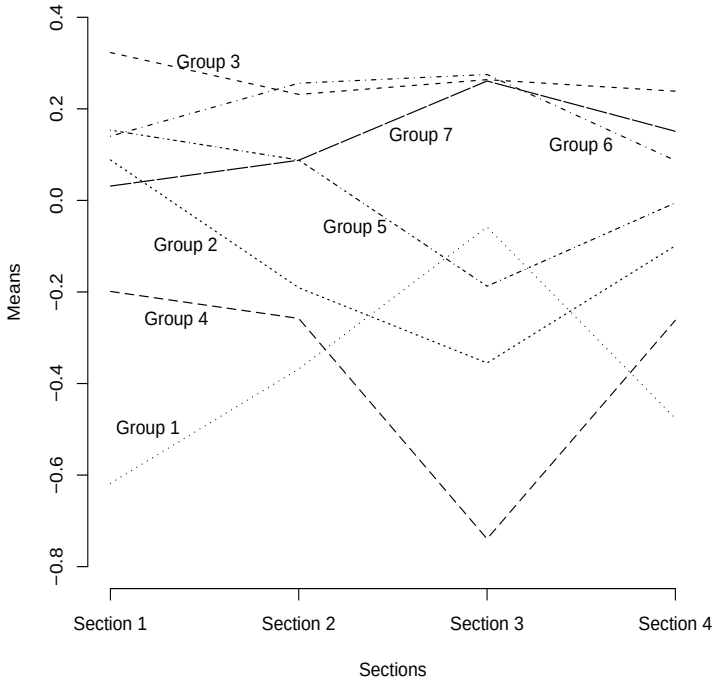

Figure 1. **The standardized subgroup means for the first form of Assessment A.**

Table 1 shows the results for Forms 1 and 3 of Assessment A (the results for the other two forms are similar). In the correlation matrices, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal. The results for all examinees are shown first, followed by the results for the gender-based subgroups and the ethnicity-based subgroups. For each subgroup, the values of PRMSE_{sg}, PRMSE_{xg}, PRMSE_{sxg}, β_{sg}, β_{xg}, and β_{sxg}, PRMSE_{sg*}, PRMSE_{xg*}, and PRMSE_{sxg*} are shown in Table 1.

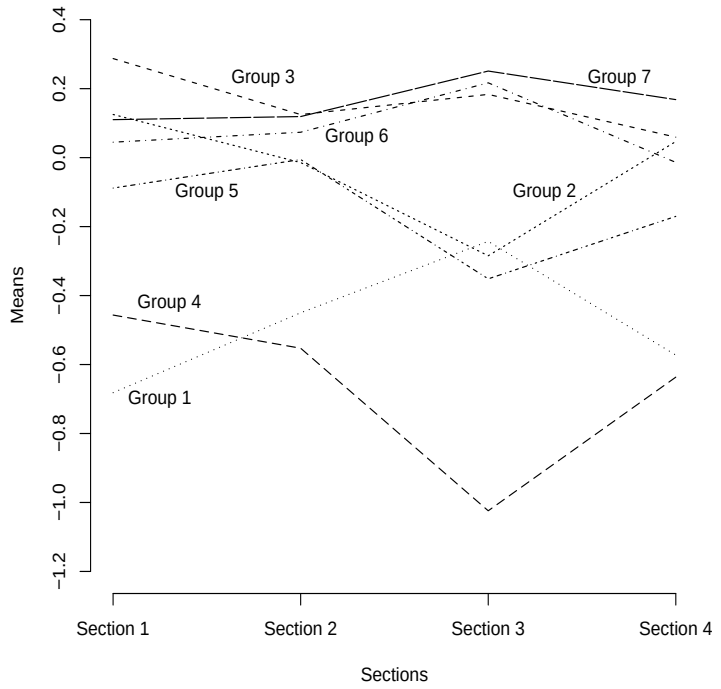

Figure 2. **The standardized subgroup means for the third form of Assessment A.**

The sizes of the subgroups are also shown. If the PRMSE of a subscore (or its reliability) is larger than the corresponding PRMSE of the total score, the former is written in bold font. For both of these forms, the covariance matrix of the true section scores was not positive definite for a few language groups (such as Language Group 3); for these groups, the PRMSE_x and PRMSE_{sx} were set equal to the reliability of the total score.

There is little variation between the subgroups, either with respect to PRMSEs or added value of the section scores. The third section has added value for the full sample and for all the language subgroups. The other section scores do not have added value for any group except for Language Group 4 for Form 3, for which the first section score has added value.

For Assessment A, the differences between PRMSE_{sg} and PRMSE_{sg*} are always close to zero. The same is true for the differences between PRMSE_{sxg} and PRMSE_{sxg*}. The departures from parallelism of profiles in Figures 1 and 2 do not have much impact on the PRMSE of the augmented subscores. The differences between PRMSE_{xg} and PRMSE_{xg*} are substantial for the third section score for some language groups, such as Language Group 2; note that PRMSE_{xg} is

Table 1

Proportional Reduction in Mean-Squared Errors (PRMSEs) for

Forms 1 and 3 of Assessment A

	Form 1 sections				Form 3 sections			
	1	2	3	4	1	2	3	4
Maximum score	42	34	24	10	42	34	24	10
All examinees (sizes: 14,496, 14,186)								
Correlation	1.00	0.76	0.54	0.74	1.00	0.78	0.60	0.76
	0.92	1.00	0.66	0.76	**0.91**	1.00	0.69	0.78
	0.63	**0.77**	1.00	0.69	**0.68**	**0.79**	1.00	0.72
	0.94	**0.97**	**0.85**	1.00	**0.93**	**0.96**	**0.87**	1.00
PRMSE_{s}	0.84	0.84	**0.88**	0.74	0.87	0.85	**0.88**	0.77
PRMSE_{x}	0.86	0.90	0.62	0.94	0.87	0.90	0.67	0.92
PRMSE_{sx}	0.88	0.91	0.89	0.94	0.90	0.91	**0.90**	0.93
Language Group 1 (sizes: 1,363, 1,207)								
PRMSE_{sg}	0.81	0.81	**0.87**	0.74	0.84	0.84	**0.89**	0.79
PRMSE_{xg}	0.87	0.90	0.64	0.88	0.85	0.89	0.67	0.88
PRMSE_{sxg}	0.88	0.91	0.89	0.90	0.88	0.91	0.90	0.91
β_{sg}	0.12	0.07	0.01	0.15	0.10	0.07	0.03	0.14
β_{xg}	0.04	-0.01	-0.10	0.06	0.03	-0.01	-0.05	0.05
β_{sxg}	0.18	-0.03	-0.07	0.09	0.13	-0.03	-0.04	0.08
PRMSE_{sg*}	0.80	0.81	**0.87**	0.72	0.83	0.84	**0.89**	0.77
PRMSE_{xg*}	0.87	0.90	0.63	0.88	0.85	0.89	0.66	0.88
PRMSE_{sxg*}	0.85	0.90	0.88	0.88	0.86	0.90	0.90	0.90
Language Group 2 (sizes: 1,927, 2,197)								
PRMSE_{sg}	0.85	0.83	**0.84**	0.73	0.85	0.82	**0.81**	0.71
PRMSE_{xg}	0.87	0.89	0.63	0.94	0.86	0.89	0.62	0.90
PRMSE_{sxg}	0.89	0.90	0.87	0.94	0.89	0.90	0.84	0.90
β_{sg}	-0.02	0.03	0.06	0.03	-0.02	0.00	0.05	-0.01
β_{xg}	-0.03	0.02	0.13	0.01	-0.02	0.00	0.17	-0.04
β_{sxg}	-0.12	0.08	0.09	0.01	-0.09	0.01	0.11	-0.05
PRMSE_{sg*}	0.85	0.83	**0.84**	0.72	0.85	0.82	**0.80**	0.71
PRMSE_{xg*}	0.87	0.89	0.59	0.93	0.86	0.89	0.56	0.90
PRMSE_{sxg*}	0.88	0.90	0.85	0.94	0.88	0.90	0.82	0.90
Language Group 3 (sizes: 440, 518)								
PRMSE_{sg}	0.83	0.85	**0.84**	0.66	0.88	0.88	**0.88**	0.73
PRMSE_{xg}	0.87	0.89	0.61	0.93	0.91	0.94	0.73	0.92
PRMSE_{sxg}	0.88	0.91	0.86	0.93	0.92	0.94	0.90	0.92
β_{sg}	-0.06	-0.04	-0.04	-0.08	-0.04	-0.02	-0.02	-0.02
β_{xg}	-0.01	0.01	-0.02	0.02	-0.01	0.01	-0.01	0.08
β_{sxg}	-0.05	0.03	-0.02	0.02	-0.06	0.04	-0.01	0.12
PRMSE_{sg*}	0.82	0.85	**0.83**	0.64	0.88	0.88	**0.88**	0.73
PRMSE_{xg*}	0.87	0.89	0.59	0.93	0.91	0.94	0.73	0.92
PRMSE_{sxg*}	0.88	0.90	0.86	0.93	0.92	0.94	0.90	0.92

	Form 1 sections				Form 3 sections			
	1	2	3	4	1	2	3	4
Language Group 4 (sizes: 1,020, 792)								
PRMSE_{sg}	0.83	0.83	**0.89**	0.72	**0.82**	0.84	**0.90**	0.76
PRMSE_{xg}	0.86	0.90	0.66	0.93	0.81	0.89	0.70	0.90
PRMSE_{sxg}	0.88	0.91	0.90	0.93	0.86	0.90	0.91	0.92
β_{sg}	0.04	0.05	0.09	0.08	0.07	0.09	0.12	0.18
β_{xg}	-0.02	-0.01	0.14	-0.03	-0.02	-0.01	0.14	0.05
β_{sxg}	-0.06	-0.04	0.13	-0.04	-0.07	-0.02	0.14	0.07
PRMSE_{sg*}	0.83	0.82	**0.88**	0.71	0.81	0.83	**0.88**	0.72
PRMSE_{xg*}	0.86	0.90	0.64	0.93	0.80	0.89	0.67	0.90
PRMSE_{sxg*}	0.87	0.91	0.89	0.93	0.85	0.90	**0.89**	0.91
Language Group 5 (sizes: 2,577, 1,194)								
PRMSE_{s}	0.82	0.84	**0.89**	0.75	0.84	0.85	**0.89**	0.73
PRMSE_{x}	0.84	0.90	0.69	0.92	0.85	0.89	0.67	0.91
PRMSE_{sxg}	0.87	0.91	0.90	0.93	0.88	0.91	0.90	0.91
β_{sg}	-0.03	-0.02	0.02	0.00	0.01	0.00	0.04	0.05
β_{xg}	-0.02	-0.01	0.07	0.04	0.00	-0.02	0.08	0.05
β_{sxg}	-0.08	-0.03	0.06	0.06	-0.01	-0.09	0.07	0.06
PRMSE_{sg*}	0.82	0.84	**0.89**	0.75	0.83	0.85	**0.88**	0.73
PRMSE_{xg*}	0.83	0.89	0.67	0.92	0.84	0.89	0.66	0.91
PRMSE_{sxg*}	0.86	0.91	0.90	0.92	0.88	0.90	0.89	0.91
Language Group 6 (sizes: 1,032, 699)								
PRMSE_{sg}	0.81	0.81	**0.83**	0.68	0.87	0.87	**0.84**	0.74
PRMSE_{xg}	0.85	0.86	0.62	0.92	0.87	0.89	0.66	0.91
PRMSE_{sxg}	0.87	0.88	0.85	0.92	0.90	0.91	0.87	0.91
β_{sg}	-0.03	-0.05	-0.04	-0.03	-0.01	-0.01	-0.03	0.00
β_{xg}	0.01	-0.01	-0.05	0.08	0.00	0.00	-0.07	0.06
β_{sxg}	0.03	-0.06	-0.04	0.11	0.02	0.00	-0.05	0.09
PRMSE_{sg*}	0.80	0.81	**0.82**	0.68	0.87	0.86	**0.84**	0.74
PRMSE_{xg*}	0.85	0.86	0.61	0.92	0.87	0.89	0.64	0.91
PRMSE_{sxg*}	0.87	0.88	0.85	0.92	0.90	0.91	0.86	0.91
Language Group 7 (sizes: 6,136, 699)								
PRMSE_{sg}	0.84	0.83	**0.86**	0.73	0.87	0.85	**0.87**	0.76
PRMSE_{xg}	0.88	0.90	0.64	0.94	0.88	0.90	0.68	0.91
PRMSE_{sxg}	0.89	0.91	0.88	0.94	0.91	0.91	0.89	0.92
β_{sg}	-0.01	-0.02	-0.03	-0.05	-0.02	-0.02	-0.03	-0.05
β_{xg}	0.01	0.00	-0.06	-0.04	0.00	0.00	-0.04	-0.02
β_{sxg}	0.04	0.01	-0.05	-0.06	0.01	0.01	-0.04	0.03
PRMSE_{sg*}	0.84	0.83	**0.86**	0.73	0.87	0.85	**0.86**	0.76
PRMSE_{xg*}	0.88	0.90	0.64	0.94	0.88	0.90	0.68	0.91
PRMSE_{sxg*}	0.89	0.91	0.88	0.94	0.91	0.91	0.88	0.92
PRMSE_{s+}	0.84	0.84	**0.88**	0.74	0.87	0.86	**0.89**	0.78
PRMSE_{x+}	0.87	0.90	0.69	0.94	0.88	0.90	0.72	0.93
PRMSE_{sx+}	0.89	0.91	0.90	0.94	0.91	0.92	0.90	0.93

much less than PRMSE_{sg} for the third section score for all subgroups.

The differences are small between PRMSE_{s} and PRMSE_{s+} and between PRMSE_{sx} and PRMSE_{sx+}. The differences between PRMSE_{x} and PRMSE_{x+} are substantial for the third section score.

The values of PRMSE^{x}_{xg} and PRMSE^{x}_{xg*} are the same (and roughly equal to the total score reliability for the full sample) for all the subgroups for all the forms of Assessment A. The values of PRMSE^{x}_{x+} are the same as the total test reliability for both the gender-based subgroups and ethnicity-based subgroups for all the forms.

Table 2 shows the weights on the section scores and the total score in the computation of the augmented subscores for Forms 1 and 3 of Assessment A.

Table 2

Weights on the Total Score and the Section Score in the Computation of the Augmented Subscores for Forms 1 and 3 of Assessment A

		Form 1 sections				Form 3 sections			
Group		1	2	3	4	1	2	3	4
Overall	Weight on Section	0.37	0.24	0.76	0.09	0.44	0.28	0.74	0.15
	Weight on Total	0.23	0.23	0.03	0.08	0.21	0.22	0.04	0.08
Language Group 1	Weight on Section	0.25	0.15	0.73	0.22	0.39	0.28	0.75	0.28
	Weight on Total	0.27	0.26	0.04	0.07	0.21	0.22	0.04	0.07
Language Group 2	Weight on Section	0.38	0.23	0.68	0.08	0.42	0.20	0.63	0.04
	Weight on Total	0.24	0.24	0.04	0.08	0.22	0.25	0.04	0.09
Language Group 3	Weight on Section	0.27	0.32	0.69	-0.05	0.30	0.20	0.68	-0.04
	Weight on Total	0.28	0.21	0.04	0.09	0.27	0.25	0.05	0.09
Language Group 4	Weight on Section	0.35	0.19	0.76	-0.02	0.45	0.28	0.75	0.19
	Weight on Total	0.24	0.24	0.03	0.09	0.18	0.23	0.04	0.07
Language Group 5	Weight on Section	0.38	0.26	0.74	0.13	0.40	0.31	0.74	0.06
	Weight on Total	0.21	0.23	0.04	0.08	0.21	0.22	0.04	0.08
Language Group 6	Weight on Section	0.28	0.32	0.66	0.00	0.43	0.35	0.66	0.08
	Weight on Total	0.27	0.20	0.04	0.09	0.21	0.21	0.04	0.08

The weight for the third section, which is shown in Table 1 to have added value for all the groups shown, is stable over the subgroups. The weight for the fourth section, which was furthest from having added value in Table 1, varies substantially, although all weights for the section score are very small. Indeed, some weights are negative (for example, Group 3). The extent of variation of weights between the subgroups for the first two sections is in-between that of the last two sections. At best, augmented subscores s_{sx} provide only limited gains over the best results from use of s_s or s_x.

Assessment B

Figures 3 and 4 show the standardized subscore means of the gender-based subgroups and the ethnicity-based subgroups for the Assessments B1 and B2. The subscore profiles of the subgroups are close to being parallel.

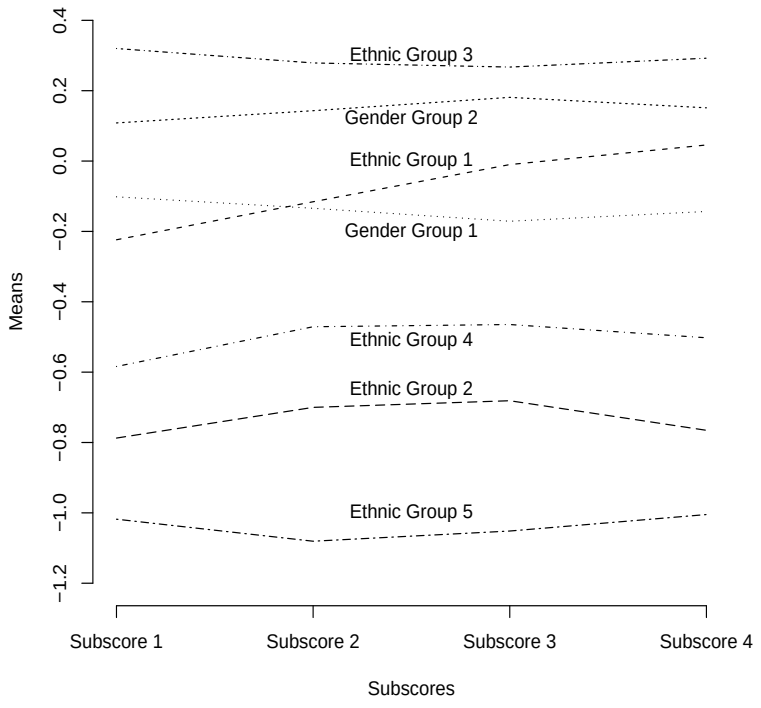

Figure 3. **The standardized subgroup means for Assessment B1.**

Table 3 shows the estimated reliability coefficients of the subscores for several subgroups for Assessments B1 and B2.

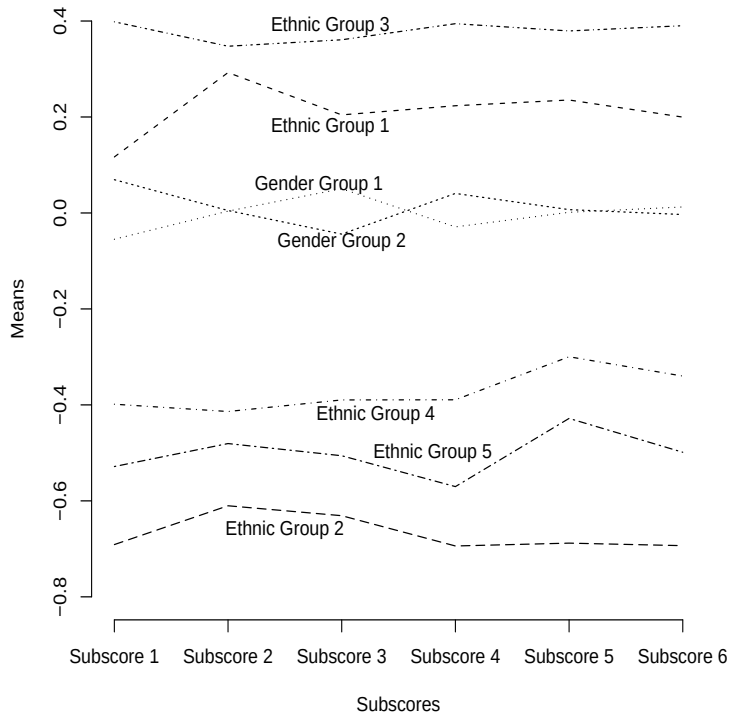

Figure 4. **The standardized subgroup means for Assessment B2.**

For the full sample and for the subgroups, some of the disattenuated correlations were larger than 1. In such a case, it is concluded, without computing any PRMSEs, that the subscores do not have added value (see, for example, Sinharay & Haberman, 2008). So PRMSEs were not computed for the full sample or for any subgroup for Assessments B1 and B2.

The values of PRMSE^x_{xg} and PRMSE^x_{xg*} are the same (and roughly equal to the total score reliability for the full sample) for all the subgroups for both Assessments B1 and B2. The values of PRMSE^x_{x+} are the same as the total test reliability for both the gender-based subgroups and ethnicity-based subgroups for both the assessments.

Assessment C

Figures 5 to 8 show the standardized subscore means of the gender-based subgroups and the ethnicity-based subgroups for Assessment C. The lines for the subgroups appear to be roughly parallel in these figures.[3] Therefore, we would expect β_{xg} and β_{sxg} to be small. However, examinees in Ethnic Group 1 score less on average than those in other ethnic subgroups on all the

Table 3

Reliabilities and Correlations for Assessments B1 and B2

	Assessment B1 subscores				Assessment B2 subscores					
	1	2	3	4	1	2	3	4	5	6
Raw score range	0-16	0-14	0-16	0-14	0-16	0-12	0-13	0-13	0-9	0-13
	All examinees (sizes: 6,563, 7,362)									
Correlation	1.00	0.79	0.70	0.75	1.00	0.73	0.73	0.76	0.74	0.79
	1.06	1.00	0.69	0.71	**0.99**	1.00	0.72	0.73	0.68	0.73
	0.95	**0.99**	1.00	0.72	**0.98**	**1.01**	1.00	0.72	0.69	0.73
	0.96	**0.96**	**0.97**	1.00	**0.99**	**0.99**	**0.98**	1.00	0.71	0.74
					1.02	**0.98**	**0.99**	**0.98**	1.00	0.75
					1.02	**0.99**	**0.98**	**0.96**	**1.03**	1.00
Reliability	0.77	0.71	0.69	0.78	0.77	0.71	0.71	0.77	0.70	0.78
	Gender Group 1 (sizes: 3,322, 3,784)									
Reliability	0.78	0.68	0.66	0.80	0.79	0.74	0.73	0.78	0.68	0.77
	Gender Group 2 (sizes: 3,224, 3,533)									
Reliability	0.75	0.61	0.65	0.77	0.76	0.68	0.69	0.75	0.66	0.76
	Ethnic Group 1 (sizes: 217, 206)									
Reliability	0.79	0.65	0.66	0.78	0.78	0.75	0.71	0.79	0.57	0.76
	Ethnic Group 2 (sizes: 1,416, 2,214)									
Reliability	0.76	0.67	0.59	0.75	0.65	0.50	0.51	0.61	0.54	0.64
	Ethnic Group 3 (sizes: 4,499, 4,388)									
Reliability	0.65	0.52	0.61	0.72	0.73	0.69	0.69	0.73	0.63	0.73
	Ethnic Group 4 (sizes: 343, 338)									
Reliability	0.73	0.63	0.60	0.75	0.67	0.62	0.58	0.69	0.61	0.69

subareas for Assessment C1—so we would expect β_{sg} to be large for this subgroup.

The results for Assessments C1 and C2 are shown in Table 4. The table demonstrates that the added value for the full sample is the same as that for the subgroups. For the two forms of Assessment C1, only the first subscore has added value, and that added value is very small. For the two forms of Assessment C2, the first and third subscores have added value. There are some

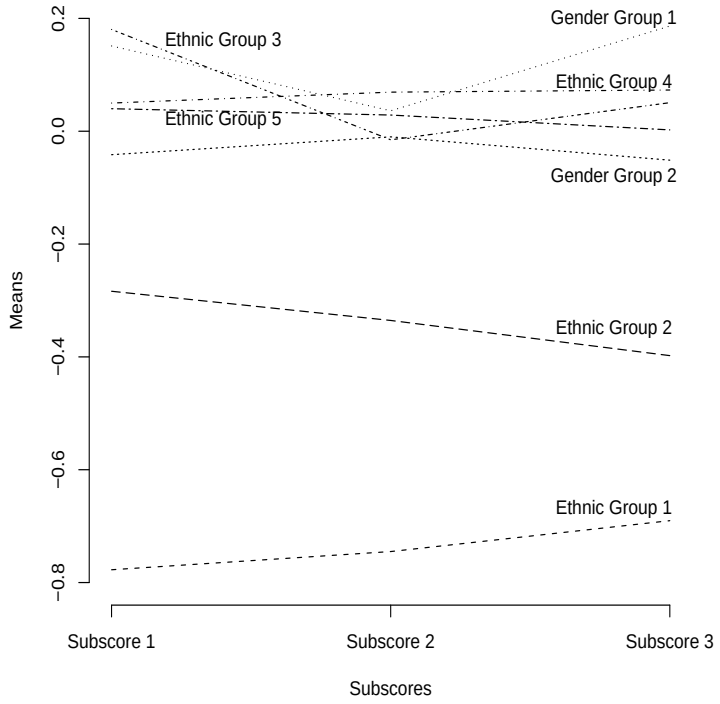

Figure 5. **The standardized subgroup means for the first form of Assessment C1.**

exceptions to the general pattern. Ethnic Group 1 does not have any subscore with added value for Form 1 of Assessment C1, for example. For any subscore, the differences between PRMSE_{sg}, PRMSE_{xg}, and PRMSE_{sxg} are similar to those between PRMSE_{sg*}, PRMSE_{xg*}, and PRMSE_{sxg*}. For example, for both forms of Assessment C1, PRMSE_{sg} is larger than PRMSE_{xg} only for the first subscores and PRMSE_{sg*} is larger than PRMSE_{xg*} for the first subscore for most of the subgroups.

For most subgroups, PRMSE_{sg} is almost the same as PRMSE_{sg*}, PRMSE_{xg} is almost the same as PRMSE_{xg*}, and PRMSE_{sxg} is almost the same as PRMSE_{sxg*}. There are some subgroups for which the normalized bias β_{sg} is different from zero and, as a consequence, PRMSE_{sg*} is somewhat smaller than PRMSE_{sg}. For example, for Ethnic Group 1, $\beta_{sg} = 0.18$ and 0.19 for the third subscore for the two forms of Assessment C1. This result is expected from Figures 5 and 6, because the profiles of Ethnic Group 1 fall lower than those of the other subgroups. As a result, PRMSE_{sg*} is smaller than PRMSE_{sg} by 0.04 for the two forms. Figures 5 and 6 show that the subscore profiles of the subgroups are most often parallel. This results in β_{xg} and β_{sxg} not

Table 4

Proportional Reduction in Mean-Squared Errors (PRMSEs) for Assessments C1 and C2

	C1 Form 1 subscores			C1 Form 2 subscores			C2 Form 1 subscores			C2 Form 2 subscores		
	1	2	3	1	2	3	1	2	3	1	2	3
Length	82	88	35	82	88	35	66	67	67	67	67	66
All examinees (sizes: 4,241, 3,870, 1,931, 19,41)												
Correlation	1.00	0.80	0.77	1.00	0.77	0.74	1.00	0.72	0.60	1.00	0.79	0.66
	0.90	1.00	0.75	**0.87**	1.00	0.74	**0.82**	1.00	0.76	**0.90**	1.00	0.78
	0.91	**0.90**	1.00	**0.89**	**0.90**	1.00	**0.68**	**0.88**	1.00	**0.75**	**0.91**	1.00
PRMSE_{s}	**0.91**	0.88	0.78	**0.90**	0.88	0.78	**0.89**	0.85	0.**87**	**0.90**	0.86	**0.87**
PRMSE_{x}	0.90	0.89	0.87	0.89	0.88	0.85	0.78	0.89	0.79	0.84	0.92	0.82
PRMSE_{sx}	0.93	0.92	0.89	0.92	0.91	0.88	0.91	0.91	0.89	0.91	0.93	0.90
Gender Group 1 (sizes: 915, 840, 628, 636)												
Correlation	1.00	0.81	0.78	1.00	0.78	0.75	1.00	0.71	0.59	1.00	0.79	0.66
	0.90	1.00	0.75	**0.88**	1.00	0.74	**0.81**	1.00	0.75	**0.90**	1.00	0.78
	0.92	**0.89**	1.00	**0.90**	**0.89**	1.00	**0.66**	**0.87**	1.00	**0.74**	**0.91**	1.00
PRMSE_{sg}	**0.92**	0.89	0.79	**0.91**	**0.88**	0.78	**0.90**	0.84	**0.88**	**0.90**	0.86	**0.87**
PRMSE_{xg}	0.91	0.90	0.87	0.89	0.88	0.86	0.77	0.87	0.78	0.83	0.92	0.81
PRMSE_{sxg}	0.94	0.92	0.89	0.93	0.91	0.88	0.91	0.90	0.89	0.91	0.93	0.90
β_{sg}	-0.01	0.00	-0.05	-0.02	0.00	-0.05	-0.02	-0.04	-0.03	-0.02	-0.03	-0.03
β_{xg}	0.00	0.01	-0.02	-0.01	0.01	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
β_{sxg}	-0.02	0.04	-0.07	-0.05	0.06	-0.06	0.00	-0.01	-0.02	0.00	0.00	-0.03
PRMSE_{sg*}	**0.92**	0.89	0.79	**0.91**	**0.88**	0.77	**0.90**	0.84	**0.87**	**0.90**	0.86	**0.87**
PRMSE_{xg*}	0.91	0.90	0.86	0.89	0.88	0.86	0.77	0.87	0.78	0.83	0.92	0.81
PRMSE_{sxg*}	0.94	0.92	0.89	0.92	0.91	0.88	0.91	0.90	0.89	0.91	0.93	0.90
Gender Group 2 (sizes: 3,326, 3,030, 1,303, 13,05)												
Correlation	1.00	0.80	0.76	1.00	0.78	0.74	1.00	0.72	0.59	1.00	0.78	0.65
	0.90	1.00	0.75	**0.88**	1.00	0.74	**0.83**	1.00	0.75	**0.90**	1.00	0.77
	0.91	**0.91**	1.00	**0.88**	**0.91**	1.00	**0.68**	**0.89**	1.00	**0.74**	**0.90**	1.00
PRMSE_{sg}	**0.91**	0.88	0.78	**0.90**	0.88	0.77	**0.89**	0.84	**0.85**	**0.89**	0.85	**0.87**
PRMSE_{xg}	0.90	0.89	0.86	0.89	0.88	0.85	0.77	0.89	0.78	0.83	0.92	0.81
PRMSE_{sxg}	0.93	0.91	0.89	0.92	0.91	0.88	0.90	0.90	0.88	0.91	0.92	0.89
β_{sg}	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.02	0.02	0.01	0.02	0.02
β_{xg}	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{sxg}	0.01	-0.01	0.02	0.01	-0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.01
PRMSE_{sg*}	**0.91**	0.88	0.78	**0.90**	**0.88**	0.77	**0.89**	0.84	**0.85**	**0.89**	0.85	**0.86**
PRMSE_{xg*}	0.90	0.89	0.86	0.88	0.88	0.85	0.77	0.89	0.78	0.83	0.92	0.81
PRMSE_{sxg*}	0.93	0.91	0.89	0.92	0.91	0.88	0.90	0.90	0.88	0.91	0.92	0.89
PRMSE_{s+}	**0.91**	0.88	0.78	**0.90**	**0.88**	0.78	**0.89**	0.85	**0.87**	**0.90**	0.86	**0.87**
PRMSE_{x+}	0.90	0.89	0.87	0.89	0.88	0.86	0.77	0.88	0.78	0.84	0.92	0.82
PRMSE_{sx+}	0.93	0.92	0.89	0.92	0.91	0.88	0.91	0.91	0.89	0.91	0.93	0.90

	C1 Form 1 subscores			C1 Form 2 subscores			C2 Form 1 subscores			C2 Form 2 subscores		
	1	2	3	1	2	3	1	2	3	1	2	3
Ethnic Group 1 (sizes: 187, 166, 133, 129)												
Correlation	1.00	0.84	0.77	1.00	0.77	0.73	1.00	0.75	0.62	1.00	0.78	0.64
	0.93	1.00	0.75	**0.87**	1.00	0.78	**0.88**	1.00	0.77	**0.90**	1.00	0.77
	0.93	**0.92**	1.00	**0.88**	**0.94**	1.00	**0.70**	**0.91**	1.00	**0.73**	**0.90**	1.00
PRMSE_{sg}	0.91	0.89	0.75	**0.89**	0.87	0.79	**0.88**	0.83	**0.87**	**0.88**	0.85	**0.86**
PRMSE_{xg}	0.92	0.92	0.88	0.87	0.89	0.87	0.80	0.92	0.80	0.82	0.92	0.80
PRMSE_{sxg}	0.94	0.93	0.89	0.92	0.91	0.89	0.90	0.92	0.89	0.90	0.92	0.89
β_{sg}	0.07	0.09	0.18	0.10	0.11	0.19	-0.03	-0.02	0.02	-0.03	-0.02	0.00
β_{xg}	0.00	0.00	0.01	0.01	0.00	0.01	-0.02	0.00	0.02	-0.01	0.00	0.02
β_{sxg}	0.04	0.03	0.04	0.06	0.03	0.04	-0.05	-0.02	0.07	-0.06	-0.01	0.06
PRMSE_{sg*}	0.91	0.88	0.71	**0.88**	0.85	0.75	**0.88**	0.83	**0.87**	**0.88**	0.85	**0.86**
PRMSE_{xg*}	0.92	0.91	0.87	0.87	0.89	0.87	0.80	0.92	0.80	0.82	0.92	0.80
PRMSE_{sxg*}	0.94	0.93	0.88	0.91	0.91	0.89	0.90	0.91	0.89	0.90	0.92	0.88
Ethnic Group 2 (sizes: 229, 191, 1,313, 1,278)												
Correlation	1.00	0.76	0.76	1.00	0.75	0.70	1.00	0.72	0.61	1.00	0.80	0.70
	0.85	1.00	0.74	**0.85**	1.00	0.73	**0.83**	1.00	0.75	**0.91**	1.00	0.80
	0.92	**0.90**	1.00	**0.85**	**0.90**	1.00	**0.69**	**0.88**	1.00	**0.79**	**0.92**	1.00
PRMSE_{sg}	**0.90**	**0.88**	0.76	**0.90**	**0.87**	0.76	**0.90**	0.84	**0.86**	**0.90**	0.86	**0.87**
PRMSE_{xg}	0.88	0.87	0.88	0.87	0.87	0.83	0.79	0.88	0.79	0.85	0.93	0.84
PRMSE_{sxg}	0.92	0.91	0.90	0.92	0.90	0.87	0.91	0.90	0.89	0.92	0.93	0.90
β_{sg}	0.03	0.04	0.10	0.03	0.03	0.07	0.00	0.00	-0.01	0.01	0.01	0.00
β_{xg}	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
β_{sxg}	-0.01	0.02	0.10	0.02	0.00	0.02	0.01	0.00	-0.02	0.02	0.00	-0.02
PRMSE_{sg*}	**0.90**	0.88	0.75	**0.90**	0.87	0.75	**0.90**	0.84	**0.86**	**0.90**	0.86	**0.87**
PRMSE_{xg*}	0.88	0.87	0.88	0.87	0.87	0.83	0.79	0.88	0.79	0.85	0.93	0.84
PRMSE_{sxg*}	0.92	0.91	0.88	0.92	0.90	0.86	0.91	0.90	0.89	0.92	0.93	0.90
Ethnic Group 3 (sizes: 207, 183, 485, 534)												
Correlation	1.00	0.81	0.75	1.00	0.81	0.79	1.00	0.69	0.58	1.00	0.76	0.61
	0.92	1.00	0.74	**0.89**	1.00	0.79	**0.79**	1.00	0.76	**0.87**	1.00	0.76
	0.89	**0.89**	1.00	**0.92**	**0.93**	1.00	**0.66**	**0.88**	1.00	**0.69**	**0.89**	1.00
PRMSE_{sg}	0.90	0.87	0.78	**0.92**	0.89	0.80	**0.89**	0.85	**0.87**	**0.89**	0.85	**0.87**
PRMSE_{xg}	0.91	0.90	0.87	0.91	0.90	0.89	0.75	0.88	0.79	0.80	0.91	0.79
PRMSE_{sxg}	0.93	0.91	0.88	0.94	0.92	0.91	0.90	0.90	0.89	0.90	0.92	0.89
β_{sg}	-0.02	0.00	-0.01	-0.02	-0.01	0.00	0.00	0.00	0.01	-0.01	-0.01	0.00
β_{xg}	-0.01	0.01	0.00	-0.01	0.00	0.02	-0.01	0.00	0.01	-0.01	0.00	0.01
β_{sxg}	-0.06	0.06	0.02	-0.04	0.01	0.07	-0.02	0.00	0.02	-0.03	0.00	0.03
PRMSE_{sg*}	**0.90**	0.87	0.78	0.92	0.89	0.80	**0.89**	0.85	**0.87**	**0.89**	0.85	**0.87**
PRMSE_{xg*}	0.91	0.90	0.84	0.91	0.90	0.89	0.75	0.88	0.79	0.80	0.91	0.79
PRMSE_{sxg*}	0.92	0.91	0.88	0.93	0.92	0.90	0.90	0.90	0.89	0.90	0.92	0.89

	C1 Form 1 subscores			C1 Form 2 subscores			C2 Form 1 subscores			C2 Form 2 subscores		
	1	2	3	1	2	3	1	2	3	1	2	3
Ethnic Group 4 (sizes: 2,845, 2,596, -, -)												
Correlation	1.00	0.79	0.76	1.00	0.76	0.72						
	0.90	1.00	0.74	**0.86**	1.00	0.72						
	0.90	**0.90**	1.00	**0.88**	**0.88**	1.00						
PRMSE_{sg}	**0.90**	0.87	0.77	**0.89**	0.86	0.76						
PRMSE_{xg}	0.90	0.89	0.86	0.88	0.87	0.83						
PRMSE_{sxg}	0.93	0.91	0.88	0.91	0.90	0.87						
β_{sg}	0.00	-0.01	-0.02	-0.01	-0.01	-0.03						
β_{xg}	0.00	0.00	0.00	0.00	0.00	-0.01						
β_{sxg}	0.00	-0.01	-0.02	0.00	0.00	-0.02						
PRMSE_{sg*}	**0.90**	0.87	0.77	**0.89**	0.86	0.76						
PRMSE_{xg*}	0.90	0.89	0.86	0.88	0.86	0.83						
PRMSE_{sxg*}	0.93	0.91	0.88	0.91	0.90	0.87						
Ethnic Group 5 (sizes: 773, 734, -, -)												
Correlation	1.00	0.80	0.77	1.00	0.78	0.75						
	0.89	1.00	0.74	**0.87**	1.00	0.76						
	0.91	**0.89**	1.00	**0.89**	**0.91**	1.00						
PRMSE_{sg}	**0.92**	**0.89**	0.78	**0.91**	0.89	0.78						
PRMSE_{xg}	0.91	0.89	0.86	0.89	0.88	0.86						
PRMSE_{sxg}	0.94	0.92	0.89	0.93	0.92	0.89						
β_{sg}	0.00	0.00	0.00	0.01	0.01	0.03						
β_{xg}	0.00	0.00	0.00	0.00	0.00	0.01						
β_{sxg}	-0.01	0.00	0.02	0.01	-0.01	0.04						
PRMSE_{sg*}	**0.92**	0.89	0.78	**0.91**	0.89	0.78						
PRMSE_{xg*}	0.91	0.89	0.86	0.89	0.88	0.86						
PRMSE_{sxg*}	0.93	0.92	0.89	0.93	0.92	0.89						
PRMSE_{s+}	**0.91**	0.88	0.78	**0.90**	**0.88**	0.78	**0.89**	0.85	**0.87**	**0.90**	0.86	**0.87**
PRMSE_{x+}	0.90	0.89	0.86	0.89	0.88	0.85	0.78	0.88	0.79	0.84	0.92	0.82
PRMSE_{sx+}	0.93	0.92	0.89	0.92	0.91	0.88	0.91	0.91	0.89	0.91	0.93	0.90

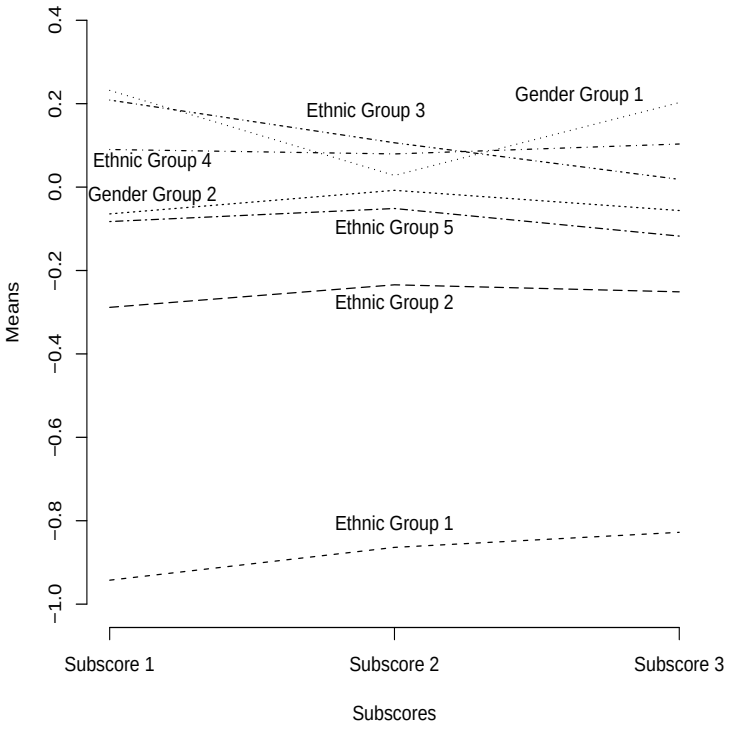

Figure 6. **The standardized subgroup means for the second form of Assessment C1.**

departing much from zero. As a consequence, the differences between PRMSE_{xg} and PRMSE_{xg*} are 0.01 or 0.00, and the same is true for the differences between PRMSE_{sxg} and PRMSE_{sxg*}. These numbers exemplify the case in which it is easier to obtain invariance of the augmented subscore than invariance of the subscores.

The values of PRMSE^{x}_{xg} and PRMSE^{x}_{xg*} are the same (and roughly equal to the total score reliability for the full sample) for all the subgroups for all the forms (results not shown). The values of PRMSE^{x}_{x+} are the same as the total test reliability for all subgroups for all the forms.

The differences are small between PRMSE_{s} and PRMSE_{s+}, and between PRMSE_{x} and PRMSE_{x+}, and between PRMSE_{sx} and PRMSE_{sx+}. Therefore, knowledge of examinee subgroups fails to improve the prediction of true subscores.

Table 5 shows the weights on the section scores and the total score in the computation of the augmented subscores for Assessments C1 and C2. The results for the subgroups do not differ much from the results for the full sample. The weight on the section is larger than that on the total score, with a larger difference of weights for Assessment C2. There is some variation for the small

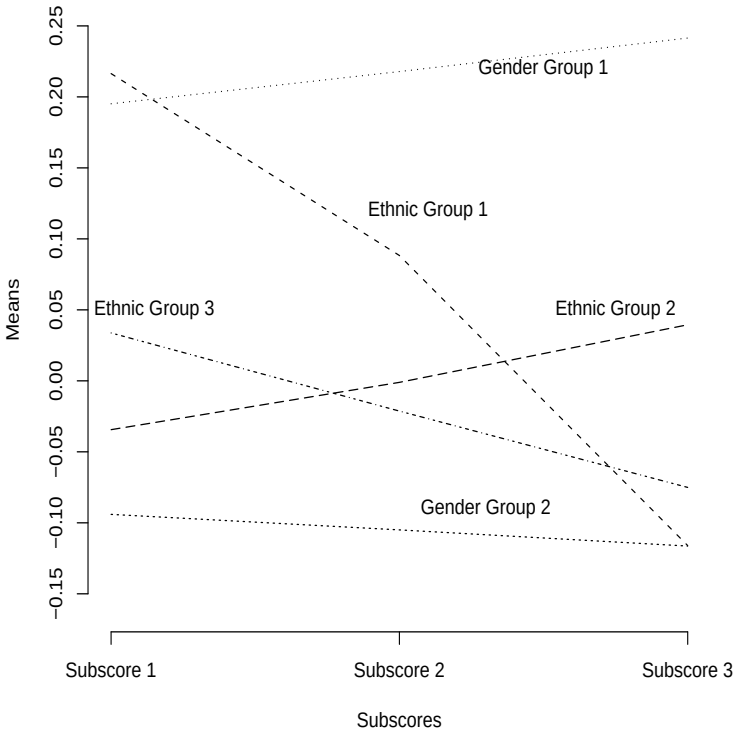

Figure 7. **The standardized subgroup means for the first form of Assessment C2.**

subgroups, the most extreme being the weights for Ethnic Group 1 for Section 2 of Assessment C2. In this case, the total score has a higher weight (0.25) than the section score (0.15). Gains from use of weighted averages are quite modest when compared to the best choice of s_s or s_x.

Assessment D

Figures 9 and 10 show the standardized subscore means of the gender-based subgroups and the ethnicity-based subgroups for Assessment D. Gender Group 1 departs somewhat from the parallel profile pattern in both figures. Ethnic Group 2 shows a notable departure from the parallel pattern in Figure 10, but the group size is rather small. In Figure 9, the profile of Ethnic Group 1 lies far below that of all the other subgroups.

Table 6 shows the results for Assessments D1 and D2. The second subscore of Assessment D1 has added value for all the subgroups for all the forms. The first subscore of Assessment D1 and the third subscore of Assessment D2 never have added value. The added value of the other subscores varies over the subgroups and/or over forms. For example, the third

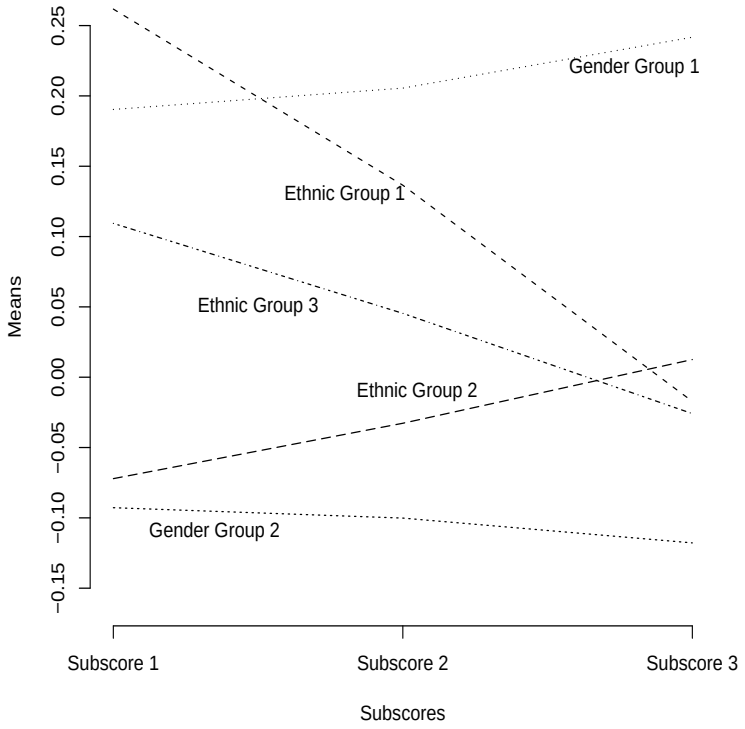

Figure 8. **The standardized subgroup means for the second form of Assessment C2.**

subscore of Assessment D1 does not have added value for the full sample, Gender Groups 1 and 2, and Ethnic Group 1 and 4 but has added value for Ethnic Groups 2 and 3.

There are some subgroups for which the normalized bias β_{sg} is different from zero and, as a result, PRMSE_{sg*} is smaller than PRMSE_{sg}. For example, PRMSE_{sg*} is smaller than PRMSE_{sg} by between 0.04 and 0.11 for the four subscores for Ethnic Group 1 for Assessment D1. Figure 9 shows that the examinees of Ethnic Group 1 score lower on an average than the other subgroups in all areas. However, β_{sxg} is close to zero for Ethnic Group 1 for Assessment D1, and, as a result, PRMSE_{sxg*} is smaller than PRMSE_{sxg} by at most 0.02. There is a somewhat large difference between PRMSE_{sxg} and PRMSE_{sxg*} for the third subscore for Gender Group 1 for Assessment D1, which is a result of the value of β_{sxg} being large (0.23). This phenomenon is expected, because the lines for the two genders are not parallel in Figure 9.

The values of PRMSE^{x}_{xg} and PRMSE^{x}_{xg*} are close for all the subgroups for both assessments D1 and D2. The values of PRMSE^{x}_{x+} are close to the total test reliability for all subgroups for both D1 and D2.

Table 5

Weights on the Total Score and the Section Score in the Computation of the Augmented Subscores for One Form Each for Assessments C1 and C2

		C1 Form 1 sections			C2 Form 1 sections		
Group		1	2	3	1	2	3
Overall	Weight on Section	0.49	0.40	0.28	0.71	0.33	0.61
	Weight on Total	0.20	0.21	0.12	0.09	0.19	0.11
Gender Group 1	Weight on Section	0.50	0.42	0.29	0.73	0.36	0.65
	Weight on Total	0.20	0.20	0.12	0.08	0.18	0.10
Gender Group 2	Weight on Section	0.49	0.39	0.27	0.69	0.31	0.59
	Weight on Total	0.20	0.22	0.12	0.09	0.20	0.11
Ethnic Group 1	Weight on Section	0.42	0.34	0.20	0.65	0.15	0.60
	Weight on Total	0.23	0.24	0.12	0.10	0.25	0.12
Ethnic Group 2	Weight on Section	0.56	0.50	0.20	0.70	0.33	0.59
	Weight on Total	0.17	0.18	0.13	0.09	0.19	0.11
Ethnic Group 3	Weight on Section	0.44	0.34	0.23	0.71	0.38	0.62
	Weight on Total	0.22	0.23	0.11	0.08	0.18	0.11
Ethnic Group 4	Weight on Section	0.48	0.38	0.28			
	Weight on Total	0.21	0.22	0.12			
Ethnic Group 5	Weight on Section	0.52	0.46	0.29			
	Weight on Total	0.19	0.19	0.12			

The differences are negligible between PRMSE_s and PRMSE_{s+}, between PRMSE_x and PRMSE_{x+}, and between PRMSE_{sx} and PRMSE_{sx+}. Knowledge of examinee subgroups does not lead to better prediction of true subscores.

Table 7 shows the weights on the section scores and the total score in the computation of the augmented subscores for Assessments D1 and D2. Other than the small Ethnic Group 2, the weights do not vary much over the subgroups. The weight on the section is almost always larger

Table 6

Proportional Reduction in Mean-Squared Errors

for Assessments D1 and D2

	Assessment D1 subscores				Assessment D2 subscores		
	1	2	3	4	1	2	3
All examinees (sizes: 6,641, 5,134)							
PRMSE_{s}	0.63	**0.86**	0.75	0.78	0.87	0.84	0.82
PRMSE_{x}	0.72	0.75	0.76	0.82	0.87	0.85	0.89
PRMSE_{sx}	0.78	0.88	0.83	0.86	0.91	0.89	0.90
Gender Group 1 (sizes: 677, 341)							
PRMSE_{sg}	0.66	**0.87**	0.77	**0.83**	**0.88**	**0.85**	0.82
PRMSE_{xg}	0.74	0.74	0.80	0.81	0.86	0.84	0.88
PRMSE_{sxg}	0.79	0.89	0.85	0.88	0.91	0.89	0.90
β_{sg}	0.03	-0.03	-0.17	-0.10	0.00	-0.03	0.02
β_{xg}	0.12	0.02	-0.09	-0.03	0.01	-0.03	0.03
β_{sxg}	0.25	0.03	-0.23	-0.10	0.02	-0.09	0.09
PRMSE_{sg*}	0.65	**0.87**	0.74	0.82	**0.88**	**0.85**	0.82
PRMSE_{xg*}	0.72	0.74	0.79	0.81	0.86	0.84	0.88
PRMSE_{sxg*}	0.73	0.88	0.80	0.86	0.91	0.88	0.89
Gender Group 2 (sizes: 5,964, 4,893)							
PRMSE_{sg}	0.62	**0.86**	0.74	0.77	0.87	0.84	0.82
PRMSE_{xg}	0.74	0.76	0.76	0.81	0.87	0.85	0.89
PRMSE_{sxg}	0.78	0.88	0.82	0.85	0.91	0.89	0.90
β_{sg}	0.00	0.00	0.02	0.01	0.00	0.00	0.00
β_{xg}	-0.02	0.00	0.01	0.00	0.00	0.00	0.00
β_{sxg}	-0.03	0.00	0.03	0.01	0.00	0.01	-0.01
PRMSE_{sg*}	0.62	**0.86**	0.74	0.77	**0.87**	**0.84**	0.82
PRMSE_{xg*}	0.74	0.76	0.75	0.81	0.87	0.85	0.89
PRMSE_{sxg*}	0.78	0.88	0.82	0.85	0.91	0.89	0.90
PRMSE_{sg+}	0.63	**0.86**	0.76	0.78	**0.87**	**0.84**	0.82
PRMSE_{xg+}	0.74	0.76	0.77	0.82	0.87	0.85	0.89
PRMSE_{sxg+}	0.78	0.88	0.83	0.86	0.91	0.89	0.90
Ethnic Group 1 (sizes: 642, 1,172)							
PRMSE_{sg}	**0.63**	**0.79**	0.64	**0.74**	**0.84**	0.78	0.77
PRMSE_{xg}	0.61	0.69	0.70	0.71	0.84	0.81	0.84
PRMSE_{sxg}	0.72	0.82	0.76	0.80	0.88	0.85	0.86
β_{sg}	0.33	0.18	0.28	0.24	0.08	0.12	0.13
β_{xg}	0.02	0.04	0.00	0.02	0.00	0.02	0.01
β_{sxg}	0.06	0.12	0.04	0.08	0.02	0.06	0.05
PRMSE_{sg*}	0.52	**0.75**	0.55	0.68	**0.83**	0.77	0.75
PRMSE_{xg*}	0.60	0.66	0.68	0.71	0.83	0.80	0.84
PRMSE_{sxg*}	0.70	0.80	0.74	0.79	0.88	0.84	0.86

	Assessment D1 subscores				Assessment D2 subscores		
	1	2	3	4	1	2	3
Ethnic Group 2 (sizes: 156, 417)							
PRMSE_{sg}	0.67	**0.85**	**0.75**	0.76	**0.87**	**0.81**	0.81
PRMSE_{xg}	0.73	0.68	0.68	0.84	0.85	0.80	0.85
PRMSE_{sxg}	0.79	0.87	0.81	0.87	0.90	0.86	0.87
β_{sg}	0.18	0.07	0.10	0.11	0.08	0.07	0.12
β_{xg}	0.04	0.01	-0.01	0.01	0.02	-0.04	0.03
β_{sxg}	0.10	0.03	0.00	0.05	0.08	-0.08	0.11
PRMSE_{sg*}	0.64	**0.85**	**0.74**	0.75	**0.87**	0.80	0.79
PRMSE_{xg*}	0.72	0.68	0.68	0.84	0.84	0.77	0.84
PRMSE_{sxg*}	0.77	0.86	0.80	0.86	0.89	0.84	0.86
Ethnic Group 3 (sizes: 5,251, 3,193)							
PRMSE_{sg}	0.56	**0.83**	**0.73**	0.75	0.83	0.83	0.77
PRMSE_{xg}	0.69	0.70	0.72	0.79	0.84	0.84	0.86
PRMSE_{sxg}	0.74	0.85	0.80	0.83	0.88	0.88	0.87
β_{sg}	-0.07	-0.03	-0.03	-0.04	-0.06	-0.06	-0.09
β_{xg}	-0.01	-0.01	0.00	0.00	-0.01	0.00	-0.02
β_{sxg}	-0.03	-0.02	0.00	-0.01	-0.03	-0.01	-0.05
PRMSE_{sg*}	0.55	**0.83**	**0.73**	0.75	**0.82**	0.82	0.76
PRMSE_{xg*}	0.68	0.70	0.72	0.79	0.83	0.84	0.86
PRMSE_{sxg*}	0.73	0.85	0.80	0.83	0.88	0.88	0.87
Ethnic Group 4 (sizes: 592, 452)							
PRMSE_{sg}	0.73	**0.90**	0.82	0.81	**0.87**	**0.84**	0.81
PRMSE_{xg}	0.77	0.82	0.83	0.87	0.85	0.82	0.88
PRMSE_{sxg}	0.83	0.92	0.88	0.89	0.90	0.88	0.89
β_{sg}	0.08	0.03	0.01	0.02	0.05	0.04	0.08
β_{xg}	0.02	0.02	-0.02	-0.02	0.01	-0.01	0.02
β_{sxg}	0.07	0.05	-0.05	-0.05	0.04	-0.02	0.06
PRMSE_{sg*}	0.71	**0.90**	**0.81**	0.81	**0.87**	**0.84**	0.80
PRMSE_{xg*}	0.76	0.82	0.83	0.87	0.85	0.82	0.88
PRMSE_{sxg*}	0.82	0.91	0.87	0.89	0.90	0.88	0.89
PRMSE_{s+}	0.65	**0.86**	**0.76**	0.79	**0.88**	0.85	0.83
PRMSE_{x+}	0.73	0.76	0.76	0.82	0.88	0.86	0.89
PRMSE_{sx+}	0.78	0.88	0.83	0.86	0.91	0.89	0.90

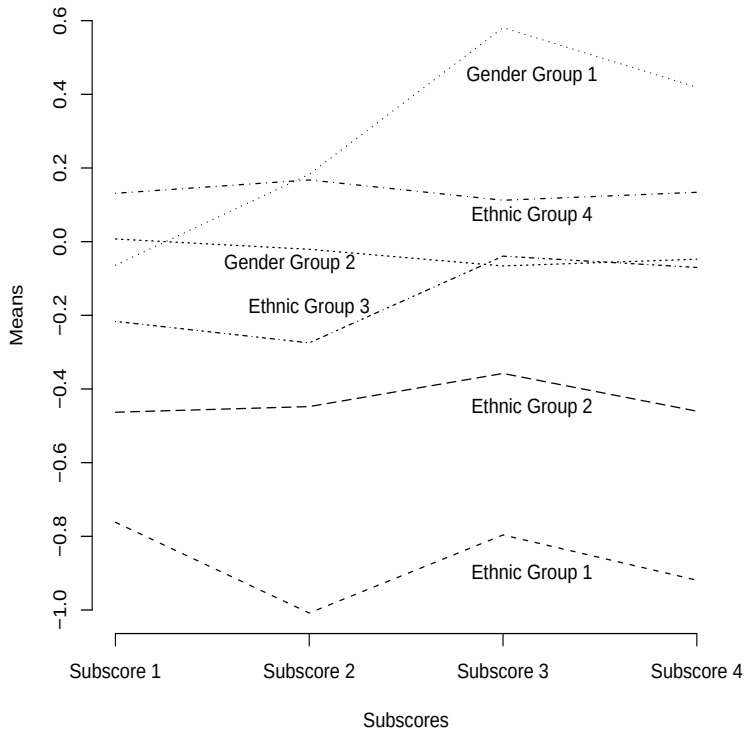

Figure 9. **The standardized subgroup means for Assessment D1.**

than that on the total score. The gain from use of augmented subscores is somewhat variable. For the first, third, and fourth subscores of Assessment D1 and for the first and second subscores of Assessment D2, the overall results favor augmented subscores by a relatively substantial amount. Results for Gender Group 1 are less favorable for the first and third subscores of D1. The overall augmented subscore appears rather effective for both assessments for all groups.

Conclusions

The results from the four assessments show that whether a subscore has added value over the total score is unaffected by whether the analysis is performed at the overall level or at the subgroup level.

For the data examined, the answers to the questions we asked in the beginning of the paper seem to be the following:

- For the assessments considered here, the prediction of the true subscore is not improved by knowledge of the subgroup membership of the examinees.

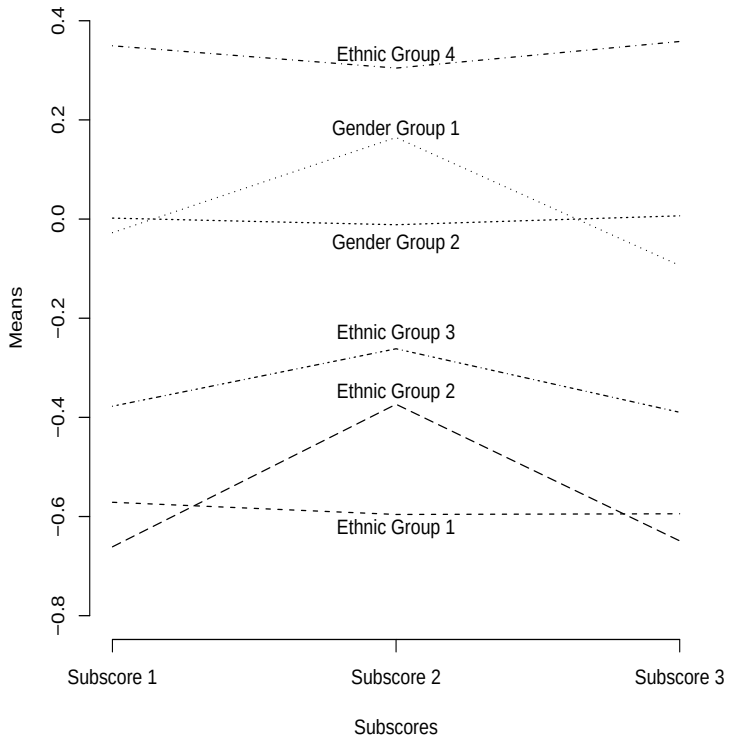

Figure 10. **The standardized subgroup means for Assessment D2.**

- Whether the observed subscore is a better estimate than the observed total score of the true subscore for each subgroup is specific to each test. We found that the answers are (a) "yes" for one out of the four parts for Assessment A, (b) "no" for all the subscores of Assessment B, (c) "yes" for one out of the three subscores on an average for Assessment C1 and "yes" for two out of the three subscores for Assessment C2, (d) "yes" for two out of the four subscores on an average for Assessment D1 and "yes" for one out of the three subscores on an average for Assessment D2.

- The added value of the subscores does not differ over subgroups. For almost all data sets, either a subscore has added value for all subgroups or does not have added value for any subgroup.

The invariance criterion was more likely to hold for augmented subscores than for subscores or total scores. In several cases, PRMSE_{sg} was somewhat different from PRMSE_{sg*} because of a non-zero value of β_{sg}, but PRMSE_{sxg} was very close to PRMSE_{sxg*}. This finding regarding

Table 7

Weights on the Total Score and the Section Score in the Computation of the Augmented Subscores for Assessments D1 and D2

		Assessment D1 sections				Assessment D2 sections		
Group		1	2	3	4	1	2	3
Overall	Weight on section	0.27	0.64	0.40	0.36	0.45	0.42	0.24
	Weight on total	0.10	0.10	0.12	0.15	0.17	0.17	0.22
Gender Group 1	Weight on section	0.30	0.67	0.38	0.47	0.52	0.47	0.28
	Weight on total	0.10	0.09	0.13	0.13	0.16	0.16	0.20
Gender Group 2	Weight on section	0.25	0.63	0.38	0.34	0.44	0.41	0.24
	Weight on total	0.11	0.10	0.13	0.15	0.18	0.18	0.22
Ethnic Group 1	Weight on section	0.38	0.58	0.31	0.44	0.44	0.38	0.26
	Weight on total	0.09	0.10	0.12	0.12	0.18	0.17	0.20
Ethnic Group 2	Weight on section	0.32	0.69	0.50	0.26	0.52	0.45	0.34
	Weight on total	0.11	0.08	0.10	0.18	0.16	0.14	0.19
Ethnic Group 3	Weight on section	0.24	0.63	0.41	0.35	0.40	0.39	0.21
	Weight on total	0.10	0.09	0.12	0.15	0.17	0.20	0.21
Ethnic Group 4	Weight on section	0.36	0.65	0.41	0.30	0.50	0.49	0.22
	Weight on total	0.10	0.10	0.13	0.16	0.16	0.15	0.22

invariance, along with the research finding that the augmented subscores mostly lead to better diagnostic information than subscores (Sinharay, 2010), shows that the augmented subscores are viable alternatives to subscores.

References

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

Livingston, S. A., & Rupp, S. L. (2004). *Performance of men and women on multiplechoice and constructed-response tests for beginning teachers* (ETS Research Report No. RR-04-48). Princeton, NJ: ETS.

Lyren, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, and Evaluation*, *14*(4), 1–10.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266–285.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (ETS Research Memorandum No. RM-08-18). Princeton, NJ: ETS.

Stricker, L. J. (1993). *Discrepant LSAT subscores* (Technical Report No. 93-01). Newtown, PA: Law School Admission Council.

Notes

[1]Here, for example, PRMSE_s is computed just like PRMSE_{sg}, but using all examinees in the sample.

[2]Note that it is possible to perform the standardization by using the standard deviation of the true subscore instead of that of the observed subscore. The figures would look very similar to Figures 1 and 2 in that case.

[3]The only exceptions are Gender Group 1 and Ethnic Group 3 for both forms of Assessment C1 and Ethnic Groups 1 and 3 for both forms of Assessment C2. These subgroups are small, however, except for Gender Group 1 for Assessment C1.