British Journal of Mathematical and Statistical Psychology (2013), 66, 452–469
© 2012 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Does subgroup membership information lead to better estimation of true subscores?

Shelby J. Haberman[1] and Sandip Sinharay[1,2]*

[1]Educational Testing Service, Princeton, New Jersey, USA
[2]CTB/McGraw-Hill, Monterey, California, USA

Haberman (2008) suggested a method to determine if subtest scores have added value over the total score. The method is based on classical test theory and considers the estimation of the true subscores. Performance of subgroups, for example, those based on gender or ethnicity, on subtests is often of interest. Researchers such as Stricker (1993) and Livingston and Rupp (2004) found that the difference in performance between the subgroups often varies over the different subtests. We suggest a method to examine whether the knowledge of the subgroup membership of the examinees leads to a better estimation of the true subscores. We apply our suggested method to data from two operational testing programmes. The knowledge of the subgroup membership of the examinees does not lead to a better estimation of the true subscore for the data sets.

1. Introduction

Haberman (2008) suggested a method based on classical test theory (CTT) to examine whether subscores have added value over total scores. The method focuses on the estimation of the true subscore from the observed subscore and the scores on other parts of the test. A subscore has added value when the mean square error (MSE) of the regression of the true subscore on the observed subscore is smaller than the MSE of the regression of the true subscore on the observed total score. Haberman (2008), Haberman, Sinharay, and Puhan (2009), Ling (2009), Lyren (2009), Puhan, Sinharay, Haberman, and Larkin (2010) and Sinharay (2010) applied the method of Haberman (2008) to data sets from a variety of testing programmes.

Several researchers found that the difference in performance between the subgroups (e.g., those based on gender or ethnicity) varied over the different subtests (Livingston & Rupp, 2004; Stricker, 1993). Stricker (1993) studied the performance of subgroups on the subtests of the Law School Admissions Test (LSAT) and found, for example, that the difference between male examinees and female examinees for the analytical reasoning

*Correspondence should be addressed to Sandip Sinharay, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA (e-mail: sandip_sinharay@ctb.com).
Sandip Sinharay conducted this study and wrote this report while on staff at Educational Testing Service. He is currently at CTB/McGraw-Hill in Monterey, CA.

DOI:10.1111/j.2044-8317.2012.02061.x

subscore of LSAT is considerably smaller than the difference for the logical reasoning subscore. Livingston and Rupp (2004) found that women tend to perform better, relative to men, on constructed-response tests than on multiple-choice (MC) tests in Praxis Principles of Learning and Teaching tests for secondary school teachers and in subject-knowledge tests for social studies teachers, science teachers, and middle school mathematics teachers.

The analysis of Haberman (2008) considers the estimation of the true subscore based on scores on other parts of the test, but does not incorporate any subgroup information. Therefore, it is important to examine whether the knowledge of subgroup membership of the examinees leads to a better estimation of the true subscore (compared to no knowledge of the subgroup membership) in applications of the method of Haberman (2008). In other words, does the subgroup information have any 'impact' on the estimation of true subscores? Note here that if subgroup membership indeed leads to a better estimation, then we would not recommend the use of subgroup membership to estimate true subscore, but we would recommend a thorough examination of the test content and the test-taking population to find out why the subgroup membership leads to a better estimation of the true subscore. This is a test fairness procedure similar to the study of score equity assessment (Dorans, 2004) or equating invariance (Dorans & Holland, 2000), where one examines whether the equating conversion differs between subgroups; presence of difference would lead not to reporting of different equating conversions but to follow-up analyses (e.g., Dorans, 2004, p. 65).[1]

The next section provides a description of the method of Haberman (2008). Section 3 extends the method of Haberman (2008) to examine whether the knowledge of subgroup membership of the examinees lead to a better estimation of the true subscore. Section 4 describes data from two operational tests that report subscores. Section 5 describes the findings from the application of our suggested methods to the data. Section 6 provides conclusions and recommendations.

2. Haberman's method

This section describes the approach of Haberman (2008) to determine whether and how to report subscores. Let us denote the observed subscore and the observed total score of an examinee by s and x, respectively. Assume that s and x have positive and finite variances, positive reliability coefficients less than 1, and correlation less than 1. Assume that a simple random sample of n examinees is available for estimation. In typical applications, n is somewhat greater than 1,000. Let $\bar{s}$ be the sample mean of the observed subscores, $\bar{x}$ the sample mean of the observed total scores, V_x the estimated variance of the observed total score x, and $\mathrm{Cov}(s_t, x)$ the estimated covariance of the true subscore s_t and the observed total score x. Let $\hat{\rho}_s^2$ be the estimated reliability coefficient of the subscore. Haberman (2008), taking a CTT viewpoint, assumed that a reported subscore is intended to be an estimate of the true subscore s_t and considered the following estimates of s_t based on linear regression:

- $s_s = \bar{s} + \hat{\rho}_s^2(s - \bar{s})$ based on the observed subscore.
- $s_x = \bar{s} + c(x - \bar{x})$ based on the observed total score, where $c = \mathrm{Cov}(s_t, x)/V_x$ depends on the estimated reliability coefficients and standard deviations of the subscore and the total score, and the estimated correlation between the subscores.

[1] It is also similar to differential prediction (e.g., Dorans, 2004), where one studies whether the prediction of a criterion score based on several test scores is the same over subgroups of interest.

- $s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x})$, that is, a weighted average of the observed subscore and the observed total score. The constants a and b are given by

$$a = \frac{\sigma(s)}{\sigma(x)} \rho(s_t, s)\tau,$$

$$b = \rho(s_t, s)[\rho(s_t, s) - \rho(s, x)\tau],$$

and

$$\tau = \frac{\rho(x_t, x)\rho(s_t, x_t) - \rho(s, x)\rho(s_t, s)}{1 - \rho^2(s, x)},$$

where $\sigma(u)$ and $\rho(u,v)$ respectively denote the standard deviation of u and the correlation between u and v. The estimate s_{sx} is a special case of the augmented subscore suggested by Wainer, Sheehan, and Wang (2000). Sinharay (2010) noted that s_{sx} is normally very highly correlated with the augmented subscore (Wainer *et al.*, 2000). Therefore, any result that we find on s_{sx} in this paper will hold for the augmented subscore of Wainer *et al.* (2000) as well. The quantity s_{sx} will be referred to as the *augmented subscore* throughout this paper.

To compare s_s, s_x, and s_{sx} as estimates of the true subscore s_t, Haberman (2008) suggested the use of the proportional reduction in mean square error (PRMSE). For example,

$$\text{PRMSE of } s_x = 1 - \frac{\text{MSE of } s_x}{\text{MSE of } \bar{s}} = 1 - \frac{\text{MSE of } s_x}{\text{Estimate of Var}(s_t)}.$$

The larger the PRMSE, the more accurate is the estimate in the sense that the MSE is smaller. We denote the PRMSE for s_s, s_x, and s_{sx} by PRMSE_s, PRMSE_x, and PRMSE_{sx}, respectively. Similarly, we denote the MSE for s_s, s_x, and s_{sx} by MSE_s, MSE_x, and MSE_{sx}, respectively. The PRMSE is conceptually similar to reliability, and, for the approximation s_s, PRMSE_s is exactly equal to the reliability coefficient of the observed subscore s. The MSE for s_s is conceptually similar to variance of measurement. Haberman (2008) recommended the following strategy to decide whether a subscore or an augmented subscore has added value:

- Declare that a subscore has added value over the total score only if PRMSE_s is larger than PRMSE_x, because the subscore will provide more accurate diagnostic information (in the form of a lower MSE in estimating the true subscore) than the observed total score in that case. Sinharay, Haberman, and Puhan (2007) discussed why this strategy is reasonable and how this ensures that a subscore satisfies professional standards.
- Declare that an augmented subscore has added value only if PRMSE_{sx} is substantially larger than both PRMSE_s and PRMSE_x.

Haberman (2008) and Sinharay *et al.* (2007) explained that a subscore is more likely to have added value when (1) it has high reliability; (2) the total score has low reliability; and (3) it is distinct from other subscores. Note that the methodology does not involve any assumptions except those of CTT.

3. Methods

To examine whether the knowledge of the subgroup membership of the examinees leads to a better estimation of the true subscore, there is a need to extend the CTT-based method of Haberman (2008). This extension considers estimates of true subscores that incorporate subgroup information and estimates that do not incorporate subgroup information and proposes an approach to compare them using appropriate MSEs and PRMSEs. As in score equity assessment or differential prediction analysis (Dorans, 2004), the analysis is intended to lead to follow-up analyses rather than to provide an alternative approach to reporting of subscores.

3.1. Estimates that incorporate subgroup information

Here, we consider linear estimates of true subscores that incorporate the subgroup information in addition to the scores on other parts of the test. Consider examinee subgroups g from 1 to $N_g \geq 2$. Groups will vary with the testing programme. For example, one might consider two subgroups based on gender of the examinees, four subgroups based on the first language of the examinees, or three subgroups based on the race/ethnicity of the examinees. Let $p_g > 0$ be the fraction of the sample in subgroup g. Let the conditional expected value of the measurement error s_e be 0 for examinees in subgroup g. For examinees in subgroup g, let $\hat{\rho}^2_{sg}$ be the estimate of ρ^2_{sg}, the reliability of observed subscore s, let $\bar{s}_g$ be the sample mean for observed subscore s, and let $\bar{x}_g$ be the sample mean for observed total score x. For examinees in subgroup g, let V_{sg} be the estimated variance of s, let V_{xg} be the estimated variance of the observed total score x, let $\mathrm{Cov}_g(s,x)$ be the estimated covariance of the observed subscore s and the observed total score x, and let $\mathrm{Cov}_g(s_t,x)$ be the estimated covariance of the true subscore s_t and the observed total score x.

We consider the following estimates of s_t for subgroup g:

- $\bar{s}_g$, defined in the previous paragraph.
- $s_{sg} = \bar{s}_g + \hat{\rho}^2_{sg}(s - \bar{s}_g)$ based on the observed subscore.
- $s_{sxg} = \bar{s}_g + a_g(s - \bar{s}_g) + b_g(x - \bar{x}_g)$ that is a weighted average of the observed subscore and the observed total score, where a_g and b_g are given by

$$a_g = \frac{\sqrt{V_{sg}}}{\sqrt{V_{xg}}}\rho_g(s_t,s)\tau_g,$$

$$b_g = \rho_g(s_t,s)[\rho_g(s_t,s) - \rho_g(s,x)\tau_g],$$

and

$$\tau_g = \frac{\rho_g(x_t,x)\rho_g(s_t,x_t) - \rho_g(s,x)\rho_g(s_t,s)}{1 - \rho_g^2(s,x)},$$

where, for example, $\rho_g(s_t,s)$ denotes the correlation between s_t and s computed from the examinees in subgroup g.[2]

[2] It is also possible to consider the estimate $s_{xg} = \bar{s}_g + c_g(x - \bar{x}_g)$. Results for s_{xg} are not shown here and can be obtained from the authors on further request.

Note that all the above estimates use the subgroup information. However, we would like to emphasize again that these estimates are not intended to be used in operational subscore reporting, for their use would violate basic fairness principles. They are employed here in the same spirit as in score equity analysis (Dorans, 2004) to examine the impact of subgroups on different estimates of the true subscore s_t.

For each subgroup g, the subgroup-specific estimate $\bar{s_g}$ leads to the subgroup-specific mean square error MSE_g, which is an estimate of

$$E\left((s_t - \mu_{sg})^2 | G = g\right),$$

the variance of s_t for examinees in subgroup g. Similarly, s_{sg}, and s_{sxg} lead to the respective subgroup-specific mean square errors MSE_{sg}, and MSE_{sxg}. For estimating s_t, the subgroup-specific estimates s_{sg} and s_{sxg} have the respective proportional reductions in mean square error (compared to the subgroup-specific estimate $\bar{s_g}$) of PRMSE_{sg} and PRMSE_{sxg}, where

$$\mathrm{PRMSE}_{sg} = 1 - \frac{\mathrm{MSE}_{sg}}{\mathrm{MSE}_g}$$

is the reliability of the subscore computed only using examinees in subgroup g. Let RMSE_g denote the square root of MSE_g, so that RMSE_g estimates the standard deviation of the true subscore s_t for examinees in subgroup g.

3.2. Estimates that do not incorporate subgroup information

The formulas for the estimates s_{sg} and s_{sxg} include not only the scores from the other parts of the test, but also the subgroup information. In addition, the computation of their proportional reductions in mean square error, PRMSE_{sg} and PRMSE_{sxg}, is based only on examinees of subgroup g. To examine if the use of subgroup information leads to a better estimation of the true subscore, we have to compare s_{sg} and s_{sxg} to estimates of s_t whose formulas do not include subgroup information. In addition, the computation of their corresponding PRMSEs has to be based only on examinees of subgroup g.

The function s_s is an estimate of s_t that is based on subscore s and ignores subgroup information. Thus, s_s is comparable to s_{sg}. For examinees in subgroup g, it is shown in the appendix that the estimated MSE of s_s computed from examinees of subgroup g is given by

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

The bias, arising out of ignoring of subgroup information, of s_s for subgroup g is

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

and the normalized bias is

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

This bias is small if either the subgroup means for the subscore vary little for different subgroups or if the reliability of the subscore is high. The corresponding subgroup-specific PRMSE because of the use of s_s instead of $\bar{s_g}$ as an approximation for s_t is

$$\text{PRMSE}_{sg^*} = 1 - \frac{\text{MSE}_{sg^*}}{\text{MSE}_g} = \text{PRMSE}_{sg} - \frac{(\hat{\rho}_{sg}^2 - \hat{\rho}_s^2)^2}{\hat{\rho}_{sg}^2} - \beta_{sg}^2. \tag{4}$$

The function s_{sx} is an estimate of s_t that is based on subscore s, total score x and ignores subgroup information. Thus, s_{sx} is comparable to s_{sxg}. For subgroup g, the estimated mean square error of s_{sx} is

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

where the bias of s_{sx} for subgroup g is

$$B_{sxg} = a(\bar{s}_g - \bar{s}) + b(\bar{x}_g - \bar{x}) - (\bar{s}_g - \bar{s}) \tag{5}$$

and its normalized value is

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

Once again, the bias is likely to be small if subgroups with higher subscore means also have higher means of total scores, which would happen if the subgroups have parallel subscore profiles. The corresponding subgroup-specific PRMSE because of the use of s_{sx} instead of $\bar{s}_g$ as an approximation for s_t is

$$\text{PRMSE}_{sxg^*} = \text{PRMSE}_{sxg} - \frac{(a_g - a)^2 V_{sg} + 2(a_g - a)(b_g - b)\text{Cov}_g(s,x) + (b_g - b)^2 V_{xg}}{\text{MSE}_g} - \beta_{sxg}^2. \tag{7}$$

Note that MSE_{sg^*}, MSE_{sxg^*}, PRMSE_{sg^*}, and PRMSE_{sxg^*} are computed only using subgroup g. Therefore, to ascertain whether use of the subgroup information leads to a better estimation of the true subscore, one can compare MSE_{sg^*} and MSE_{sxg^*} to MSE_{sg} and MSE_{sxg}, or, alternatively, compare PRMSE_{sg^*} and PRMSE_{sxg^*} to PRMSE_{sg} and PRMSE_{sxg}. For example, if PRMSE_{sg^*} is substantially smaller than PRMSE_{sg} or PRMSE_{sxg^*} is substantially smaller than PRMSE_{sxg}, one can conclude that the use of the subgroup information leads to a better estimation of the true subscore and a follow-up analysis might reveal why that happened.

One can also examine the normalized biases β_{sg} and β_{sxg} because some of the differences between PRMSE_{sg^*} and PRMSE_{sg}, or between PRMSE_{sxg^*} and PRMSE_{sxg}, involve these normalized biases. For β_{sg} to be small, one needs the subscore means of the subgroups to be close to each other or the subscore reliability to be very high. On the other hand, for β_{sxg} to be small, it is enough that the subscore profiles of the subgroups are parallel, which would usually happen for tests that are nearly one-dimensional. Thus β_{sxg} would be small more often than β_{sg}. That means that the difference between PRMSE_{sxg^*} and PRMSE_{sxg} will be smaller more often than the difference between PRMSE_{sg^*} and PRMSE_{sg}. Thus it would be easier to achieve subgroup invariance of the augmented subscores than subgroup invariance of the subscores. We will later show examples of cases when β_{sxg} is small (i.e., the difference between PRMSE_{sxg^*} and PRMSE_{sxg} is small) but β_{sg} is not (i.e., the difference between PRMSE_{sg^*} and PRMSE_{sg} is not).

3.3. Overall measure of the effect of subgroups on estimation of true subscores

Some overall summary of the effects of subgroups can be obtained with the respective overall estimated mean square errors

$$\mathrm{MSE}_{s+}=\sum_{g=1}^{N_g} p_g \mathrm{MSE}_{sg} \tag{8}$$

and

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

Thus the resulting proportional reductions in MSE are

$$\mathrm{PRMSE}_{s+}=1-\mathrm{MSE}_{s+}/\mathrm{MSE}_{s} \tag{9}$$

and

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

Large differences between the results for subgroup-specific estimates and corresponding overall estimates warrant attention. For example, larger differences between PRMSE_{s+} and PRMSE_{s1} would indicate there is something unique about the performance on the subscore s of subgroup 1 and would require a follow-up analysis to determine why. In addition, large differences between PRMSE_{s} and PRMSE_{s+}, or between PRMSE_{sx} and PRMSE_{sx+} might suggest that the use of subgroup information leads to better estimation of the true subscore and would require a follow-up analysis. The follow-up analysis would be specific to the test. For example, suppose that the use of subgroup information leads to better estimation of the true subscore for a test of English for non-native speakers. In this case, a follow-up analysis could examine if it is due to factors such as (1) the similarity/dissimilarity of the native languages of the subgroups; (2) the variation in the mode of English instruction of the subgroups; or (3) the cultural bias in some of the test items. If the third factor turns out to be the cause, a redesign of the test could solve the problem in the future. Thus, the application of the method described above and the subsequent follow-up analyses, if any, constitute a fairness procedure that can be adopted by testing companies that report subscores.

Table 1 shows for each PRMSE discussed above (1) the estimate of which it is a PRMSE; (2) the estimate compared to which the proportional reduction in the PRMSE is computed; and (3) how the PRMSE is used to make conclusions for a data set. For example, the second row lists PRMSE_{s}, that is, the proportional reduction in mean square error of the estimate s_s of s_t, is computed in comparison to the estimate $\bar{s}$, and is compared to PRMSE_{x} to determine if a subscore has added value. Note that all the PRMSEs shown in the table correspond to estimates of s_t. All the PRMSEs with a subscript g are computed only from examinees of subgroup g.

3.4. How large is large?

In order to apply the above methodology, some guidance is needed as to when a PRMSE can be interpreted as substantially larger than another and also when a normalized bias

Table 1. List of all PRMSEs

PRMSE	PRMSE of	Compared to	Use of the PRMSE
PRMSE_x	s_x	$\bar{s}$	
PRMSE_s	s_s	$\bar{s}$	The subscore has added value if $\mathrm{PRMSE}_s > \mathrm{PRMSE}_x$
PRMSE_{sx}	s_{sx}	$\bar{s}$	s_{sx} has added value if $\mathrm{PRMSE}_{sx} \gg \max(\mathrm{PRMSE}_x, \mathrm{PRMSE}_s)$
PRMSE_{sg}	s_{sg}	$\bar{s}_g$	
PRMSE_{sxg}	s_{sxg}	$\bar{s}_g$	
PRMSE_{sg^*}	s_s	$\bar{s}_g$	Large differences between PRMSE_{sg^*} and PRMSE_{sg} should lead to a follow-up analysis
PRMSE_{sxg^*}	s_{sx}	$\bar{s}_g$	Large differences between PRMSE_{sxg^*} and PRMSE_{sxg} should lead to a follow-up analysis
PRMSE_{s+}	$s_{sg}\forall g$	$\bar{s}_g\forall g$	Large differences between PRMSE_{s+} and PRMSE_{sg} or between PRMSE_{s+} and PRMSE_s should lead to a follow-up analysis
PRMSE_{sx+}	$s_{sxg}\forall g$	$\bar{s}_g\forall g$	Large differences between PRMSE_{sx+} and PRMSE_{sxg} or between PRMSE_{sx+} and PRMSE_{sx} should lead to a follow-up analysis

Note. The symbols$\gg$ and $\forall$ respectively mean 'substantially larger' and 'for all'.

(β_{sg} or β_{sxg}) is small. This interpretation differs depending on the nature of the assessment and the data. However, we have found from experience that, roughly speaking, one PRMSE (say, PRMSE_1) is appreciably larger than another (PRMSE_2) when PRMSE_1 reduces the distance of PRMSE_2 from 1.0 by at least 10%, that is, when

$$\mathrm{PRMSE}_1 - \mathrm{PRMSE}_2 > 0.1(1 - \mathrm{PRMSE}_2).$$

For example, if PRMSE_2 is 0.80, then PRMSE_1 has to be at least 0.82 to be called substantially larger than PRMSE_2 (because the difference between 0.82 and 0.80 is 10% larger than the distance of 0.80 from 1.0). To interprete the normalized biases, note from equations (4) and (7) that the differences between two PRMSEs involves the squares of these biases.[3] A natural consequence is that a normalized bias can be called small when it leads to a difference between two PRMSEs that is smaller than $0.1(1 - \mathit{PRMSE}_2)$. Therefore, we can recommend that, roughly speaking, β_{sg} is small when its square is smaller than $0.1(1 - \mathrm{PRMSE}_{sg})$. For example, if PRMSE_{sg} is 0.80, then β_{sg} has to be less than 0.1414 to be called small.

4. Data

We obtained data from two testing programmes that report subscores operationally. The descriptions of the tests are provided below. No further details are provided about the data sets due to confidentiality restrictions.

4.1. Test A

Test A is a battery of tests that gauges achievement in several disciplines. Each test under the battery is intended for students who have majored in or have extensive background in

[3] We also found that in the right-hand sides of equations (4) and (7), the contributions of the second terms are usually negligible.

that specific area. We considered the two titles under Test A with the largest volumes – we refer to them here as Tests A1 and A2. We analysed data from one recently administered test form for each of these two tests. The sample sizes were 4,242 for Test A1 and 1,932 for Test A2. In Test A1, which has approximately 205 MC items, the examinees receive two subscores. Some questions (about 17%) are not part of a reported subscore but contribute to the total reported score on Test A1 – we treat these items as contributing to a third subscore for the test. In Test A2, which has approximately 200 MC items, the examinees receive three subscores. We performed a three-subscore analysis for both these tests. We show results for five subgroups based on ethnicity for Test A1 and three subgroups based on ethnicity for Test A2. One of the ethnicity-based subgroups for each test included the small ethnicity-based subgroups (those with less than 100 examinees) and those who did not provide their ethnicity; we thought that it was important to study if there is anything unique about this group rather than omitting this group from our analysis.

4.2. Test B

Test B is a battery of teacher-certification tests. We considered two titles under Test B, referred to as Tests B1 and B2. Test B1 is designed for prospective teachers in primary to upper elementary school grades. It consists of 120 MC items divided into four broad categories of equal length. A subscore is operationally reported for each category. Test B2 is used to measure skills necessary for prospective and practising paraprofessionals. The 75 MC items contribute to three subscores (25 items each). We analysed data from one recently administered test form of Tests B1 and B2. The sample sizes were 6,643 for Test B1 and 5,270 for Test B2. We show results for four subgroups based on ethnicity. One of the subgroups for each test includes the small ethnicity-based subgroups (those with less than 100 examinees) and those who did not provide their ethnicity.

5. Results

5.1. Test A

Figures 1 and 2 show the standardized subscore means of the subgroups for Test A. In these figures, the standardized subscore means for each ethnic group are connected by a line. 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 sample standard deviation and then computed the mean of these standardized values.[4] In these figures the lines for the subgroups appear to be approximately parallel, except for ethnic group 3 for Test A1 and ethnic groups 1 and 3 for Test A2. However, all these subgroups are rather small. So, from our earlier discussion on β_{sg} and β_{sxg}, we would expect β_{sxg} to be small. However, ethnic group 1 is weaker than the other ethnic subgroups on all the subareas for Test A1 – so we would expect β_{sg} to be substantial for this subgroup.

Tables 2 and 3 show the results for Tests A1 and A2. The total test reliability for the full sample is .95 and .94, respectively, for the two tests. 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 subgroups. For each subgroup, the values of PRMSE$_{sg}$, PRMSE$_{sxg}$, β_{sg},

[4] 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.

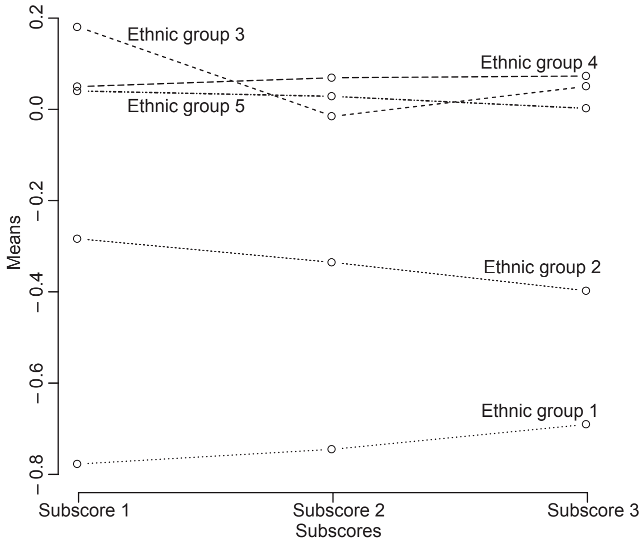

Figure 1. Standardized subgroup means for the first form of Test A1.

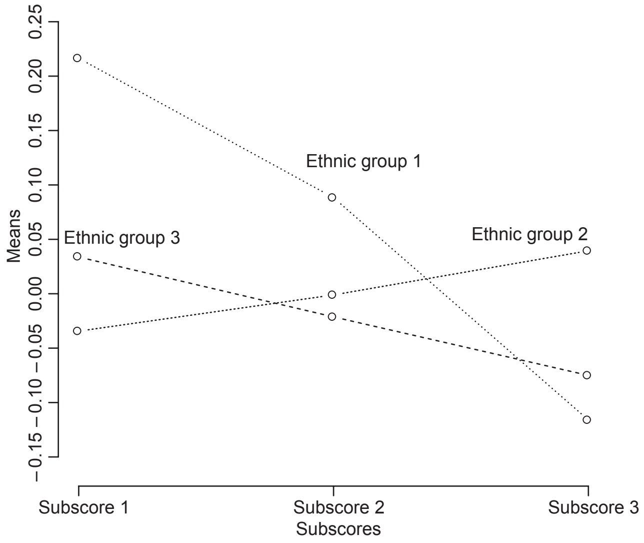

Figure 2. Standardized subgroup means for the first form of Test A2.

β_{sxg}, PRMSE$_{sg^*}$, and PRMSE$_{sxg^*}$ are shown in Tables 2 and 3. The sizes of the subgroups are also shown.

For Test A1, only the first subscore has added value for the full sample and the added value is very small. For Test A2, the first and third subscores have added value, by a substantial margin, for the full sample. Augmented subscores have added value for both Tests A1 and A2 for the full sample.

In addition, Tables 2 and 3 show that for most subgroups, especially the large ones, PRMSE$_{sg}$ is close to PRMSE$_{sg^*}$ and PRMSE$_{sxg}$ is close to PRMSE$_{sxg^*}$. There are

Table 2. Results for Test A1

	Subscore		
	1	2	3
Length →	82	88	35
All examinees (size: 4241)			
Correlation	1.00	0.80	0.77
	0.90	1.00	0.75
	0.91	**0.90**	1.00
PRMSE$_s$	**0.91**	0.88	0.78
PRMSE$_x$	0.90	0.89	0.87
PRMSE$_{sx}$	0.93	0.92	0.89
Ethnic Group 1 (size: 187)			
PRMSE$_{sg}$	0.91	0.89	0.75
PRMSE$_{sxg}$	0.94	0.93	0.89
β_{sg}	0.07	0.09	0.18
β_{sxg}	0.04	0.03	0.04
PRMSE$_{sg^*}$	0.91	0.88	0.71
PRMSE$_{sxg^*}$	0.94	0.93	0.88
Ethnic Group 2 (size: 229)			
PRMSE$_{sg}$	0.90	0.88	0.76
PRMSE$_{sxg}$	0.92	0.91	0.90
β_{sg}	0.03	0.04	0.10
β_{sxg}	−0.01	0.02	0.10
PRMSE$_{sg^*}$	0.90	0.88	0.75
PRMSE$_{sxg^*}$	0.92	0.91	0.88
Ethnic Group 3 (size: 207)			
PRMSE$_{sg}$	0.90	0.87	0.78
PRMSE$_{sxg}$	0.93	0.91	0.88
β_{sg}	−0.02	0.00	−0.01
β_{sxg}	−0.06	0.06	0.02
PRMSE$_{sg^*}$	0.90	0.87	0.78
PRMSE$_{sxg^*}$	0.92	0.91	0.88
Ethnic Group 4 (size: 2,845)			
PRMSE$_{sg}$	0.90	0.87	0.77
PRMSE$_{sxg}$	0.93	0.91	0.88
β_{sg}	0.00	−0.01	−0.02
β_{sxg}	0.00	−0.01	−0.02
PRMSE$_{sg^*}$	0.90	0.87	0.77
PRMSE$_{sxg^*}$	0.93	0.91	0.88
Ethnic Group 5 (sizes: 773)			
PRMSE$_{sg}$	0.92	0.89	0.78
PRMSE$_{sxg}$	0.94	0.92	0.89
β_{sg}	0.00	0.00	0.00
β_{sxg}	−0.01	0.00	0.02
PRMSE$_{sg^*}$	0.92	0.89	0.78
PRMSE$_{sxg^*}$	0.93	0.92	0.89
PRMSE$_{s+}$	0.91	0.88	0.78
PRMSE$_{sx+}$	0.93	0.92	0.89

Table 3. Results for Test A2

	Subscore		
	1	2	3
Length →	67	67	66
All examinees (size: 1931)			
Correlation	1.00	0.72	0.60
	0.82	1.00	0.76
	0.68	**0.88**	1.00
PRMSE$_s$	**0.89**	0.85	**0.87**
PRMSE$_x$	0.78	0.89	0.79
PRMSE$_{sx}$	0.91	0.91	0.89
Ethnic Group 1 (size: 133)			
PRMSE$_{sg}$	0.88	0.83	0.87
PRMSE$_{sxg}$	0.90	0.92	0.89
β_{sg}	−0.03	−0.02	0.02
β_{sxg}	−0.05	−0.02	0.07
PRMSE$_{sg^*}$	0.88	0.83	0.87
PRMSE$_{sxg^*}$	0.90	0.91	0.89
Ethnic Group 2 (size: 1,313)			
PRMSE$_{sg}$	0.90	0.84	0.86
PRMSE$_{sxg}$	0.91	0.90	0.89
β_{sg}	0.00	0.00	−0.01
β_{sxg}	0.01	0.00	−0.02
PRMSE$_{sg^*}$	0.90	0.84	0.86
PRMSE$_{sxg^*}$	0.91	0.90	0.89
Ethnic Group 3 (size: 485)			
PRMSE$_{sg}$	0.89	0.85	0.87
PRMSE$_{sxg}$	0.90	0.90	0.89
β_{sg}	0.00	0.00	0.01
β_{sxg}	−0.02	0.00	0.02
PRMSE$_{sg^*}$	0.89	0.85	0.87
PRMSE$_{sxg^*}$	0.90	0.90	0.89
PRMSE$_{s+}$	0.89	0.85	0.87
PRMSE$_{sx+}$	0.91	0.91	0.89

some subgroups, mostly small ones, for which β_{sg} is substantially different from zero and, as a result, PRMSE$_{sg}$ is somewhat larger than PRMSE$_{sg^*}$. For example, for ethnic group 1, $\beta_{sg} = 0.18$ for the third subscore for Test A1. This result is expected from Figure 1, for the profiles of ethnic group 1 lie well below those of the other subgroups. As a result, the difference between PRMSE$_{sg}$ and PRMSE$_{sg^*}$ is 0.04, which is substantial. Because the subscore profiles of the subgroups are mostly parallel (see Figure 1), β_{sxg} is not far from zero for the test. As a result, the difference between PRMSE$_{sxg}$ and PRMSE$_{sxg^*}$ is only 0.01, which is not substantial. These numbers exemplify the case in which the invariance of the augmented subscore is easier to achieve than the invariance of the subscores.

The differences are small between PRMSE$_s$ and PRMSE$_{s+}$ and between PRMSE$_{sx}$ and PRMSE$_{sx+}$. Therefore, knowledge of examinee subgroups does not lead to appreciably better estimation of true subscores.

5.2. Test B

Figures 3 and 4 show the standardized subscore means of the subgroups for Test B. In Figure 4, ethnic group 2 exhibits a notable departure from the parallel pattern, but the sample size is rather small. In Figure 3, the profile for ethnic group 1 lies lower than that of all the other subgroups.

Tables 4 and 5 show the PRMSEs for Tests B1 and B2. The total test reliability for the full sample is .92 and .94, respectively, for the two tests. For the full sample, the second

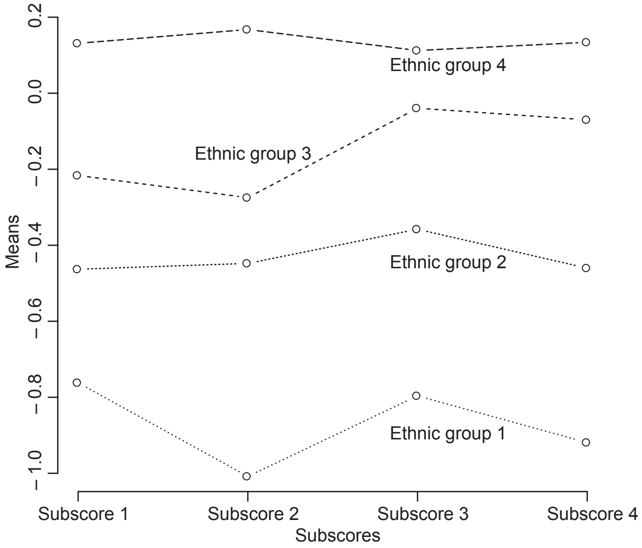

Figure 3. Standardized subgroup means for Test B1.

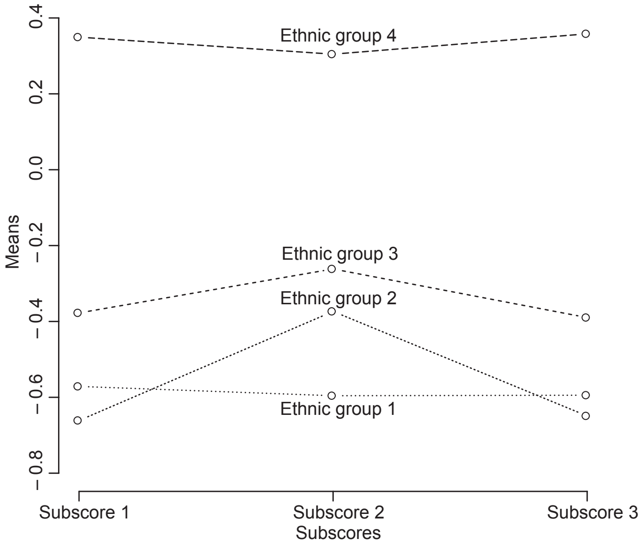

Figure 4. Standardized subgroup means for Test B2.

Table 4. Results for Test B1

	Subscore			
	1	2	3	4
Number of items →	30	30	30	30
All examinees (size: 6,641)				
Correlation	1.00	0.57	0.52	0.55
	0.77	1.00	0.55	0.63
	0.76	**0.68**	1.00	0.69
	0.78	**0.77**	**0.90**	1.00
PRMSE$_{s}$	0.63	**0.86**	0.75	0.78
PRMSE$_{x}$	0.72	0.75	0.76	0.82
PRMSE$_{sx}$	0.78	0.88	0.83	0.86
Ethnic Group 1 (size: 642)				
PRMSE$_{sg}$	0.63	0.79	0.64	0.74
PRMSE$_{sxg}$	0.72	0.82	0.76	0.80
β_{sg}	0.33	0.18	0.28	0.24
β_{sxg}	0.06	0.12	0.04	0.08
PRMSE$_{sg^*}$	0.52	0.75	0.55	0.68
PRMSE$_{sxg^*}$	0.70	0.80	0.74	0.79
Ethnic Group 2 (size: 156)				
PRMSE$_{sg}$	0.67	0.85	0.75	0.76
PRMSE$_{sxg}$	0.79	0.87	0.81	0.87
β_{sg}	0.18	0.07	0.10	0.11
β_{sxg}	0.10	0.03	0.00	0.05
PRMSE$_{sg^*}$	0.64	0.85	0.74	0.75
PRMSE$_{sxg^*}$	0.77	0.86	0.80	0.86
Ethnic Group 3 (size: 5,251)				
PRMSE$_{sg}$	0.56	0.83	0.73	0.75
PRMSE$_{sxg}$	0.74	0.85	0.80	0.83
β_{sg}	−0.07	−0.03	−0.03	−0.04
β_{sxg}	−0.03	−0.02	0.00	−0.01
PRMSE$_{sg^*}$	0.55	0.83	0.73	0.75
PRMSE$_{sxg^*}$	0.73	0.85	0.80	0.83
Ethnic Group 4 (size: 592)				
PRMSE$_{sg}$	0.73	0.90	0.82	0.81
PRMSE$_{sxg}$	0.83	0.92	0.88	0.89
β_{sg}	0.08	0.03	0.01	0.02
β_{sxg}	0.07	0.05	−0.05	−0.05
PRMSE$_{sg^*}$	0.71	0.90	0.81	0.81
PRMSE$_{sxg^*}$	0.82	0.91	0.87	0.89
PRMSE$_{s+}$	0.65	0.86	0.76	0.79
PRMSE$_{sx+}$	0.78	0.88	0.83	0.86

subscore of Test B1 has added value and none of the three subscores of Test B2 has added value. Augmented subscores have added value for both Tests B1 and B2 for the full sample.

There are some subgroups for which β_{sg} is substantially different from zero and, as a result, PRMSE$_{sg}$ is somewhat larger than PRMSE$_{sg^*}$. For example, for the four subscores for ethnic group 1 for Test B1, β_{sg}s are between 0.18 and 0.33 and PRMSE$_{sg}$ is larger than the

Table 5. Results for Test B2

	Subscore		
	1	2	3
Number of items →	25	25	25
All examinees (size: 5,134)			
Correlation	1.00	0.74	0.78
	0.86	1.00	0.74
	0.92	**0.89**	1.00
PRMSE$_s$	0.87	0.84	0.82
PRMSE$_x$	0.87	0.85	0.89
PRMSE$_{sx}$	0.91	0.89	0.90
Ethnic Group 1 (size: 1,172)			
PRMSE$_{sg}$	0.84	0.78	0.77
PRMSE$_{sxg}$	0.88	0.85	0.86
β_{sg}	0.08	0.12	0.13
β_{sxg}	0.02	0.06	0.05
PRMSE$_{sg^*}$	0.83	0.77	0.75
PRMSE$_{sxg^*}$	0.88	0.84	0.86
Ethnic Group 2 (size: 417)			
PRMSE$_{sg}$	0.87	0.81	0.81
PRMSE$_{sxg}$	0.90	0.86	0.87
β_{sg}	0.08	0.07	0.12
β_{sxg}	0.08	−0.08	0.11
PRMSE$_{sg^*}$	0.87	0.80	0.79
PRMSE$_{sxg^*}$	0.89	0.84	0.86
Ethnic Group 3 (size: 3,193)			
PRMSE$_{sg}$	0.83	0.83	0.77
PRMSE$_{sxg}$	0.88	0.88	0.87
β_{sg}	−0.06	−0.06	−0.09
β_{sxg}	−0.03	−0.01	−0.05
PRMSE$_{sg^*}$	0.82	0.82	0.76
PRMSE$_{sxg^*}$	0.88	0.88	0.87
Ethnic Group 4 (size: 452)			
PRMSE$_{sg}$	0.87	0.84	0.81
PRMSE$_{sxg}$	0.90	0.88	0.89
β_{sg}	0.05	0.04	0.08
β_{sxg}	0.04	−0.02	0.06
PRMSE$_{sg^*}$	0.87	0.84	0.80
PRMSE$_{sxg^*}$	0.90	0.88	0.89
PRMSE$_{s+}$	0.88	0.85	0.83
PRMSE$_{sx+}$	0.91	0.89	0.90

corresponding PRMSE$_{sg^*}$ by between 0.04 and 0.11. Note in Figure 3 that the profile of ethnic group 1 lies well below the profiles of the other subgroups. However, for all the subscores of Test B1, β_{sxg} is close to zero for ethnic group 1, and, as a result, PRMSE$_{sxg}$ is close to PRMSE$_{sxg^*}$.

The differences are small between PRMSE$_s$ and PRMSE$_{s+}$ and between PRMSE$_{sx}$ and PRMSE$_{sx+}$. Thus, information on examinee subgroups does not lead to better estimation of true subscores.

6. Conclusions

For the subgroups studied, knowledge of the subgroup membership of the examinees does not lead to a better estimation of the true subscore for the tests considered here. In other words, estimation of the true subscore is nearly invariant over the subgroups. This implies that while reporting, for example, the augmented subscores for any test considered in this paper,[5] one does not have to worry about the lack of invariance of the function providing the augmented subscores over subgroups. This is an interesting finding given that the subgroups were often quite different in their average performance on subscores for our data sets. For example, Figure 2 shows a big difference, roughly of one standard deviation, between ethnic group 4 and ethnic group 1 for Test A.

Augmented subscores, which estimate true subscores based on both the subscore and the total scores, were more likely to satisfy the invariance criterion compared to estimation based on only subscores or only total scores. In several cases, there was a substantial difference between PRMSE_{sg} and PRMSE_{sg^*} caused by a non-zero value of β_{sg}, but there was a negligible difference between PRMSE_{sxg} and PRMSE_{sxg^*}. The PRMSEs of augmented subscores are mostly quite high and hence there is not much room for further improvement from the inclusion of subgroup membership information. Together with the finding that the augmented subscores often lead to more accurate diagnostic information than subscores (e.g., Sinharay, 2010; Sinharay & Haberman, 2008), this finding regarding invariance makes the augmented subscores quite attractive to those interested in reporting diagnostic scores.

We reported the results for two operational tests that report subscores and involve subgroups. Similar results were obtained for data from three other testing programmes – an English proficiency test, a teacher licensing test, and a battery of tests that measure school and individual student progress.[6]

Although results in this report suggest that subgroup membership has little impact on approaches to reporting of subscores, there is no guarantee that this conclusion applies to all data. For example, the results may be different for a test for which a subgroup is better than another on some subtests but is worse on some other subtests. Therefore, it is a prudent practice to employ the methods we have described prior to a decision to report subscores to inform decisions and to verify impact on examinees of decisions on reporting practices. It is also appropriate to monitor testing programmes to verify periodically that earlier decisions on reporting practices remain appropriate.

There are several related issues that can be examined in further research. First, we analysed data from a few operational tests. One could examine data from more and preferably different types of tests. We reported results for subgroups based on ethnicity. Subgroups based on other factors such as income could be considered. We performed analyses involving the individual-level subscores. It is possible to perform similar analyses involving aggregate-level subscores, for example, average subscores of different schools, in the same way as in Haberman *et al.* (2009). However, those analyses would involve a different sets of techniques that are possible topics for further research.

[5] Augmented subscores were found to have added value for these tests.

[6] Results for these tests are not shown and are available on request from the authors.

Acknowledgements

The authors thank Neil J. Dorans, Skip Livingston, Frank Rijmen, Matthias von Davier, Thom Baguley, and the anonymous reviewers for their advice. The authors gratefully acknowledge the help of Kim Fryer with proofreading. Any opinions expressed in this publication are those of the authors and not necessarily of ETS.

References

Dorans, N. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41*, 43–68. doi:10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J., & Holland P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306. doi:10.1111/j.1745-3984.2000.tb01088.x

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229. doi:10.3102/1076998607302636

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95. doi:10.1348/000711007X248875

Ling, G. (2009, April). *Why the major field (business) test does not report subscores of individual test-takers – reliability and construct validity evidence*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Livingston, S. A., & Rupp, S. L. (2004). *Performance of men and women on multiple choice and constructed-response tests for beginning teachers* (ETS Research Report No. 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. doi:10.1080/08957347.2010.486287

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174. doi:10.1111/j.1745-3984.2010.00106.x

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

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28. doi:10.1111/j.1745-3992.2007.00105.x

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

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140. doi:10.1111/j.1745-3984.2000.tb01079.x

Received 28 February 2012; revised version received 11 July 2012

Appendix A: Proof of the relationship expressed in equation (1)

The minimum of

$$E\big([\alpha + \beta s - s_t]^2 | G = g\big)$$

with respect to α and β is achieved by $\alpha_{\min} = \mu_{sg}(1 - \rho_{sg}^2)$ and $\beta_{\min} = \rho_{sg}^2$. If

$$f_{sg} = \alpha_{\min} + \beta_{\min} s = \mu_{sg} + \rho_{sg}^2(s - \mu_{sg}),$$

then

$$E(f_{sg} - s_t | G = g) = 0 \tag{A1}$$

and

$$E(s(f_{sg} - s_t)|G = g) = 0. \tag{A2}$$

The mean square error for subgroup g of s_s, MSE_{g^*}, is an estimate of

$$\begin{aligned} &E\big(\big[E(s) + \rho_s^2[s - E(s)] - s_t\big]^2 | G = g\big) \\ &= E(\{E(s) + \rho_s^2[s - E(s)] - f_{sg} + f_{sg} - s_t\}^2 | G = g), \end{aligned} \tag{A3}$$

$$= E(\{E(s) + \rho_s^2[s - E(s)] - f_{sg}\}^2 | G = g) + E((f_{sg} - s_t)^2 | G = g). \tag{A4}$$

This is because equations (A1) and (A2) imply that the cross-product term in the expansion of equation (A3),

$$\begin{aligned} &2E\big(\{E(s) + \rho_s^2[s - E(s)] - f_{sg})\}\{f_{sg} - s_t\} | G = g\big) \\ &= 2E\big(\big[E(s)(f_{sg} - s_t)\big] | G = g\big) + 2E\big(\big[\rho_s^2 s(f_{sg} - s_t)\big] | G = g\big) \\ &- 2E\big(\big[\rho_s^2 E(s)(f_{sg} - s_t)\big] | G = g\big) - 2E\big(\big[f_{sg}(f_{sg} - s_t)\big] | G = g\big), \end{aligned}$$

is equal to 0.

Note that f_{sg} is estimated by s_{sg} with the corresponding mean square error MSE_{sg}. Therefore, the second term of equation (A4) is estimated by MSE_{sg}. Because V_{sg} estimates $\text{Var}(s \mid G = g)$, $\bar{s}$ estimates $E(s)$, $\bar{s}_g$ estimates μ_{sg}, and $B_{sg} = -(1 - \hat{\rho}_s^2)(\bar{s}_g - \bar{s})$, the first term of equation (A4) is equal to

$$\begin{aligned} &E\left(\left\{(s - \mu_{sg})(\rho_s^2 - \rho_{sg}^2) + (E(s) - \mu_{sg})(1 - \rho_s^2)\right\}^2 | G = g\right) \\ &= (\rho_s^2 - \rho_{sg}^2)^2 \text{Var}(s | G = g) + (1 - \rho_s^2)^2 [\mu_{sg} - E(s)]^2, \end{aligned}$$

and is estimated by

$$(\hat{\rho}_s^2 - \hat{\rho}_{sg}^2)^2 V_{sg} + B_{sg}^2.$$

Author Query Form

Journal: BMSP
Article: 2061

Dear Author,
During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.
Many thanks for your assistance.

Query reference	Query	Remarks
1	**AUTHOR: Please check whether the insertion of opening brace in equation A1 ok.**	
2	**AUTHOR: Please provide complete affiliation details.**	
3	**AUTHOR: Please provide DOI number for all journal type references.**	