

An Empirical Bayes Approach to Subscore Augmentation: How Much Strength Can We Borrow?

Michael C. Edwards
The Ohio State University

Jack L. Vevea
The University of California at Santa Cruz

This article examines a subscore augmentation procedure. The approach uses empirical Bayes adjustments and is intended to improve the overall accuracy of measurement when information is scant. Simulations examined the impact of the method on subscale scores in a variety of realistic conditions. The authors focused on two popular scoring methods: summed scores and item response theory scale scores for summed scores. Simulation conditions included number of subscales, length (hence, reliability) of subscales, and the underlying correlations between scales. To examine the relative performance of the augmented scales, the authors computed root mean square error, reliability, percentage correctly identified as falling within specific proficiency ranges, and the percentage of simulated individuals for whom the augmented score was closer to the true score than was the nonaugmented score. The general findings and limitations of the study are discussed and areas for future research are suggested.

Keywords: ability estimation, empirical Bayes, item response theory, subscore augmentation

As the popularity of testing grows in the realm of education, tests that were originally designed for one purpose are frequently being pressed into service for others. For example, tests designed to produce reliable scores for ranking individuals may also be expected to provide diagnostic information to interested parties (e.g., the test taker, parents, teachers, etc.). It is difficult to construct a test that can achieve both of these goals adequately.

A potential solution to this challenge involves the reporting of subscores. While the overall scores are used to rank individuals, scores on smaller subsections of the test are used to provide feedback specific to a narrow content area. This approach, although intuitively appealing, has a drawback. A test may be constructed so that the overall score is reliable, but such a test may not provide reliable subscores, which are based on less information than the overall score.

This work was supported by the North Carolina Department of Public Instruction. The authors thank Howard Wainer and David Thissen for their comments on earlier drafts of this article.

There are three ways we could address this problem. First, we could increase the length of each subscale until it provides reliable scores. This solution is unrealistic given time and length constraints often present in large-scale educational testing. Another option is to use adaptive technologies, in the form of computerized adaptive testing (CAT). Such a strategy enables tailored testing to achieve reliable measurement on a subject-by-subject basis. The realities of educational testing make this solution impractical under many circumstances (see Wainer, 2000). The third possibility is to increase the reliability of diagnostic subscores by incorporating information from the rest of the test. This final procedure is the focus of the current article.

Multivariate Empirical Bayes Estimation

Nearly 50 years after Kelley's (1927) seminal work, what are now called empirical Bayes (EB) methods resurfaced in psychology in the form of multivariate generalizability theory in Cronbach, Gleser, Nanda, and Rajaratnam's *The Dependability of Behavioral Measurements* (1972). Equation 10.3 in chapter 10 of that book is a multivariate form of Kelley's regressed estimates. Although the language is slightly different (Cronbach et al. discuss "universe scores" instead of true scores), the underlying theory is the same. The authors note that using all the information in the "profile" will produce a superior estimate of the subscale true score whenever the subscales are correlated. They also point out that the new, augmented estimates will always have smaller error variance than the nonaugmented estimates.

Kelley's (1927) univariate regression of true score on observed score can be rewritten for the multivariate case as

$$\hat{\tau} = \mathbf{x} + \mathbf{B}(\mathbf{x} - \mathbf{x}),$$

where \mathbf{x} is a vector of subscale means, \mathbf{x} is a vector of subscale scores, and \mathbf{B} is a matrix of reliability-based weights. The least-squares estimate of \mathbf{B} is the product of the true score covariance matrix and the inverse of the observed score covariance matrix,

$$\mathbf{B} = \Sigma_t \Sigma_x^{-1},$$

where Σ_t is the true score covariance matrix and Σ_x is the observed score covariance matrix. Given data, both of these covariance matrices may be estimated: Σ_t by S_t , and Σ_x by S_x . S_x is an estimate of the observed score covariance matrix and is easily obtained from the data. As shown by Wainer et al. (2001), S_t can be computed from the estimated covariance matrix of the observed scores using

$$s_{vv'}^t = s_{vv'}^x \text{ for } v \neq v',$$

and

$$s_{vv'}^t = \rho_v s_{vv'}^x \text{ for } v = v',$$

where $s_{vv'}^t$ is the vv' element in the covariance matrix \mathbf{S}_t , $s_{vv'}^x$ is the corresponding element in the observed covariance matrix (\mathbf{S}_x), and ρ_v is the reliability of subscore v .

Using Subscore Augmentation

The multivariate augmentation procedure can be used with observed scores (Vevea, Billeaud, & Nelson, 1998), item response theory (IRT) scale scores for response patterns, and IRT scale scores for summed scores. Differences in the procedure for each are largely computational; the interested reader is directed to Wainer et al.'s (2001) treatment of this issue. The current article focuses on the two summed-score-based procedures: observed scores and IRT scale scores for summed scores.

Although IRT scale scores for summed scores do not incorporate as much information as scale scores for response patterns, they preserve IRT's nonlinear relationship between number correct and proficiency while eliminating the awkwardness of a test on which equal summed scores may map to different proficiency estimates. In addition, IRT scale scores are comparable across different forms of a test, so they are widely used in large-scale testing with multiple forms.

The need for subscale augmentation arose when the North Carolina Department of Public Instruction (NCDPI) decided to use its end-of-grade (EOG) tests to provide diagnostic subscores in curricular areas. The EOG tests are primarily designed to allow reliable rankings of individuals; they were not designed to provide diagnostic subscores. When the decision was made to use them for diagnosis, problems arose because of the relatively small size, and hence low reliability, of some of the subscales.

The goal of the current article is to demonstrate the utility of using subscore augmentation under such circumstances. The simulation study described below examines the method implemented with the summed-score IRT approach used in the NCDPI context. In addition, we present parallel simulations using multivariate EB methods for simple summed scores. These latter results may be of more interest to persons considering the use of EB methods in small-scale contexts where the number of respondents cannot support reliable IRT estimates of item characteristics.

Method

We conducted simulations to examine the effect of augmenting summed scores and IRT scale scores for summed scores. The general simulation design elements are the same for the two types of scores.

Simulation Design

The goal of the current study is to assess the performance of subscore augmentation at various levels of subscale number, subscale length (reliability), and subscale correlation. Although it is known in advance that with an appropriate weighting

scheme the empirical Bayes estimates will outperform scores that have not been adjusted, the extent and practical significance of the difference has yet to be determined.

The design includes tests with two or four subscales. Items were simulated using the same item parameter distributions as Wainer, Bradlow, and Du (2000), which are based on marginal distributions from SAT data. Using standard notation for the 3PL IRT model, the distributions are: $a \sim N(0.8, 0.2^2)$, $b \sim N(0, 1)$, and $c \sim N(0.2, 0.03^2)$. We truncated the distributions at three standard deviations above and below the mean to avoid the unusual parameter values discussed by Harwell, Stone, Hsu, and Kirisci (1996). We sampled individual ability parameters (thetas) from a standard normal distribution.

One of the most important aspects of the empirical Bayes estimation method is the correlation among the subscales. We chose three levels of correlation to reflect relatively uncorrelated subscales ($r = 0.3$), moderately correlated subscales ($r = 0.6$), and highly correlated subscales ($r = 0.9$).

In this simulation, average subscale reliability is a function of number of items on a subscale. We chose four subscale lengths so that two of them (5 and 10 items) are relatively unreliable (0.43 and 0.59, respectively), and two of them (20 and 40 items) are relatively reliable (0.75 and 0.85, respectively). These reliability estimates are the average observed squared correlations between scale scores and generating ability parameters (thetas) over 100 replications. The 5-, 10-, and 20-item subscales have been seen in actual work with the NCDPI, which is a principal reason for their inclusion. We included the 40-item subscale as a ceiling condition.

The levels within conditions chosen for the two-subscale case result in a total of 30 cells. Crossing each combination of subscale item counts (5×5 , 5×10 , etc.) with each correlation results in 48 possible combinations. Of those combinations 18 are redundant (e.g., for two 5-item subscales, Subscale 1 correlated 0.3 with Subscale 2 is the same condition as Subscale 2 correlated 0.3 with Subscale 1), so only 30 of the 48 possible cells provide unique information. In the four-subscale condition, nine different combinations of subscale length were examined at the three correlation levels, resulting in a total of 27 cells. Table 1 shows the details of how these design factors are crossed.

One hundred replications were used in every cell throughout the simulation because that represents a good balance between data yield and computational time. The sample size for each cell was 2,000, a number that was large enough to yield stable IRT parameter estimates and small enough to converge relatively quickly.

Analytic Strategy

To evaluate the relative performance of augmented versus nonaugmented scores we focused on four measures. Root mean squared error (RMSE) and reliability are reasonable ways to compare the relative performance of scores. RMSE is the square root of the average squared difference between estimated scores and true scores. In the case of IRT simulations, true scores are defined as the simulated individual thetas. In the summed-score analyses, true scores are the expected summed scores, given

TABLE 1
Simulation Design

Length of Subscale A	Two-Subscale Design				Four-Subscale Design		
	Length of Subscale B				Length of Subscale B, C, & D		
	5	10	20	40	5,5,5	10,10,10	20,20,20
5	0.3				0.3	0.3	0.3
	0.6				0.6	0.6	0.6
	0.9				0.9	0.9	0.9
10	0.3	0.3			0.3	0.3	0.3
	0.6	0.6			0.6	0.6	0.6
	0.9	0.9			0.9	0.9	0.9
20	0.3	0.3	0.3		0.3	0.3	0.3
	0.6	0.6	0.6		0.6	0.6	0.6
	0.9	0.9	0.9		0.9	0.9	0.9
40	0.3	0.3	0.3	0.3			
	0.6	0.6	0.6	0.6			
	0.9	0.9	0.9	0.9			

Note: Entries in the table indicate level of correlation between subscales at which the simulation was performed in each cell.

the simulated individual thetas and the generating values of the IRT parameters. That is, for an individual with ability equal to θ_j , the expected summed score is

$$E(X_j) = \sum_{i=1}^{\text{items}} \left\{ c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_j - b_i)]} \right\}.$$

We computed reliability as the square of the correlation between true and estimated scores.

Users of these methods will not know the true scores and, hence, will be unable to compute reliability in this manner, so we also examined estimates of reliability for each type of score. In the case of the EB augmented scores, we used a new procedure to produce a reliability estimate. Wainer et al. (2001) computed reliability for the augmented scores as

$$r_v^2 = \frac{a_{vv}}{c_{vv}},$$

where the numerator is an approximation of the v th subscale's unconditional true score variance, and the denominator is an estimate of unconditional score variance of the estimates for the v th subscale. The numerator is taken from the diagonal of the matrix

$$\mathbf{A} = \mathbf{S}_t (\mathbf{S}_x)^{-1} \mathbf{S}_t (\mathbf{S}_x)^{-1} \mathbf{S}_t,$$

and the denominator is taken from the diagonal of the matrix

$$\mathbf{C} = \mathbf{S}_t (\mathbf{S}_x)^{-1} \mathbf{S}_t.$$

Research conducted since the publication of *Test Scoring* has shown this estimate of reliability for augmented scores to be positively biased (Edwards, 2002). When estimating reliability for augmented scores we used (and recommend)

$$r_v^2 = 1 - \frac{a_{vv}^*}{s_{vv}^t},$$

taking the numerator of the fraction from

$$\mathbf{A}^* = \mathbf{S}_t - \mathbf{S}_t \mathbf{S}_x^{-1} \mathbf{S}_t,$$

and the denominator from the estimated true score matrix.

In addition to RMSE and reliability, we computed the percentage of simulated individuals for whom the augmented score was closer to the true score than was the nonaugmented score. This percentage adds to our knowledge of the relative accuracy of the various scores, especially for situations that involve classification.

One final measure of performance involved accuracy in assigning individuals to groups based on their scores. Many tests in education and certification contexts make use of cut scores to determine levels of proficiency. In some cases, this leads to pass–fail decisions for schoolchildren; in other cases, it leads to a placement in some achievement level or identifies a need for remediation. In all cases, it is of interest to compare the performance of augmented subscores in these decisions with the performance of their nonaugmented counterparts.

The analyses of percentage correctly identified focused on a smaller subset of the cells for the two-subscale portion of the simulation. The notation $A \times B$ denotes the augmentation of a subscale with A items using information from a subscale with B items; for example, 5×40 denotes the case where a 5-item subscale has been improved by EB augmentation using a 40-item subscale. Four different combinations of subscales (5×40 , 5×5 , 20×20 , and 40×5) were examined at the three correlation levels (0.3, 0.6, and 0.9) used throughout the study. For the four-subscale case, we examined three different combinations of subscales ($5 \times 20 \times 20 \times 20$, $5 \times 5 \times 5 \times 5$, and $20 \times 5 \times 5 \times 5$) at the three correlation levels mentioned above.

We conducted this analysis by dividing the simulees into four “ability groupings” based on their true and estimated scale scores (*expected a posteriori* scale scores, or EAPs, were used as IRT scale score estimates throughout this study). We selected three cutscores (–1.96, 0, 1.036) to place 2.5% of simulees in the lowest category, 47.5% in the 2nd category, 35% in the 3rd category, and 15% in the highest category. For analyses of this issue in the context of traditional summed scores (rather than summed-score EAPs), we set cutscores by computing the expected summed scores of individuals whose thetas were –1.96, 0, or 1.036, using the simulated IRT parameters to compute the expectation.

Results

IRT Scale Scores for Summed Scores

In the course of analyzing the results it became apparent that the two-subscale simulations would be sufficient to convey the general trends observed throughout the simulation. In the interest of brevity we do not present the four-subscale results here, although they are available from the authors upon request.

Item Parameter Estimation

All of the augmentation procedures discussed in this article were implemented using the AUGMENT software¹ (Vevea et al., 2002). In addition to performing the various augmentation procedures, AUGMENT estimates item parameters using the Bock and Aitkin (1981) EM algorithm.

RMSE

As expected, in all cases the augmented EAPs had lower RMSE than the non-augmented EAPs. Figure 1 shows the overall trends in the difference in RMSE between the nonaugmented and augmented EAPs. The four plots in Figure 1 highlight the effect of augmentation on a subscale as a function of the correlation between subscales and the length of the subscale contributing ancillary information.

The top left panel and lower right panel of Figure 1 deserve special scrutiny because these are the conditions under which we expect to see the most and the least improvement, respectively, from the augmentation procedure. We see from the top left panel of Figure 1 that when the subscale being augmented is small (5 items), the number of items in the second subscale is large (40 items), and the correlation between the two subscales is high ($r = 0.9$) the RMSE of the augmented EAPs (0.507) is approximately 33% smaller than the RMSE of the nonaugmented EAPs (0.757).

The lower right panel of Figure 1 shows that when the subscale being augmented is large (40 items), the number of items in the second subscale is small (5 items), and the correlation between the two subscales is small ($r = 0.3$), then the RMSE of the augmented EAPs (0.386) is not appreciably lower than the RMSE of the non-augmented EAPs (0.387). This is not a surprising finding, but it highlights the fact that even in the case in which we expect the least improvement from the augmentation procedure we still observe *some* improvement.

It is also interesting to assess the benefit of employing the augmentation procedure in less extreme cases. Two such “less extreme cases” are the 10×10 and 20×20 conditions, where the correlation between the subscales is 0.6. In the 10×10 case, we see an approximate 5% decrease in RMSE when using the augmented subscores. In the 20×20 case, the RMSE of the augmented subscores is roughly 4.5% smaller. The practical significance of this difference depends largely on the use of the subscores.

Reliability

For reliability, we were able to examine the “true” reliability (the squared correlation between a theta estimate and the generating theta) as well as the two

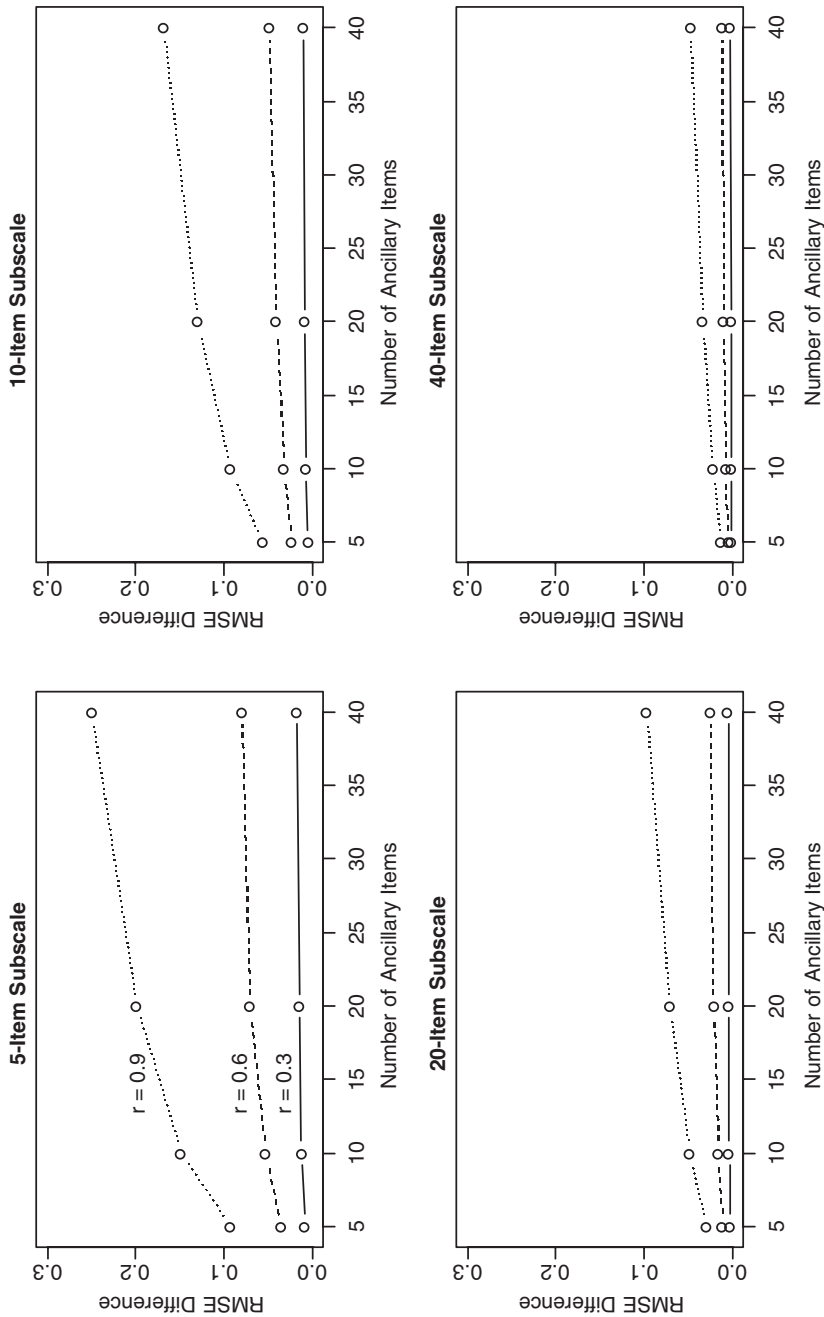


FIGURE 1. Difference in RMSE between nonaugmented and augmented IRT scale scores for summed scores as a function of correlation between subscales and number of ancillary items. The solid line, for example, indicates the difference between the RMSE for nonaugmented scores and augmented scores when the subscales are correlated 0.3.

estimates of reliability employed in this portion of the study. The two estimates of reliability, IRT marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984) and the augmented score reliability, both performed very well in this simulation. On average, over the 100 replications in each cell, the estimated reliability was never more than 0.01 away from the true reliability. Given the accuracy of reliability estimation for both augmented and nonaugmented scores, we focus on the differences in “true” reliability between the scores.

The augmented scores exhibited higher true reliability in all cases, but these differences were not always of practical significance. Figure 2 displays the reliabilities for the various configurations examined in the two-subscale portion of this simulation. An examination of Figure 2 reveals several general trends. The size of the difference in reliability between the augmented and nonaugmented scores depends on the size of the subscale being augmented, the size of the subscale used to provide ancillary information, and the correlation between the two subscales. As with RMSE, the greatest advantage when using the augmented scores was seen when the subscale being augmented was small (5 items), the subscale providing ancillary information was large (40 items), and the correlation between the two subscales was high ($r = 0.9$). In this extreme case, the reliability of the augmented scores was more than 1.5 times that of the nonaugmented scores.

Percentage of Simulees for Whom Augmented Scores Were More Accurate

To provide additional information about the performance of the augmented subscores, we computed the percentage of simulees for whom augmented subscores were more accurate (i.e., the distance between the augmented estimate and the generating value was smaller than the distance between the nonaugmented estimate and the generating value). The results, summarized in Table 2, show that, in all cases examined in the current study, the augmented scores were more accurate for a larger percentage of the simulees than the nonaugmented scores. The size of this difference ranges from 1% to 16%, with an average over all conditions of about 5%. These findings provide evidence that the gains in RMSE previously discussed are not attributable to a few extreme scores, but reflect an accurate picture of a general pattern of improvement seen when using augmented subscores.

Percentage Correctly Identified

The percentage correctly identified analyses focused on a smaller subset of the cells run for the two-subscale portion of the simulation. Four different combinations of subscales (40×5 , 20×20 , 5×5 , and 5×40) were examined at the three correlation levels (0.3, 0.6, and 0.9) used throughout the study. The results are summarized in Table 3.

As expected, the augmented subscores placed a higher percentage of simulees in the correct ability group in all cases analyzed. The magnitude of this difference ranges from small (0.03%) to quite large (13.43%). Although it is extremely unlikely that anyone would ever try to place individuals into four ability groups based on five items, it is interesting to see how the augmentation procedure performs in the worst- and best-case scenarios. A more realistic situation is the 20×20 subscale

