Measuring the Power of Learning.™

Research Report
ETS RR–18-01

A Comparison of Score Aggregation Methods for Unidimensional Tests on Different Dimensions

Jianbin Fu

Yuling Feng

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Comparison of Score Aggregation Methods for Unidimensional Tests on Different Dimensions

Jianbin Fu[1] & Yuling Feng[2]

1 Educational Testing Service, Princeton, NJ
2 University of South Carolina, Columbia, SC

In this study, we propose aggregating test scores with unidimensional within-test structure and multidimensional across-test structure based on a 2-level, 1-factor model. In particular, we compare 6 score aggregation methods: average of standardized test raw scores (M1), regression factor score estimate of the 1-factor model based on the correlation matrix of test raw scores (M2), overall ability from a unidimensional generalized partial credit model (GPCM) based on the items from all tests (M3), average of ability estimates from individual tests based on GPCM (M4), regression factor score of the 1-factor model based on the correlation matrix of ability estimates from individual tests based on GPCM (M5), and general ability from the testlet model (M6). The 4 design factors considered in the simulation study are ability correlation between tests (.3, .5, .7, .8, and .9), test length (10, 20, 30, and 60 items), number of tests (2 and 4), and factor loading distribution (equal and unequal). The comparisons are also conducted on a real test data set with 2 tests. On the basis of the results, M1 and M4 are recommended for 2 tests, and M2, M5, and M6 are recommended for 3 or more tests. Several issues regarding attaining aggregate score reliability for intended uses and score aggregation types distinguished by test dimensionality are discussed, and practical suggestions for score aggregation are provided.

Keywords Score aggregation; factor analysis; item response theory; testlet model

doi:10.1002/ets2.12194

In practice, several tests may be administered to a group of test takers on different occasions during a time period, and these test scores may need to be aggregated to obtain a final score as an indicator of a test taker's overall performance. For example, in K–12 testing, several assessments may be administered to test takers over the course of a school year, and information from these tests might then be aggregated to evaluate students' overall academic performance for the purpose of school, teacher, and student accountability.

Multiple test administrations during a time period may be needed for various reasons. For example, equivalent test forms may be administered on different occasions to measure students' growth and/or increase test reliability. Alternatively, tests measuring related constructs may be administered right after each construct is taught to provide a timely measure of the extent of students' understanding of the construct. Combining the results of multiple assessments during a school year, for accountability purposes, is a key component of the *CBAL*® learning and assessment tool research initiative supported by Educational Testing Service (ETS). Combining results from more than one testing occasion is also a key feature of the tests developed within the Race to the Top Assessment Program by the two state consortia, SMARTER Balanced Assessment Consortium and Partnership for Assessment of Readiness for College and Careers.

In this study, we compared different methods for aggregating scores from tests with unidimensional within-test structure and multidimensional across-test structure. In the first section, the statistical methods for this type of score aggregation are introduced based on the two-level, one-factor model. In the second section, the simulation study design and results are described. The third section presents the comparisons of the score aggregation methods based on a sample of actual CBAL writing data. Finally, in the fourth section, results are summarized, several issues regarding attaining aggregate score reliability for intended uses and score aggregation types distinguished by test dimensionality are discussed, and practical suggestions for score aggregation are provided.

Corresponding author: J. Fu, E-mail: jfu@ets.org

Statistical Methods for Score Aggregation

Let us assume there are K tests and that test k has I_k test takers and J_k items. Denote X_{ijk} as test taker i's item score on item j on test k with possible integer scores from 0 to this item's max score, $M_{jk}-1$. Assume X_{ijk} has an underling continuous variable, X^*_{ijk}, following a standard normal distribution. There are M_{jk} threshold parameters, τ_{jm}, and define $X_{ijk}=m$ if $\tau_{jm} < X^*_{ijk} \leq \tau_{jm+1}$, where $m=0, 1, \ldots, M_{jk}-1, \tau_{j0}=-\infty$, and $\tau_{jM_{jk}}=\infty$. In this study, we assume each test has a unique dimension. Then, item scores on test k can be fitted by the one-factor model for ordered-categorical variables:

$$X^*_{ijk} = \gamma_{jk} s_{ik} + \varepsilon_{ijk}, \tag{1}$$

where s_{ik} is test taker i's score on factor k, which has mean 0 and standard deviation 1; γ_{jk} is the factor loading (regression coefficient) of item j in test k on factor k; and ε_{ijk} is the error term, which is independent of s_{ik} and the error terms of the other items and has mean 0 and variance $1-\gamma^2_{jk}$. Alternatively, each test can be fitted by a unidimensional item response theory (IRT) model, for example, the one-, two-, or three-parameter logistic model; the graded-response model; or the generalized partial credit model (GPCM; Baker & Kim, 2004). The two-parameter logistic model and the graded-response model have been shown to be equivalent to the one-factor model for ordered-categorical variables (Equation (1); Wirth & Edwards, 2007). The GPCM (Muraki, 1992; Muraki & Carlson, 1995) uses a different item response function from the graded-response model; however, the models are virtually indistinguishable empirically in terms of the estimates of the probabilities of item responses conditional on a latent skill (Ostini & Nering, 2005). In this study, the GPCM is employed, and its item response function for test k is written as

$$p_{ijmk} = P\left(X_{ijk}=m|\theta_{ik}, a_{jk}, \boldsymbol{\beta}_{jk}\right) = \frac{\exp\left(a_{jk}\theta_{ik}m + \beta_{jmk}\right)}{\sum_{v=0}^{M_{jk}-1}\exp\left(a_{jk}\theta_{ik}v + \beta_{jvk}\right)}, \tag{2}$$

where $\beta_{j0k} \equiv 0$; θ_{ik} is test taker i's latent skill on dimension k with mean 0 and standard deviation 1, corresponding to the factor score s_{ik} in the one-factor model; p_{ijmk} is the probability of getting score m on item j in test k conditional on θ_{ik} and the item parameters; M_{jk} is the number of score categories on item j in test k with integer item scores 0, 1, 2, $\ldots$, M_j-1; a_{jk} is item j's discrimination (slope) parameter on dimension k; β_{jmk} is item j's intercept on score m and dimension k; and $\boldsymbol{\beta}_{jk}$ is the vector with elements β_{jmk}.

A reasonable way to create a final score from the K factor/skill scores is to estimate a common factor/ability score of the K factor/skill scores at the second level. The second-level model can be written as

$$s_{ik} = \gamma_k s_{ig} + \varepsilon_{ik} \tag{3}$$

or

$$\theta_{ik} = \gamma_k \theta_{ig} + \varepsilon_{ik}, \tag{4}$$

where s_{ig} (θ_{ig}) is test taker i's (unknown) common factor (general ability) score with mean 0 and standard deviation 1, which can be treated as the final aggregation score to be estimated; γ_k is the factor loading (regression coefficient) of test k's factor/skill scores on the common factor; and ε_{ik} is the error term, which is independent of s_{ig} and the error terms of the other tests and has mean 0 and variance $1-\gamma^2_k$. The correlation between two factor/skill scores is

$$\rho_{s_{ik},s_{ik'}} = \gamma_k \gamma_{k'}, \tag{5}$$

and the correlation between a factor/skill score s_{ik} (θ_{ik}) and the common factor/ability score s_{ig} (θ_{ig}) is the factor loading γ_k.

If all the factor loadings are equal to 1 in Equations (3) or (4), then all the tests are dependent on the same general factor/ability,[1] in which case, aggregation methods for tests with unidimensional within- and across-test structures can be used (for a comparison of such methods, see Fu, 2011). The present study concerns the case in which the factor loadings are not all equal to 1; that is, the tests measure different dimensions. However, if the factor loading of a test is too low, for example, smaller than .3, we may remove this test score from the set of test scores to be aggregated because it makes little contribution to the aggregate score. In such an instance, it may be better to report this test score separately.

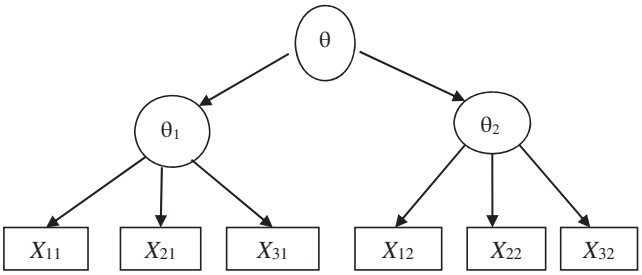

Figure 1 A two-level item response theory model with the one-factor model at the second level. X_{jk} indicates item score on item j in test k.

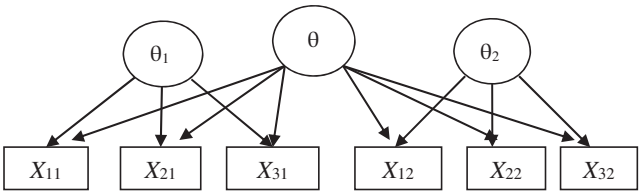

Figure 2 Testlet model. X_{jk} indicates item score on item j in test k.

Combining Equations (2) and (4), we have a two-level GPCM as shown in Figure 1 (Sheng & Wikle, 2008; Yao, 2010). Rijmen (2010) showed that the two-level GPCM is equivalent to the testlet model (Bradlow, Wainer, & Wang, 1999; Li, Bolt, & Fu, 2006). The testlet model is presented graphically in Figure 2 and can be written as

$$p_{ijmk} = P\left(X_{ijk} = m \left| \theta_{ik}, a_{jg}, C_k, \boldsymbol{\beta}_{jk}\right.\right) = \frac{\exp\left[ma_{jg}\left(\theta_{ig} + C_k\theta^*_{ik}\right) + \beta_{jkm}\right]}{\sum_{v=0}^{M_{jk}-1}\exp\left[va_{jg}\left(\theta_{ig} + C_k\theta^*_{ik}\right) + \beta_{jkv}\right]}, \tag{6}$$

where a_{jg} is item j's discrimination (slope) on general ability θ_{ig}; θ^*_{ik} is the skill specific to testlet k with mean 0 and standard deviation 1 and θ^*_{ik}s are independent of each other as well as of the general ability θ_{ig}; C_k is the common item discrimination parameter (slope) for items in testlet k on the skill specific to testlet k; and $a_{jg}C_k$ is item j's discrimination on the specific skill θ^*_{ik}. Note that in our case, we treat one test as a testlet. It is easily seen that Equations (2) and (4) are equivalent to Equation (6) by the following relationships:

$$a_{jg} = a_{jk}\gamma_k, \tag{7}$$

$$C_k = \frac{\sqrt{1-\gamma_k^2}}{\gamma_k}, \tag{8}$$

$$\theta^*_{ik} = \frac{\varepsilon_{ik}}{C_k}. \tag{9}$$

It should be pointed out that when there are only two tests, the one-factor model at the second level as well as the testlet model are not identified, as the factor loading estimates are not unique. When there are only three tests, the one-factor model is just identified, meaning that the estimates of factor loadings have unique defined solutions. When there are more than three tests, the models are identified, as there is one set of best solutions for estimates of model parameters based on a selected fit function, for example, maximizing the likelihood function (Kline, 1998).

The common estimation method for IRT models including the testlet model is the maximum marginal likelihood estimation method with expectation and maximization (MML-EM; Bock & Aitkin, 1981). Recently, Haberman (2013) employed the stabilized Newton–Raphson algorithm for the maximum marginal likelihood estimation of IRT models. There are three different estimation methods for test takers' skills/abilities: maximum likelihood estimation, expected a

posteriori (EAP), and maximum a posteriori (Baker & Kim, 2004). The reliability of EAP estimates of a skill/ability is estimated as (Haberman & Sinharay, 2010; Wainer et al., 2001)

$$\widehat{R}^2_{\theta} = \frac{var\left(\widehat{\theta}\right)}{var\left(\widehat{\theta}\right) + \sum_{i=1}^{I} \widehat{\text{var}}\left(\theta_i\right)/I}, \tag{10}$$

where $var\left(\widehat{\theta}\right)$ is the variance of EAP estimates of a skill/ability in the sample with I test takers and $\widehat{\text{var}}\left(\theta_i\right)$ is the estimated posterior variance of test taker i's EAP estimate.

The common factor scores can be estimated using several methods (for a review from practical perspectives, see DiStefano, Zhu, & Mindrila, 2009). We consider two methods here: the average of standardized input variables (i.e., item scores or factor score estimates on individual tests in our case) and the regression score estimate. The average scores do not depend on factor loading estimates, while the regression scores utilize estimates of factor loadings as well as test score correlations. Using the one-factor model at the second level (Equations (3) or (4)) as an example, the regression scores are estimated as

$$\widehat{\mathbf{S}}^g_{I\times 1} = \widehat{\mathbf{S}}^k_{I\times K}\widehat{\mathbf{B}}^{-1}_{K\times K}\widehat{\boldsymbol{\gamma}}_{K\times 1}, \tag{11}$$

where I is the total number of test takers, $\widehat{\mathbf{S}}^g$ is the column vector of estimates of all students' common factor scores, $\widehat{\mathbf{S}}^k$ is the matrix of all students' factor score estimates, $\widehat{\mathbf{B}}^{-1}$ is the inverse of the matrix of correlations among the K factor score estimates, and $\widehat{\boldsymbol{\gamma}}$ is the column vector of factor loading estimates. Both methods are actually the weighted sums of standardized input variables: For average scores, weights are equal to $1/K$, whereas for regression scores, weights are equal to $\widehat{\mathbf{B}}^{-1}\widehat{\boldsymbol{\gamma}}$. The weights of regression scores maximize the correlation between estimated and true common factor scores and thus the reliability of aggregate score estimates. In this sense, IRT ability estimates are similar to regression score estimates. Note that reliability is the squared correlation between estimated and true common factor scores. For a set of weights w_k, $k = 1, 2, \ldots K$, the reliability of aggregate scores is computed by substituting estimates of the factor loadings and/or weights into Equation (12):

$$R^2 = \frac{\left(\sum_{k=1}^{K} w_k\gamma_k\right)^2}{\sum_{k=1}^{K} w_k^2 + 2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} w_k\gamma_k w_{k'}\gamma_{k'}}. \tag{12}$$

Note that Equation (12) is equal to McDonald's (1985) ω coefficient if all w_ks are equal, and additionally, if all γ_ks are equal, Equation (12) is equal to Cronbach's alpha (Zinbarg, Revelle, Yovel, & Li, 2005).

Theoretically, the testlet model or the two-level, one-factor model is appropriate for aggregating test scores with unidimensional within-test structure and multidimensional across-test structure. However, in practice, people may prefer to use alternative score aggregation methods for practical reasons. The following are some examples:

1. Use the sum or average of standardized test raw scores as an aggregate score. This method actually uses the average method for factor score estimates on the first- and second-level factor models. This method is simplest; however, if items raw scores have different factor loadings in the first-level factor model, and/or test raw scores have different factor loadings in the second-level factor model, the reliability of these aggregate scores will be much lower than that of the aggregate scores that take different factor loadings into account.
2. The aggregate score is the regression factor score estimate on the second-level factor model with standardized test raw scores as input variables. The method uses average factor score estimates at the first level and regression factor score estimates at the second level. This method will generate aggregate scores with higher reliability than the first method if test raw scores have different factor loadings in the second-level factor model.
3. Each test is calibrated by a unidimensional IRT model, and aggregate score (i.e., general ability) is the average of skill scores on individual tests. This method uses IRT ability estimates at the first level and average factor score estimates at the second level. This method will generate aggregate scores with higher reliability than the first method if item raw scores have different factor loadings at the first level.

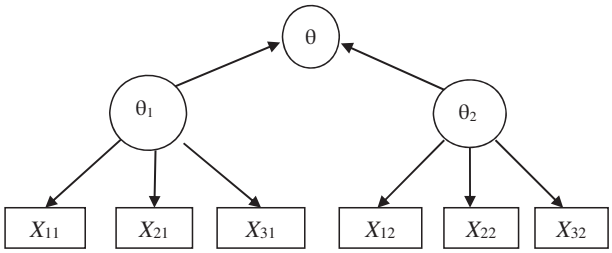

Figure 3 A two-level item response theory model with the regression model at the second level. X_{jk} indicates item score on item j in test k.

4. The first level is the same as the third method: Each test is calibrated by a unidimensional IRT model. At the second level, the aggregate score is the regression factor score estimate of skill scores on individual tests. This method is similar to the testlet model or the two-level, one-factor model and will generate aggregate scores with higher reliability than the third method will if test raw scores have different factor loadings at the second level.
5. All tests are calibrated together by a unidimensional IRT model, and the aggregate score is the ability estimate from the unidimensional IRT model. This method assumes that all tests are parallel tests. Thus, if test raw scores have different factor loadings at the second level, this method is less reliable than the fourth method; however, it is more reliable than the third method, because in this method, all items directly contribute with weights to the general ability.

Another method that may be used to aggregate skill scores on different dimensions to obtain an overall ability score was explored in the literature in the context of generating an overall score from subscores in a test. This approach also uses a two-level IRT model. The first level is a multidimensional IRT model where each item in a test loads only on the dimension specific to that test and test dimensions are allowed to be correlated. At the second level, this model uses a weighted sum of skill scores on each test as overall ability. To determine the weights, Sheng and Wikle (2008) used a regression model in which skill scores on individual tests were independent variables and overall ability was the dependent variable (see Figure 3). However, this model is not mathematically identified because the overall ability and regression coefficients are unknown. Yao (2010) used another method to determine the weights. In particular, weights were chosen so that the overall ability as a linear combination of skill scores had minimum variance and thus maximum information (Yao & Schwartz, 2006). However, whether this criterion is reasonable in practice for score aggregation is questionable for the reason stated below. By this criterion, test takers with different skill scores might have different weights in score aggregation because the variance of overall ability is a function of the variances and covariances of skill scores on individual tests, and for different skill scores, their variances and covariances are also different. It is possible that a small difference in the skill scores between two students will result in a large difference in their overall ability. Consequently, this approach was not considered in the current study.

Simulation Study

Score Aggregation Methods

We compared six score aggregation methods for unidimensional tests on different dimensions. The first is simply based on raw scores; the others are based on estimates from the specified models:

1. *Method 1* (M1): The aggregate score is the weighted sum of standardized test raw scores. The weights are predefined. In the current study, all test weights are equal and sum to 1 so that the aggregate score is the average of the standardized test raw scores.
2. *Method 2* (M2): The aggregate score is an estimate of the regression factor score of the one-factor model based on the correlation matrix of test raw scores.
3. *Method 3* (M3): The overall ability score is an estimate obtained from a unidimensional GPCM model based on the items from all tests.
4. *Method 4* (M4): The skill scores on each test are estimates obtained from a unidimensional GPCM model based on the items from the specific test only, and then the overall ability score is the weighted sum of the standardized skill

scores on the individual tests. The weights are predefined. In the current study, all test weights are equal and sum to 1 so that the aggregate score is the average of the standardized skill scores on individual tests.

5. *Method 5* (M5): The overall ability score is an estimate of the regression factor score of the one-factor model based on the correlation matrix of estimated test skill scores; skill scores on an individual test are estimated previously, based on a unidimensional GPCM.
6. *Method 6* (M6): The overall ability score is estimated from the testlet model (Equation (6)).

All the preceding models were discussed in the previous section. The average of test raw scores or skill scores (M1 and M4) and the overall ability from the GPCM model (M3) are convenient ways to aggregate scores and are commonly used in practice. Note that M5 is based on the higher order GPCM model with the two levels estimated separately, whereas M6 is equivalent to the same model with the two levels estimated concurrently. It is not only theoretically interesting but also of practical importance to compare the separate and concurrent estimations of the higher order GPCM model. The concurrent estimation is estimating a multidimensional IRT model (testlet model), so the estimation burden could be heavy, and sometimes the estimation may not be stable and accurate (Fu, 2009). Alternatively, if separate estimation is comparable to the concurrent estimation, then it provides a very convenient way to aggregate scores from previously estimated skill scores on individual tests. In addition, in some cases, tests administered on different occasions may have to be calibrated separately, because scores for individual tests have to be reported to test takers after each administration. The regression score estimates based on test raw scores (M2) are even easier to generate than M5, so this method was also included in the comparisons.

Simulation Design

The purpose of this comparison study was to investigate which method provided the best aggregate score estimation in terms of reliability under various conditions. The design factors considered here were (a) number of tests, (b) individual test length, (c) correlation between tests, and (d) factor loading distribution at the second level. These factors were studied for the following reasons:

1. The number of tests has a significant impact on the accuracy of aggregate score estimation. In general, more tests lead to more accurate aggregate score estimation, because more information regarding aggregate scores is available, assuming these tests measure a well-defined appropriate behavior domain (McDonald, 1985). In addition, as mentioned previously, the number of tests is related to the identification of the one-factor model: When aggregate scores are based only on two tests, which occurs in practice, the one-factor model is not identified. Therefore two levels of the number-of-tests factor were considered here: two tests and four tests. Because the one-factor models were not identified under the two-test condition, M2, M5, and M6 were not included in this condition.
2. Correlations between test scores affect aggregate score reliability through their relationships with factor loadings and weights for regression factor scores. The product of the factor loadings of two test scores is the correlation between the two test scores (see Equation (5)). For three or more tests, the correlation matrix among test scores determines the estimates of factor loadings and weights for regression factor scores. High correlations lead to high factor loadings and thus more accurate aggregate score estimates across all six compared methods. In this study, the correlations between test skill scores were set to five levels, .3, .5, .7, .8, and .9, to represent low, intermediate, and high correlations between test skill scores. (However, in the four-test condition, the correlations between test skill scores may not be the same across-test pairs; see Point 4.) More correlations at the high level were considered here because the correlations between tests are expected to be high in real-life testing, as the measured constructs are usually strongly related to each other. For example, in a similar context, Sinharay (2010) surveyed more than 20 operational tests and found that the disattenuated (raw) subscore correlations within each test were at least .7. In addition, use of aggregate scores in practice demands high reliability, which directly relates to the strength of the correlations between test scores.
3. Test length also has a significant impact on the accuracy of score estimation, with longer tests generally leading to more accurate score estimation, because longer tests generally have higher reliabilities and thus lead to higher correlations between two tests for a given pair of constructs. The test length per test was considered at four levels: 10, 20, 30, and 60.

4. Factor loading distribution at the second level, that is, whether factor loadings are similar or quite different among test scores, also affects the accuracy of aggregate score estimates. If all factor loadings are the same, then the reliabilities of aggregate score estimates from the average score method (M1 and M4) and the regression factor score method (M2, M5, and M6) are also the same (as seen in Equations (11) and (12)). As noted previously, correlations between test scores that measure similar constructs in real-life testing programs are usually high. Therefore, in practice, the factor loadings are often high and similar. However, it is possible for factor loadings to be quite different among tests. In this case, the regression score method is expected to provide more accurate aggregate score estimates than the average score method, because the regression score method produces aggregate score estimates with maximum reliability. Therefore, in this study, equal and unequal factor loadings were considered. As noted before, factor loadings are indeterminate when only two tests are aggregated; however, they are determined by correlations between test scores for three or more tests. For the equal condition, all factor loadings (γ_k) under a correlation level between test skill scores were equal to the square root of the correlation. For the unequal condition under a correlation level between test skill scores, half of the factor loadings were .95, and the other half were equal to the correlation between test skill scores divided by .95. Note that under the four-test condition, we assumed the correlations of test skill scores between Tests 1 and 3 and between Tests 2 and 4 were equal to the associated correlation level (i.e., .3, .5, .7, or .8). We also assumed that Tests 1 and 2 had factor loadings of .95, which led to the correlations between Tests 1 and 2 being equal to .90 and the correlations between Tests 1 and 4 and between Tests 2 and 3 being equal to the associated correlation level also. Because, at a correlation of .9, all factor loadings were very close, the unequal factor loading condition did not include the correlation level of .9.

Each design cell had 50 replications with two tests and with four tests. Each test had a sample size of 3,000 and included 70% dichotomous items and 30% three-score category (0, 1, and 2) polytomous items. This is similar to the percentages of dichotomous items and polytomous items in an actual CBAL test. Thus the maximum possible test raw scores for the 10-, 20-, 30-, and 60-item tests were 13, 26, 39, and 78, respectively. The item responses were generated assuming the underlying model was the higher order GPCM model (Equations (2) and (4)) for the following two reasons. First, the higher order GPCM model provides a convenient way to generate data consistent with the score aggregate scenario under study. Second, although data generated by the higher order GPCM will favor the testlet model, we wanted to see if simpler methods can produce comparable results with the more complicated model (i.e., the testlet model). The overall ability scores θ_{ig}, $i = 1, \ldots, 3{,}000$, were generated from a standard normal distribution. For test taker i, given an overall ability θ_{ig}, the test taker's test skill score θ_{ik} was sampled from the $N\left(\gamma_k\theta_{ig}, \sqrt{1-\gamma_k^2}\right)$ distribution. Factor loadings (γ_k) were derived from correlations between test skill scores, as described earlier. Item parameters were generated based on the estimates from an actual CBAL test administration. Item discrimination (slope) parameters a_{jk} were sampled from a lognormal distribution with mean $-.07$ and standard deviation .39, resulting in a_{jk}s with mean 1 and standard deviation .4. Intercept parameters β_{jkm} for Score Category 1 were sampled from $N(.3, 1)$ and for Score Category 2 were simulated from $N(-.4, 1)$. Table 1 summarizes the simulation conditions and parameters used to generate the simulated data. When generating simulated data, the true ability/skill scores for each test taker were fixed within each combination of the three design factors of number of tests, individual test length, and correlation between tests, whereas the true item parameters were fixed within each combination of the four design factors.

Estimation

The GPCM model parameters, except for those in the 60-item four-test condition, were estimated using the mdltm program (von Davier, 2008). In this program, the MML-EM is used. The GPCM model parameters in the 60-item four-test condition and all testlet models were estimated by the MIRT package (Haberman, 2013), which employs log-linear modeling and implements the maximum marginal likelihood method with the stabilized Newton–Raphson algorithm. The reason for using the MIRT package for all the testlet models and longer tests is that this program is more efficient and effective than other programs for estimating multidimensional IRT models and longer tests, based on our experience. As for ability/skill estimation, EAP estimates were obtained from the respective software.

Table 1 Simulation Study Design and Parameters to Generate Simulated Data Sets

Factor loading distribution at second level	No. tests	No. items per test	Correlation between test skill scores	Factor loadings of test skill scores on overall ability
Equal	2	10, 20, 30, 60	.30	all γs = .55
			.50	all γs = .71
			.70	all γs = .84
			.80	all γs = .89
			.90	all γs = .95
	4	10, 20, 30	.30	all γs = .55
			.50	all γs = .71
			.70	all γs = .84
			.80	all γs = .89
			.90	all γs = .95
Unequal	2	10, 20, 30, 60	.30	$\gamma_1 = .95, \gamma_2 = .32$
			.50	$\gamma_1 = .95, \gamma_2 = .53$
			.70	$\gamma_1 = .95, \gamma_2 = .74$
			.80	$\gamma_1 = .95, \gamma_2 = .84$
	4	10, 20, 30	.30[a]	$\gamma_1 = \gamma_2 = .95, \gamma_3 = \gamma_4 = .32$
			.50[a]	$\gamma_1 = \gamma_2 = .95, \gamma_3 = \gamma_4 = .53$
			.70[a]	$\gamma_1 = \gamma_2 = .95, \gamma_3 = \gamma_4 = .74$
			.80[a]	$\gamma_1 = \gamma_2 = .95, \gamma_3 = \gamma_4 = .84$

Note. $N = 3{,}000$; overall ability, $\theta_{ig} \sim N(0, 1)$; test skill score, $\theta_{ik} \sim N\left(\gamma_k \theta_{ig}, \sqrt{1-\gamma_k^2}\right)$; item discrimination parameter, $a_{jk} \sim$logNorm $(-.07, .39)$, Mean$(a_{jk}) = 1$, Std$(a_{jk}) = .4$; item intercept parameter, $\beta_{jk1} \sim N(-.3, 1)$, $\beta_{jk2} \sim N(-.4, 1)$.

[a]The true correlations of test skill scores for the following test pairs: Tests 1 and 3, Tests 1 and 4, Tests 2 and 3, and Tests 2 and 4. The true correlation of test skill scores between Tests 1 and 2 is equal to $\gamma_1 * \gamma_2 = .90$ and between Tests 3 and 4 is equal to $\gamma_3 * \gamma_4$.

Evaluation Criterion

The accuracy of the aggregate scores was evaluated by the corresponding reliability. The reliability of the aggregate score estimates in a replicate data set was computed as the squared correlation between the aggregate score estimates and the true aggregate scores. The mean of the reliabilities of the 50 replications within a design cell was the unit to compare reliabilities across design cells. Note that this reliability of aggregate score estimates is based on the known true aggregate scores, whereas for the estimated reliability based on Equations (10) or (12), the true aggregate scores are unknown.

Results

Because the data were generated by the higher order GPCM model, it is interesting to see how well the factor loadings were recovered in the testlet model (M6).[2] Figure 4 shows the distribution of the absolute biases of the factor loadings in M6 in the four-test condition, that is, the absolute difference between the true and estimated factor loadings. One can see that 97% of the factor loading estimates in M6 under the four-test condition differed from the true values by .04 or smaller. Relatively large differences (.06 or larger) appeared under the condition of test correlation equal to .3. The two largest differences, .11 and .07, were associated with the two tests with small factor loadings in the 10-item condition with the correlation between test skill scores at .3 and unequal factor loadings. This is understandable, considering that the two tests were short and had little contribution to the aggregate scores and thus their factor loadings were difficult to estimate. Except for this case, test length did not appear to be influential for factor loading estimates, as these estimates were very close across-test lengths for every design cell.

As for the mean reliabilities of aggregate score estimates in each design cell, we have the following observations:

1. The mean reliabilities from all the methods were quite similar under each design cell, with the following exceptions. First, when the correlations between test skill scores were low to intermediate (.3 or .5) and the test length was short (10 items) in the two-test condition, the mean reliabilities from M3 were lower than those obtained using M1 and M4 by a magnitude of .04. Second, under the two-test condition with unequal factor loadings and correlations between test skill scores at .3 and .5, M3 estimations were not stable, and the mean reliabilities might be higher or lower than those obtained from other methods. Third, in the four-test, 10-item, .3 correlation and equal factor

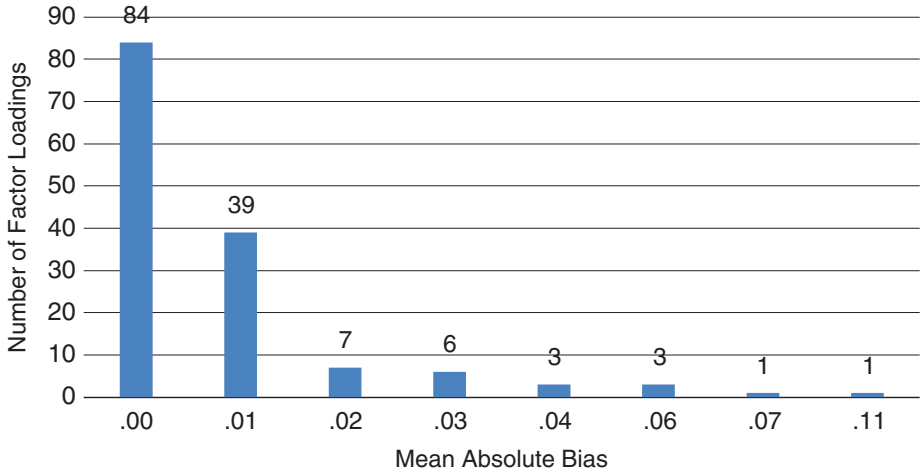

Figure 4 Distribution of mean absolute biases of factor loading estimates in method 6 in the four-test condition.

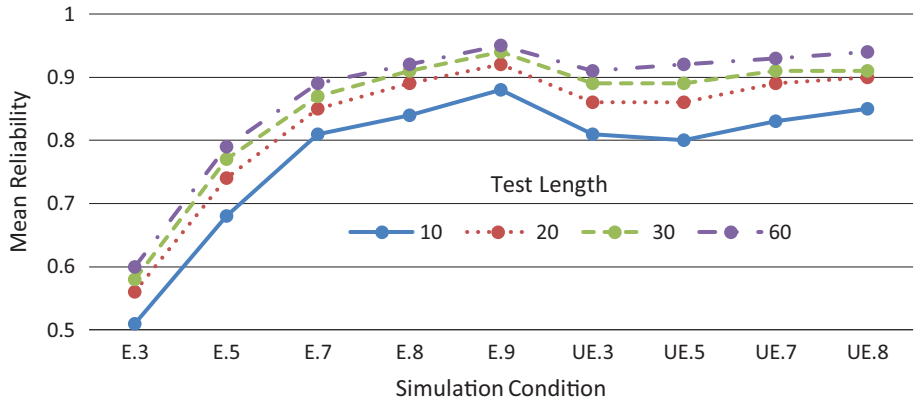

Figure 5 Mean reliability estimate of aggregate scores from method 6 by simulation conditions. In the horizontal axis labels, "E" refers to equal factor loading, "UE" refers to unequal factor loading, and the decimal refers to test correlation.

loading condition, M3 was the least accurate model, with a mean reliability lower than other methods by a maximum of .05. Fourth, when the correlations between test skill scores were .3 or .5 with unequal factor loadings under the four-test condition, the mean reliabilities obtained from the averaging methods (M1 and M4) were lower than those obtained from other methods by a maximum of .18. Under the four-test condition, the testlet mode (M6) always had the highest reliabilities (although they are equal or very close to some other methods). This is expected, as the testlet model was the true model used to generate the simulated data.

2. All the mean reliabilities grew larger with increasing correlations between test skill scores and test lengths, which is apparent in Figure 5, which shows the mean reliabilities from M6 in each design cell as an example. However, there is one exceptional case, as mentioned previously: Under the two-test condition with unequal factor loadings and correlations between test skill scores at .3 and .5, M3 estimations were not stable; the aggregate score estimates could be very good or very poor, depending on the simulated data.

3. The mean reliabilities under the four-test condition were higher than those under the two-test condition, which was expected, as the total test length increased. Figure 6 shows this pattern for M1 as an example

4. When the correlation between test skill scores was low to intermediate, the mean reliabilities under unequal factor loadings were larger than those under equal factor loadings, which was expected based on Equation (12). However, the differences in mean reliabilities decreased with increasing correlations between test skill scores, as the factor

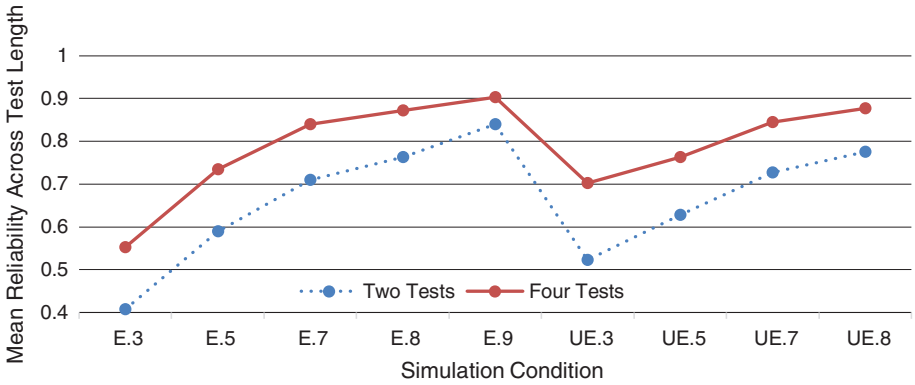

Figure 6 Mean reliability estimate of aggregate scores across test length from method 1 by simulation conditions. In the horizontal axis labels, "E" refers to equal factor loading, "UE" refers to unequal factor loading, and the decimal refers to test correlation.

loadings became more homogenous. When the correlation between test skill scores was .7, the mean reliabilities were quite close. See the mean reliability pattern from M6 in Figure 5 as an example.

5. We calculated the mean estimated reliabilities of the general ability in the testlet models based on Equation (10) as well as the mean estimated reliabilities of the aggregate scores from M2 and M5 based on Equation (12) for the four-test condition. For all three models in the four-test condition, these estimated reliabilities were very close to those based on the known true aggregate score in every design cell: The mean differences across all design cells were .011, .004, and .006 for M2, M5, and M6, respectively. This indicates that reliability estimates based on Equations (10) or (12) were quite accurate.

We calculated the mean Pearson correlations between aggregate score estimates for each design cell and found the following:

1. Under the two-test condition, the mean correlations among M1, M3, and M4 were very high (at least .96) across all conditions, except for M3 at the low to intermediate correlation between test skill scores (.3 or .5) and with the test length of 10 items.
2. Under the four-test condition, the mean correlations between M2 and M5 were at least .99, those between M1 and M4 were at least .99, and the mean correlations of M2 and M5 with M6 were at least .98 across all design cells.
3. Under the four-test condition with equal factor loadings, the mean correlations among all methods were at least .98, except for some cases with the correlation between test skill scores of .3 and the test length of 10 items, where the minimum correlation was .93.
4. Under the four-test condition with unequal factor loadings and the correlation between test skill scores being .3 or .5, the correlations, except for those mentioned in Point 2, were in general lower than those in the corresponding conditions with equal factor loadings. The correlations could be as low as .89, and the maximum difference was .09.

Real Data Example

The real data used in this study were from two CBAL writing tests administered in spring 2011: Ban Ads (BA) and Mango Street (MA). Students took both tests in either of two sequences, BA-MA or MA-BA, within an interval of approximately 2 months. A total of 514 students took the BA-MA sequence, and 409 students took the MA-BA sequence. BA focuses on persuasive/argumentative writing, whereas MA focuses on writing about literature. Both tests are based on a common scenario, and items in each test are organized under four tasks based on the nature of the questions. The first three tasks are lead-in tasks measuring critical thinking skills that are necessary for writing a good essay on a specific genre, and the fourth task is writing an essay. Both tests include dichotomous and polytomous items. BA has 25 items with a maximum possible item score of 5 and a maximum test score of 40; MA has 14 items with a maximum possible item score of 5 and a maximum test score of 32.[3] Therefore these tests were equivalent to the 31- and 25-item tests in the simulation study, respectively, in terms of test score points.

Three methods, M1, M3, and M4, were used to aggregate the two test scores in each sequence. For the unidimensional IRT model in M3 and M4, GPCM was used. Table 2 displays the correlations and reliabilities of test raw scores (Cronbach's

Table 2 Test Statistics for Real Data

Statistics	BA-MA	MA-BA
Raw score		
Raw score correlation	.56	.63
BA reliability	.80	.78
MA reliability	.84	.78
Skill score from separate calibration		
Skill score correlation	.44	.56
BA reliability	.93	.92
MA reliability	.94	.90

Note. BA = Ban Ads; MA = Mango Street.

Table 3 Correlations of Aggregate Score Estimates in Real Data

	M1	M3
BA-MA		
M3	.99	
M4	.98	.98
MA-BA		
M3	.97	
M4	.98	.98

Note. BA = Ban Ads; MA = Mango Street.

alpha) and skill scores (Equation (10)) for both test sequences. The correlations between the two tests were moderate for both test sequences. Table 3 shows the correlations of aggregate score estimates among the three methods. As can be seen, all correlations were very high (at least .97). This result is consistent with that found under similar simulated conditions.

Summary and Discussion

Main Findings

The main findings from the simulation study and real data analysis are as follows:

1. In general, reliability of aggregate scores increased when the correlation between test skill scores, test length, and number of tests to be aggregated increased.
2. In general, the aggregate scores obtained from all methods were highly correlated.
3. The factor loading estimates in the testlet model (M6) under the four-test condition were close to the true values in most cases.
4. The M3 approach did not perform well when the tests were short, the correlations between test skill scores were low to intermediate, and the factor loadings were equal. In addition, its performance was unstable under the two-test condition when the correlation between the two test skill scores was low to intermediate with unequal factor loadings.
5. M1 and M4 did not perform well under the four-test condition with low to intermediate correlations between test skill scores (.3 or .5) and unequal factor loadings.
6. In general, M1 and M4 performed best under the two-test condition.
7. Overall, M2, M5, and M6 produced similar results and performed best under the four-test condition. M2 and M5 involved much less computation and were much easier to carry out than the testlet model (M6) and thus are recommended if there are difficult issues in applying the testlet model in practice, for example, test scores need to be reported to test takers after each administration or a test data set is so big that the estimation of the testlet model is beyond the capacity of the available computers.
8. Applying M1, M3, and M4 to CBAL writing data with two tests shows that the aggregate scores computed using the three methods are highly correlated (at least .97).

The preceding results suggest that M1 and M4 might be best used for aggregating scores for two tests and M2, M5, and M6 for three or more tests. These recommendations are based on general theoretical principles, the IRT model (i.e., GPCM), simulated data, and/or real data used in this study. The results and recommendations should, however, only be generalized with caution to conditions beyond those used in this study.

Several issues regarding score aggregation are discussed in the following sections: achieving adequate reliability of aggregate scores in practice, different types of score aggregation based on test dimensionality, and suggestions for practitioners.

Reliability of Aggregate Scores

Test scores should reach adequate reliability for practical use. Generally speaking, test scores should have a reliability of at least .90 if used for high-stakes purposes and at least .80 or .85 for low-stakes purposes (Wells & Wollack, 2003). This requirement applies to use of aggregate scores also. Therefore we need to estimate the reliability of aggregate scores for purposes of planning and designing tests or determining whether the aggregate scores can be used for their intended purposes.

Given that correlation(s) between test scores and factor loadings are known, as in simulation conditions, reliability of aggregate scores can be computed by Equation (12). Table 4 shows the true test weights used to generate regression factor scores and their generated reliability for each simulated condition of correlation between test skill scores and factor loading distribution in the current simulation study. Comparing the generated reliabilities in Table 4 with the reliability estimates of M1 and M4 in the two-test condition and M2, M5, and M6 in the four-test condition, we observe that for the two-test equal-loading condition and the four-test condition, the generated reliabilities are larger than the corresponding reliability estimates, although they are very close to the corresponding reliabilities in the 60-item condition. The reason is that the simulated tests contain measurement errors so that the correlations between simulated test scores are smaller than the true correlations between test skill scores. Actually, a true correlation can be estimated by the disattenuated correlation between two simulated test scores, that is, the correlation of the two simulated test scores divided by the product of the reliabilities of the two test scores. Longer tests have higher reliabilities so that they lead to higher correlations between test scores and more stable aggregate scores. Table 5 lists the mean reliability of individual tests by test length/maximum possible test scores across all the simulated data. From Table 5, it is apparent that reliability increases with test length, as expected. Note that in the two-test condition, the weight of .5 is used for both tests in M1 and M4, which is not consistent with the true weights in some correlation between tests conditions. For the equal factor loading condition, the weights of test scores do not affect the reliability of aggregate scores as long as they are equal across tests. However, they do affect the reliability of aggregate scores when factor loadings are not equal. This is the reason why, in the two-test condition with unequal factor loading, the reliabilities of aggregate scores from M1 and M4 are much lower than the corresponding generated reliabilities across all item conditions, especially when the correlation between test skill scores is .3 or .5.

Suppose we have test data and want to determine if the aggregate scores are stable enough for intended use. In this case, we know the correlation matrix between test scores. If we have only two tests and want to use a weighted sum of test scores as an aggregate score (i.e., M1 and M4), we cannot obtain estimates of factor loadings and test weights from the data because the one-factor model for two tests is not identified. Instead, we may assign factor loadings to the two tests based on some theoretical or practical considerations to reflect the relative contribution of each test to the aggregate scores and estimate the test weights by $\widehat{\mathbf{B}}^{-1}\hat{\boldsymbol{\gamma}}$ and the reliability of aggregate scores based on Equation (12). As mentioned before, if equal factor loading across two tests is assumed, then the actual weights used to aggregate scores do not matter, as long as they are equal across the two tests. However, if factor loading and test weights are misspecified for tests with unequal loadings (if there is a way to judge that), then this will decrease the reliability of aggregate scores, as described previously for the two-test condition with unequal factor loading. If we have three or more tests, then for M2 and M5, the factor loadings of the one-factor model can be estimated from the test data, and the reliability of aggregate scores can be estimated by substituting estimates for the parameters in Equation (12); for M6, the reliability of aggregate scores can be estimated by Equation (10) after fitting a testlet model. Before estimating the reliability of aggregate scores, a test with factor loading estimates below a cut point (e.g., .3) may be removed from the aggregation.

Suppose another scenario: We do not have tests and test data as in the test planning and design stage, and we need to estimate how many items/score points should be included in a test so that the aggregate scores will reach desired

Table 4 Generated Test Weights and Reliabilities of Regression Factor Scores

		Two tests			Four tests				
Factor loading distribution	Correlation between test skill scores	Weight 1	Weight 2	Reliability of common factor scores	Weight 1	Weight 2	Weight 3	Weight 4	Reliability of common factor scores
Equal	.30	.42	.42	.46	.29	.29	.29	.29	.63
	.50	.47	.47	.67	.28	.28	.28	.28	.80
	.70	.49	.49	.82	.27	.27	.27	.27	.90
	.80	.50	.50	.89	.26	.26	.26	.26	.94
	.90	.50	.50	.95	.26	.26	.26	.26	.97
Unequal	.30	.94	.03	.90	.79	.79	−.12	−.12	.93
	.50	.92	.07	.91	.72	.72	−.13	−.13	.92
	.70	.85	.14	.91	.63	.63	−.08	−.08	.93
	.80	.77	.23	.92	.53	.53	−.01	−.01	.95

Table 5 Mean Reliability Estimates of Individual Tests

No. items per test	Max. possible test score	Raw score (Cronbach's alpha)	Skill scores (Equation (10))
10	13	.68	.71
20	26	.81	.83
30	39	.87	.88
60	78	.93	.94

reliability. In this case, we first need to estimate the correlations between tests based on relevant theory and/or previous experience. Second, we need to know the relationship between test length and test reliability. This information may come from previous experience or a simulation study pertinent to the situation under study. For example, Table 5 lists the mean reliabilities by test length based on the conditions simulated in the current study. Note that for purposes of predicting test reliability, maximum possible test score is a better indicator of test length than number of items, because in general, a polytomous item with M_i score categories is comparable to $M_i - 1$ dichotomous items in terms of the contribution to test reliability, as seen in the calculation of Cronbach's alpha. Third, with estimated theoretical or empirical correlations between test scores and test reliabilities for tests with given length, we can predict the actual correlation between two test scores by multiplying the theoretical or empirical correlation between the two test scores by the reliabilities of two tests. Fourth, by knowing the estimated correlations between test scores, we can then estimate the reliability of aggregate scores, as described earlier. Following these steps, we can estimate the appropriate reliability and test length of each test so as to reach the desired reliabilities of aggregate scores.

Test Dimensionality and Score Aggregation

When aggregating test scores, an important factor that needs to be considered is test dimensionality within and across tests. Because within-test structure may or may not be unidimensional, and/or across-test structure could be on the same or different dimensions, we can have four potential scenarios.

In the first scenario, both the within-test and across-test structures are unidimensional. For this scenario, Fu (2011) compared several score aggregation methods under the average approach and the prediction approach using simulated and real test data and taking into account students' growth between test administrations. Note that it only makes sense to measure growth when across-test structure is on the same dimension(s). For the average approach, all test scores directly contribute to aggregate scores in an average way, and three methods were compared: average of test raw scores, average of test scale scores each produced by a unidimensional IRT model, and concurrent calibration of all tests using a unidimensional IRT model—that is, M1, M4, and M3 in the current study. For the prediction approach, the aggregate scores are based on the final test scores, and other test scores are used as auxiliary information to improve the estimation

of the final test scores. Also, three methods in the prediction approach were compared: a subscore augmentation method using raw scores, a subscore augmentation method using IRT scale scores (Wainer et al., 2001), and a multilevel IRT model in which previous test scale scores were used to predict the final test scale scores at the second level. Fu found that the average of all test scale scores in the average approach and the subscore augmentation method using IRT scale scores in the prediction approach were most accurate and reliable in the simulated data.

The second scenario is that the within-test structure is unidimensional, whereas the across-test structure is multidimensional. This is the test scenario studied in this report.

In the third scenario, the within-test structure is multidimensional, whereas the across-test structure is on the same dimensions. For example, equivalent forms of a multidimensional test are administered during a school year. One way to create final aggregate scores in this case is to first create an overall score for each test using the recommended methods in this study—M4 for two tests and M5 and M6 for three or more tests—and then produce final aggregate scores from the overall scores of individual tests using M4, as suggested by Fu (2011), because these overall scores are assumed to be on the same dimension. Alternatively, we can first create an overall score for each dimension across test forms using M4 and then use the recommended methods in this study to aggregate these overall scores on different dimensions to final scores. Finally, we may use the testlet model to calibrate all the tests simultaneously to get the estimates of general ability (i.e., aggregate scores).

The final scenario is that at least one test to be aggregated is multidimensional, and across tests, the dimensions are not the same. For this case, we may first aggregate within-test subscores to an overall score and then aggregate across-test overall scores to create the final score, both using the methods recommended in this study. This approach is similar to doing twice the score aggregations of tests on different dimensions. If there are common dimensions across tests, we may first aggregate scores on each of the common dimensions using M4, as suggested by Fu (2011), and then aggregate the (overall) scores on all dimensions to create the final score, using M4 for two dimensions and M5 for three or more dimensions. This approach is basically the same as the second one suggested for the third scenario in the preceding paragraph. A special case of this scenario is that the final test includes all the dimensions in the previous tests. If it is preferable to base the final scores on the final test, then the prediction methods Fu studied can be used. For all these cases, we may also use multidimensional IRT models (possibly with a covariate structure like those models implemented in the MIRT package) to calibrate all tests concurrently and obtain aggregate scores. For the third and fourth scenarios, well-designed studies are needed in the future to make prudent recommendations.

Some Suggestions for Practitioners

When doing score aggregation, the practitioner should first estimate test dimensionalities within and across tests and identify the type of score aggregation being carried out (as one of the four scenarios discussed earlier). Because this study focuses on the case where within-test structure is unidimensional while across-test is multidimensional, we provide some practical suggestions for this type of score aggregation:

1. If only two tests are to be aggregated, sum or average the two test raw scores or IRT skill scores, and use predefined test weights, if necessary.
2. If three or more tests are to be aggregated, estimate regression factor scores of the one-factor model based on test raw scores or skill scores from an IRT model, or estimate general abilities in the testlet model if the estimation is practically feasible.
3. Follow the steps described in the "Reliability of Aggregate Scores" section to estimate reliabilities of aggregate scores and/or to determine the desired reliability and test length of each test. Omit a test from the aggregated test set if its estimated factor loading is below a cut point (e.g., .3), unless including that test is necessary for content coverage reasons. Do not aggregate scores if the aggregate scores are not stable enough for intended uses.

Acknowledgments

Thanks to James Carlson, Randy Bennett, Christine Mills, Carolyn Wentzel, Frank Rijmen, Andreas Oranje, Matthias von Davier, and Peter van Rijn for their helpful suggestions and edits on early versions of this article. I am grateful to Ayleen Gontz for her editorial assistance.

Notes

1 Longitudinal data are not appropriate to detect across-test dimensionality because test dimensionality may be confounded with growth. For example, two parallel tests administered at different times may appear to be on different dimensions because of students' uneven growth. Therefore, to check across-test dimensionality, students' abilities should be the same when they take the different tests.
2 In M2 and M5, the Pearson correlation matrices rather than the disattenuated correlation matrices were used for the second-level, one-factor models. Thus their factor loading estimates did not match the true values.
3 Item weights were not applied in counting item score points and test score points. However, they were applied in computing test raw scores and their reliabilities and correlations.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443 – 459.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153 – 158.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 20*(14), 1 – 14.

Fu, J. (2009, April). *Marginal likelihood estimation with EM algorithm for general IRT models and its implementation in R*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Fu, J. (2011, April). *A comparison of score aggregation methods*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton – Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209 – 227.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30,* 3 – 21.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159 – 176.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73 – 90.

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361 – 372.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68,* 413 – 430.

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61,* 287 – 307.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., … Thissen, D. (2001). Augmented scores: "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343 – 387). Mahwah, NJ: Erlbaum.

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Retrieved from http://testing.wisc.edu/Reliability.pdf

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58 – 79.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47,* 339 – 360.

Yao, L., & Schwartz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30,* 469 – 492.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonalds's omega: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123 – 133.

Suggested citation:

Fu, J., & Feng, Y. (2018). *A comparison of score aggregation methods for unidimensional tests on different dimensions* (Research Report No. RR-18-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12194

Action Editor: James Carlson

Reviewers: Andreas Oranje, Frank Rijmen, and Matthias von Davier