Bulletin of Educational Psychology, 2013, 45(2), 261-278
National Taiwan Normal University, Taipei, Taiwan, R.O.C.
DOI：10.6251/BEP.20130222.2

Explorations of Composite Scores under the Multivariate Proficiency Distribution Using IRT

Shun-Wen Chang
Department of Educational Psychology and Counseling
National Taiwan Normal University

Shin Teng   Chia-Feng Lu
Department of Biomedical Imaging and Radiological Sciences
National Yang-Ming University

This study was designed to explore the composite scores under the multivariate distribution of latent proficiencies using a procedure based on IRT for empirical data. The purpose was to employ the IRT models under the multivariate ability distribution in order to investigate the composite scores with the consideration that correlations existed among the examinees' proficiencies. This research followed Kolen, Wang, and Lee's (2012) method for the development of the IRT-based procedure. The five tests of the Basic Competence Test were used, with a random sample of 5,000 examinees obtained in 2008. The analyses included the descriptive statistics and frequency distributions of the composite scores, overall SEM and reliability values, as well as the CSEMs. The findings indicated that there existed strong relationships between pairs of the proficiencies of the five tests and the results of Kolen et al.'s modeling were satisfied. For large-scale testing programs that rely on two or more tests to make high-stakes decisions, the issue of combining individual scores into a single total score is critical. Through the employment of Kolen et al.'s model for deriving the composite scores under the multivariate distribution, and along with this research utilizing the real data, the results of this study have revealed more of the psychometric features of the composite scores and have also helped lay a more solid foundation for many studies to embark on the estimation of examinees' ability levels in the multivariate proficiency setting via IRT.

KEY WORDS: composite score, IRT, multivariate proficiency distribution, measurement error, test battery

In the testing environment where a test battery is administered and a single score serves as an indicator of an examinee's overall performance on the test battery, the individual test scores of the battery may be summed or averaged to form a composite scale score or simply, a composite score. The ACT Assessment provides a classic example in which the composite score is the average of the four ACT Assessment scale scores (Kolen & Brennan, 2004). The Basic Competence Test (or BCTEST) administered in Taiwan is another example where its five test scale scores are summed to create the composite (Chang, 2008). When the selection decision is based on the composite scores, a general procedure is that a high score on one test can compensate for a low score on another test.

Thorough reviews of combining and/or weighting test components can be found in the literature of both the conventional test theory (Gulliksen, 1950; Wang & Stanley, 1970) and the item response theory (IRT; Lord, 1980). From the conventional test theory perspective, Gulliksen's (1950) discussion of combining the tests is by far the most comprehensive. In the absence of an external criterion to be used in evaluating the results of the composite scores, the goal of the approaches Gulliksen described was to maximize the reliability of the composite scores. However, even if different weighting schemes are possible for attempting to establish the best composite scores possible, the separate weighted test scores are still being summed up linearly from the separate test components. Each test is still being treated independently and the correlation issue is not taken into account.

The IRT environment has also provided theoretically and conceptually appealing procedures for scoring the composite scores. Instead of true test scores, ability or proficiency estimates, or thetas, are obtained in IRT. Combining test components by simply adding raw scores together can be criticized in that it fails to recognize differences in variation across domains and items, as well differences in item importance (Rudner, 2001). By employing the IRT models, items can be simultaneously calibrated across all test components; an examinee's ability is estimated on the composite scale (Rudner). Many of the ability estimation tasks have been conducted using the concurrent estimation method.

Through the simultaneous calibration of the items in IRT, the problem of obtaining the raw number of score points over the separate tests might be prevented. However, calibrating the items simultaneously to derive theta estimates still fails to take into account the correlations among the individual test components. It is likely that separate tests are related to one another and correlations/intercorrelations exist among the examinees' proficiencies underlying the tests. Simultaneous calibration of the items from different test contents does not consider the possible relationships of the tests, even if the ability estimates are being placed on the same continuum and "optimal" composite scores can be formed on this scale.

The problem of IRT modeling not having incorporated the correlations among the component parts of a test battery is present, big, but is not easily solved due to the limitation that methods are not available for estimating the multivariate distributions of the scores on the latent composite trait scales to accommodate the intercorrelations of different tests; therefore, while computing based on IRT, most composite scores of a test battery are still being formed by assuming the independency across individual test components (Kolen, Wang, & Lee, 2012). Kolen et al. specifically pointed out that very little research has been conducted on estimating the conditional standard errors of measurement (or CSEMs) for the composite scores; they stated that the previously developed procedures for computing the composite score CSEMs based on classical test theory and related methods such as those of the IRT-based found no ways of obtaining the CSEM values conditional on composite true scale score or latent ability. This is due to no approaches being available to estimate the multivariate distributions of the composite scale scores on these latent variables.

To overcome this problem, Kolen et al. (2012) presented a general IRT-based procedure in which the multivariate proficiency distribution was established and the CSEMs and the reliability of the composite scores could be attained in association with the latent composite proficiency. By following Kolen et al.'s general procedure, a multidimensional IRT model can be employed to provide a representation of the multivariate probability distribution of item responses so the CSEMs and reliability of the composite scores can be estimated under this multivariate setting.

In addition to the benefit of taking into account the correlations/intercorrelations of separate test scores of the examinees, one great advantage of implementing Kolen et al.'s (2012) IRT modeling procedure is that the multivariate distribution offers a foundation that is applicable to a wide range of item types. Kolen et al. provided examples of the dichotomously and polytomously scored items, such as the constructed-response items, essay items and multiple true-false choice items, both unidimensional and multidimensional IRT models, and composite scores of many different combinations and/or scoring and weighting situations. Meanwhile, the problem with previously developed methods only being appropriate when composite scores are constructed via linear transformation is also solved. Kolen et al. also said that another advantage gained by using their IRT modeling is that the procedure takes account of rounding error.

In spite of the enticing attributes of Kolen et al.'s (2012) method, it is necessary to be aware of one important aspect of utilizing this IRT approach. Kolen et al. stressed that for such a method in estimating the multivariate distribution via IRT, the procedures are built upon stronger statistical assumptions and are more computationally intensive than other related procedures. The IRT assumption of local independence or conditional independence is made. Kolen et al. asserted again the meaning of the conditional independence assumption provided in Lord (1980) that conditional on the vector of abilities, $\underset{\sim}{\theta}$, the item responses of the examinees are independent. In addition, Kolen et al. emphasized that their IRT-based procedures should be considered for use in testing settings where IRT models are found to adequately fit the data of the test scores.

On the other hand, while a multivariate IRT model has a strong appeal in its own right, Kolen et al. (2012) particularly pointed out that the dimensional structure from the item parameter calibration based on a multivariate IRT model may not align well with the test structure of the test battery. They acknowledged the use of the unidimensional IRT model in some circumstances for a better representation of a latent proficiency structure with items for each of the tests loading clearly on one of the different dimensions. Kolen et al. tailored their general IRT-based procedure to a special case where a unidimensional IRT was fit to each of the separate tests and item parameters were calibrated separately for each test while also allowing examinees' proficiencies underlying these tests to be correlated.

This special case of the IRT-based procedure from Kolen et al. (2012) was adopted in the present study for establishing the multivariate distribution of latent proficiencies for the explorations of the composite scores. While each test in the battery was modeled by a single IRT proficiency, intercorrelations among the examinees' proficiencies were allowed to exist. Specifically, the BCTEST assessment was used for the development and application of the procedure. With the BCTEST data possessing unique, distinct score features of its own (both statistical and psychometric properties), how well could Kolen et al.'s procedure act for the empirical data of the BCTEST? In spite of the IRT-based approach promising a great number of fascinating features, its design and procedure are complicated and the implementation is computationally intensive. Now that the BCTEST includes a large number of the scale score points along with more individual test components of greater test length, the computational work could be expected to be even more tedious. This ought to be one practical aspect worthy of more information. Another issue of interest in the current research would be whether the CSEMs obtained through the IRT-based procedure for the BCTEST composite scores would also be approximately equal along the score scale. It is necessary to understand the extent to which the CSEMs are similar across values of the composite scores under the multivariate proficiency distribution context. How well could the IRT modeling method improve over the current equally-weighted model in measurement effectiveness based on classical test theory? Having recognized the importance of the composite score issues within the multivariate proficiency framework, a closer look at the performance of Kolen et al.'s IRT-based approach with more empirical data would be needed as well as worthwhile. The results would be valuable for better justifying the employment of the IRT-based method under the multivariate latent composite proficiency context.

The Purpose

This study was designed to explore the composite scores under the multivariate distribution of latent proficiencies using a procedure based on IRT for empirical data. A computational routine was programmed for establishing both the multivariate ability distribution of the examinees and an IRT-based procedure to examine the composite scores that combined the individual components of the test battery. The purpose was to employ the IRT models under the multivariate ability distribution $\psi(\underset{\sim}{\theta})$, in order to investigate the composite scores with the consideration that intercorrelations existed among the proficiencies of the examinees, and also to estimate and evaluate the CSEMs and other measurement properties of the composite scores. This research followed Kolen et al.'s method (2012) for the development of the IRT-based procedure. This study also incorporated the additional case of the no-correlation distribution for further investigation and comparison.

Specifically, this study attempted to achieve the following objectives:

1. to prepare the multivariate proficiency distribution in which correlations among the abilities underlying the separate test components of a test battery were recognized;

2. to establish the IRT-based procedure proposed by Kolen et al. (2012) for the explorations of the composite scores;

3. to perform Kolen et al.'s IRT-based procedure under the multivariate ability distribution using the empirical data; and

4. to evaluate and compare the statistical and psychometric features of the composite scores resulting from Kolen et al.'s IRT modeling and the no-correlation setting.

Method

The Data

This study employed the BCTEST assessment to implement and explore the IRT-based approach for examining the composite scores of the examinees under the multivariate proficiency distribution. The BCTEST is a national standardized test that measures educational achievement in Chinese (48 items), English (45 items), Mathematics (34 items), Natural Science (58 items), and Social Studies (63 items) in the context of the junior high school curriculum in Taiwan (Chang, 2006). All tests are comprised of multiple-choice items and are scored by the number correct. The raw scores of each test were converted into integer scale scores of 1 to 60 using the arcsine transformation procedure (Kolen & Hanson, 1989; Petersen, Kolen, & Hoover, 1989) with the purpose of stabilizing the measurement error variance (or in other words, equalizing the measurement precision) along the scale score continuum. Studies have demonstrated the similarity of the CSEMs along the scale score continuum for each individual test of the BCTEST based on classical test theory (Chang). The BCTEST composite scale score equals the sum of the five test scores. A random sample of 5,000 examinees drawn from the data obtained in 2008 was used.

The Process

This research proceeded by following the approach proposed by Kolen et al. (2012) for establishing the IRT framework of multivariate ability distribution for the explorations of the composite scores, while adapting to the current testing situation of the employment of the BCTEST real data obtained in 2008. The process is as follows. Due to the intricate nature of the IRT-based method, a comprehensive procedure may not be easily presented here. Therefore, the computational steps laid out below might only be brief.

Step 1. The item parameters for each of the five BCTEST tests were calibrated, respectively, based on the three-parameter logistic IRT model (or the 3PL model) using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). The five BCTEST tests are Chinese, English, Mathematics, Natural Science, and Social Studies. In this application, each of the five BCTEST tests was modeled by the 3PL IRT model, assuming that examinees' performance on each test could be described by a single IRT proficiency or ability. Each of the five tests was separately modeled by IRT but the examinees' abilities underlying the five tests were allowed to be correlated.

Step 2. The multivariate proficiency distribution $\psi(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ or $\psi(\underset{\sim}{\theta})$ (θ_1 here refers to the Chinese proficiency, for example) was established based on the EM algorithm presented in Mislevy (1984) for the $\psi(\underset{\sim}{\theta})$ estimation. Also in this step, the correlations among the five proficiencies were estimated according to the description in the Appendix of Kolen et al. (2012). The procedural steps of the EM algorithm are as follows.

Based on the EM algorithm (Mislevy, 1984), the multivariate density function of proficiency $\psi(\underset{\sim}{\theta})$, took on the form of a Gaussian distribution in the current study and was hereafter denoted as $G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ or $G(\underset{\sim}{\theta})$. Specifically, Let $G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ be the latent Gaussian distribution for person *i* with response vectors $\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}$.

The likelihood function can be represented as

$$\begin{aligned}&L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)\\&=L_{i1}(\underset{\sim}{X}_{i1}\,|\,\theta_1)L_{i2}(\underset{\sim}{X}_{i2}\,|\,\theta_2)L_{i3}(\underset{\sim}{X}_{i3}\,|\,\theta_3)L_{i4}(\underset{\sim}{X}_{i4}\,|\,\theta_4)L_{i5}(\underset{\sim}{X}_{i5}\,|\,\theta_5).\end{aligned}$$

The marginal likelihood function is

$$\begin{aligned}&h(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5})\\&=\iint_{\theta_1\cdots\theta_5} L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)d\theta_1\cdots d\theta_5.\end{aligned}$$

For a given sample of *N* persons, the total log likelihood is

$$\log L=\sum_{i=1}^{N}\log\iint_{\theta_1\cdots\theta_5} L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)d\theta_1\cdots d\theta_5.$$

Maximizing the above function under the condition

$$\iint_{\theta_1\cdots\theta_5} G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)d\theta_1\cdots d\theta_5=1$$

provides the solution

$$\begin{aligned}&G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)\\&=\frac{1}{N}\sum_{i=1}^{N}\frac{L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)}{\iint_{\theta_1\cdots\theta_5} L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)d\theta_1\cdots d\theta_5}.\end{aligned}$$

The above implicit equation can be solved iteratively as is proceeded with an EM algorithm. For iteration *n*+1,

$$\begin{aligned}&G^{(n+1)}(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)\\&=\frac{1}{N}\sum_{i=1}^{N}\frac{L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G^{(n)}(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)}{\iint_{\theta_1\cdots\theta_5} L_i(\underset{\sim}{X}_{i1},\underset{\sim}{X}_{i2},\underset{\sim}{X}_{i3},\underset{\sim}{X}_{i4},\underset{\sim}{X}_{i5}\,|\,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)G^{(n)}(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)d\theta_1\cdots d\theta_5}.\end{aligned}$$

The initial $G^{(n=0)}$ was specified as a latent Gaussian distribution with a mean vector of zero and a variance/covariance matrix with correlations of .30 between pairs of the IRT proficiencies of the five tests.

Step 3. A random multivariate normal variable $\underset{\sim}{\theta}$, was drawn from the multivariate proficiency distribution $G(\underset{\sim}{\theta})$, prepared in step 2.

Step 4. Based on the item parameters obtained in step 1 and the given $\underset{\sim}{\theta}$, the conditional distribution of number-correct scores (i.e., the conditional probability distribution of the raw scores for each test given the value of each θ) was calculated using the recursive algorithm described by Lord and Wingersky (1984). This step was proceeded with each of the five thetas in the random multivariate normal variable attained in step 3, starting with θ_1, the Chinese proficiency. The formula of Lord and Wingersky is as follows.

$$\begin{aligned}\Pr(X_r = i \mid \theta) &= \Pr(X_{r-1} = i \mid \theta)\left[1 - p_r(\theta)\right] && i = 0\\ &= \Pr(X_{r-1} = i \mid \theta)\left[1 - p_r(\theta)\right] + \Pr(X_{r-1} = i-1 \mid \theta) p_r(\theta) && 0 < i < r\\ &= \Pr(X_{r-1} = i-1 \mid \theta) p_r(\theta) && i = r,\end{aligned}$$

where r is the r^{th} -item number of the test.

Step 5. The number-correct scores were converted into scale scores. In this step, the raw scores of each test were transformed into the scale scores using the arcsine function of

$$s(i) = \frac{1}{2}\left\{\sin^{-1}\sqrt{\frac{i}{K+1}} + \sin^{-1}\sqrt{\frac{i+1}{K+1}}\right\},$$

where i is the raw score, K is the number of items in the test, and sin^{-1} is the arcsine function with its arguments expressed in radians. The arcsine transformed scores $s(i)$, were then linearly converted to a scale having the mean of 30 and a maximum score of 60. Because the test scale scores for reporting were designed to begin with the starting point of one, the computed transformed scale scores falling below one were truncated. The scale scores of each test of the BCTEST all ranged from one to 60.

Step 6. The distribution of scale scores conditional on ability was produced by following Equations (2) and (3) in Kolen, Zeng, and Hanson (1996). This step produced $G_1(S_1 \mid \theta_1)$, $G_2(S_2 \mid \theta_2)$, $G_3(S_3 \mid \theta_3)$, $G_4(S_4 \mid \theta_4)$, and $G_5(S_5 \mid \theta_5)$, respectively. Following the equations of Kolen et al. using θ_1 as an example, the true scale score at θ_1 can be attained via $\xi(\theta_1) = \sum_{i=0}^{K} s_1(i)\Pr(X = i \mid \theta_1)$ and the conditional measurement error variance of scale scores at θ_1 can be achieved via $\sigma^2[s_1(X) \mid \theta_1] = \sum_{i=0}^{K}[s_1(i) - \xi(\theta_1)]^2 \Pr(X = i \mid \theta_1)$. The CSEM of scale scores was computed by taking the square root of the latter equation for the error variance.

Step 7. For the given θ_1, θ_2, θ_3, θ_4, and θ_5, the probabilities were calculated for each of the 296 composite scores, as expressed as $g(S_C \mid \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. S_C stands for the composite scale score of the BCTEST, which is a linear function of the five test scale scores. Although there were only 296 distinct composite scale scores, ranging from 5 to 300, there existed $60^5 = 777{,}600{,}000$ possible combinations of the five scale scores for the five individual tests.

Specifically, the following six steps laid out on page 8 of Kolen et al. (2012) were followed.

1. Use the $G_1(S_1 \mid \theta_1)$, $G_2(S_2 \mid \theta_2)$, $G_3(S_3 \mid \theta_3)$, $G_4(S_4 \mid \theta_4)$, and $G_5(S_5 \mid \theta_5)$ prepared in step 6.

2. Set the initial value of $g(S_C \mid \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ equal to zero for all 296 values of s_C, where s_C represents a particular value of the composite scale score.

3. For a set of scores on the five tests (say beginning with a 1 on each test), compute the product of $g_1(s_1 \mid \theta_1) g_2(s_2 \mid \theta_2) g_3(s_3 \mid \theta_3) g_4(s_4 \mid \theta_4) g_5(s_5 \mid \theta_5)$, where s_1, s_2, s_3, s_4, and s_5 are particular values of the scale scores on each of the five tests.

4. Compute s_C for the set of scores on the five tests in step (3).

5. For s_C computed in step (4), increment $g(S_C = s_C \mid \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ by the value computed in step (3).

6. Repeat steps (3), (4), and (5) for all $60^5 = 777{,}600{,}000$ possible combinations of the five scale scores for the individual tests. Again, although there were only 296 distinct composite scale scores, $777{,}600{,}000$ combinations of the five scale scores were possible.

Step 8. In this step, both the true/expected composite scale score and the CSEM conditional on the five proficiencies were obtained. Based on Equations (2) and (3) in Kolen et al. (2012), the true composite scale score at $\underset{\sim}{\theta}$ was calculated using $\xi(\underset{\sim}{\theta}) = \mathbf{E}[s(\underset{\sim}{X} \mid \underset{\sim}{\theta})]$ and the conditional measurement error variance of composite scores was computed through $\sigma^2(s(\underset{\sim}{X} \mid \underset{\sim}{\theta})) = \sum_{\underset{\sim}{X}}^{5}\left[s(\underset{\sim}{X}) - \xi(\underset{\sim}{\theta})\right]^2 \Pr(\underset{\sim}{X} \mid \underset{\sim}{\theta})$. The CSEM at $\underset{\sim}{\theta}$ was produced by taking the square root of the equation for the error variance. The process of employing the given $\underset{\sim}{\theta}$ attained in step 3 completed the first replication.

Step 9. Return to step 3, another random multivariate normal variable was drawn again from the multivariate proficiency distribution $G(\underset{\sim}{\theta})$, and the same process was followed until step 8. The computational routine from steps 3 to 8 were repeated 10,000 times.

Step 10. For each replication, the CSEMs were regressed on the true composite scale scores from each replication (obtained in step 8) using nonlinear regression procedures.

Step 11. The reliability of the composite scale scores, $rel_{composite}$, was computed by using Formulas (4) to (7) on page 5 of Kolen et al. (2012). Adapting to the design of this study, Kolen et al.'s formulas can be expressed as follows.

The average error variance is $\sigma^2(E_s) = \sum_{\underset{\sim}{\theta}=1}^{10000} \sigma^2\left[S(\underset{\sim}{X} \mid \underset{\sim}{\theta})\right]G(\underset{\sim}{\theta})$.

The mean of the scale scores for the population is $\mu_S = \sum_{\underset{\sim}{\theta}=1}^{10000} \xi(\underset{\sim}{\theta})G(\underset{\sim}{\theta})$ and the variance of scale scores for the population is $\sigma_S^2 = \sum_{\underset{\sim}{\theta}=1}^{10000} G(\underset{\sim}{\theta})\left\{\sum_{\underset{\sim}{X}=5}^{300}[S(\underset{\sim}{X}) - \mu_S]^2 \Pr(\underset{\sim}{X} \mid \underset{\sim}{\theta})\right\}$.

The reliability of composite scores is $rel_{composite} = 1 - \dfrac{\sigma^2(E_S)}{\sigma_S^2}$.

The Comparative Study

In this research, an additional case of the no-correlation was conducted in order to offer comparative information about the performance of the IRT approach of Kolen et al. (2012). In this no-correlation setting, the same process was followed as specified above, except that instead of being correlated with one another, the five tests of the BCTEST were not related in any degree. Specifically, in step 2 of the procedure, a multivariate Gaussian distribution was also created for this latent composite proficiency context of the no-correlation, but the correlation values among the five tests were set to zero. Then, the random multivariate normal variable to be drawn in step 3 was selected from this no-correlation distribution.

Analyses

In the present study, a unidimensional IRT model was assumed to hold for each of the five BCTEST tests, but the proficiencies underlying the five tests were allowed to be correlated. A random sample of 5,000 examinees was used in the analyses. Overall, the descriptive statistics of the four moments (i.e., mean, *SD*, skewness and kurtosis) of the raw scores were reported for each of the five tests, and their frequency distributions were plotted. Both the summary statistics and the frequency distributions of the actual BCTEST composite scores were also presented.

The multivariate proficiency distribution was attained and the composite scale scores were estimated from the IRT modeling procedure. The intercorrelations between pairs of these IRT ability estimates were obtained based on the EM procedure in step 2 of the process section. The results of the iterations proceeding in this step with the EM algorithm were presented. The marginal BCTEST composite scale score distribution was formed, and both the summary statistics and the psychometric features of the composite score distribution were attained. The overall standard error of measurement (SEM) and the reliability values of the composite scores were computed, as well as their CSEMs by true composite scale score. Then, the curve of the composite score CSEMs was displayed and the extent to which the CSEMs were similar across values of the composite scores was examined. The same analyses were also conducted for the no-correlation condition. The results of both the IRT modeling and the no-correlation case were evaluated and compared.

Results and Discussion

Table 1 presents the raw score summary statistics for the various tests based on the random sample of 5,000 examinees. The various tests were composed of different numbers of four-choice items so that the total raw scores were not the same among the tests. Reported in parentheses under the respective values of the means and *SD*s in the table are the means and *SD*s in proportion-correct raw score units. It can be detected that the five tests of the BCTEST possessed different score characteristics. The Kuder Richardson 20 (or the KR20) coefficients were also reported in Table 1 for the various tests. The English test had the highest KR20, followed by Social Studies and Natural Science. The Mathematics test had the lowest KR20.

Table 1 The BCTEST Raw Score Summary Statistics for the Various Tests

	No. of items	Mean[a]	*SD*[b]	Skewness	Kurtosis	KR20
Chinese	48	32.194 (0.671)	10.628 (0.221)	-0.444	2.082	0.932
English	45	30.377 (0.675)	13.149 (0.292)	-0.421	1.572	0.966
Mathematics	34	21.672 (0.637)	8.148 (0.240)	-0.378	1.900	0.917
Natural Science	58	34.950 (0.603)	14.165 (0.244)	0.065	1.667	0.952
Social Studies	63	43.664 (0.693)	14.169 (0.225)	-0.480	2.059	0.953

Note. KR20 = Kuder Richardson 20.
[a]The values in parentheses are means in proportion-correct raw score units.
[b]The values in parentheses are *SD*s in proportion-correct raw score units.

The differences in the raw scores of the tests can also be observed in Figure 1. The raw score frequencies plotted in Figure 1 show that the five tests distributed differently from one another. Both the Chinese and Social Studies tests were negatively skewed, and the three tests of English, Mathematics and Natural Science displayed two modes on the scale.

Table 2 shows the summary statistics of the composite scale scores of the five tests of the BCTEST. The upper most portion of the table contains the results based on the actual composite scores of the random sample of the 5,000 examinees, which were the conventional linear combinations of the arcsine transformed scaled scores obtained in step 5 of the process section. The middle part of the table unveils the outcomes for Kolen et al.'s (2012) IRT-based approach and the bottom portion presents for the no-correlation setting (also under the latent composite proficiency context, but with no correlations). These results for both Kolen et al.'s IRT modeling and the no-correlation case were produced through 10,000 replications.

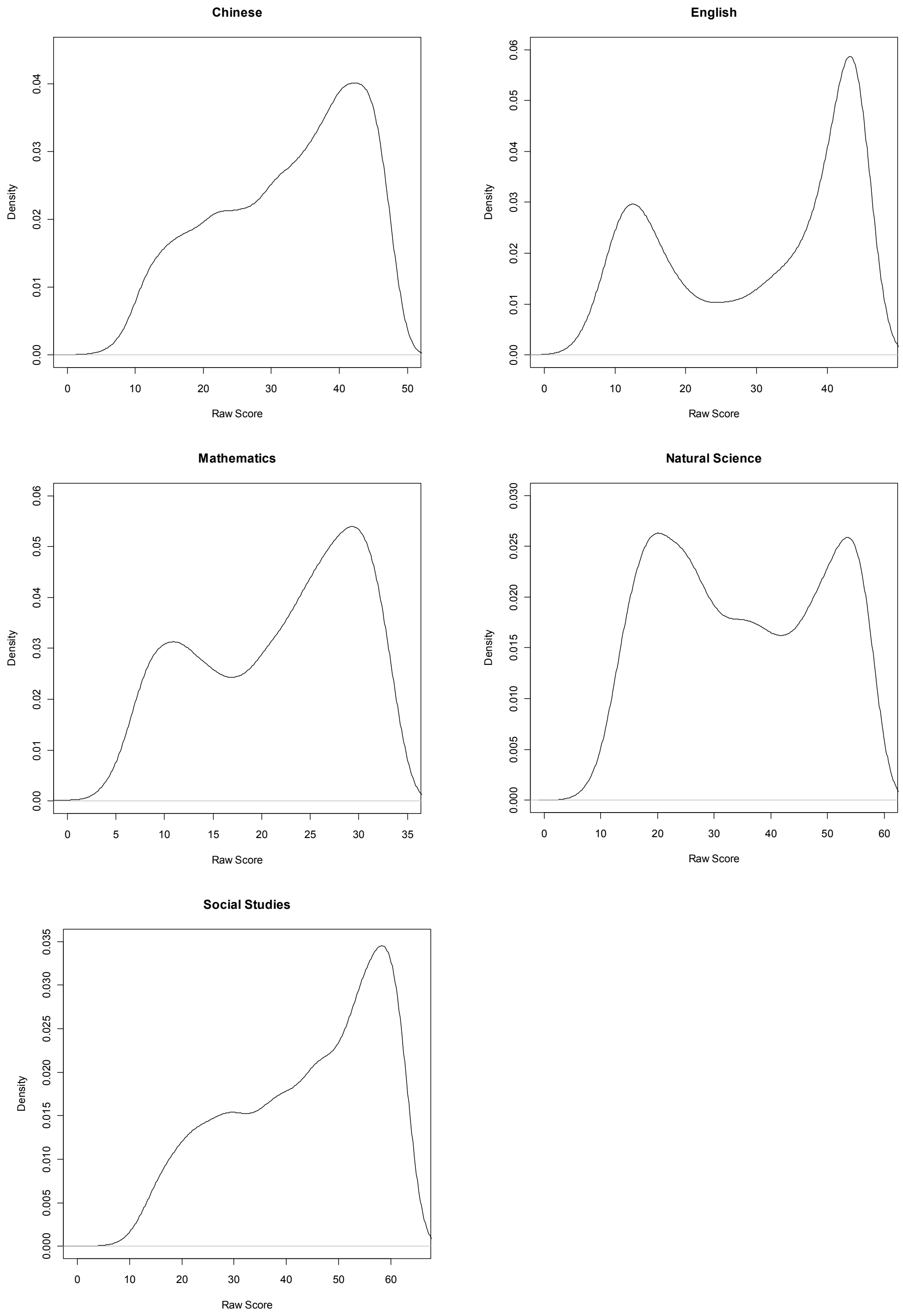


Figure 1 BCTEST raw score distributions for the various tests.

Table 2 Summary Statistics and the Overall SEM and Reliability for the BCTEST Composite Scores of the Five Tests

N	Mean	*SD*	Skewness	Kurtosis	SEM	Reliability
The actual BCTEST composites						
5000[a]	149.01	78.21	-0.0541	1.78		
Kolen, Wang, and Lee's IRT modeling						
10000[b]	145.42	76.27	0.1636	1.82	8.55	0.9874
No-correlation						
10000[b]	145.36	36.87	0.0762	2.72	8.53	0.9465

Note. Composite scale scores range from 5 to 300.

[a]The actual sample consisted of 5,000 examinees.

[b]These results were produced through 10,000 replications.

Overall, the descriptive features of the composite scores based on the IRT modeling were closer to the actual BCTEST composite scores than were those of the no-correlation to the actual BCTEST composite scores, except for the skewness value (see Table 2). The outcomes of the no-correlation were especially different from those of the actual composite scores for both the scale score *SD* and kurtosis. It can be found in Table 2 that the variability of the composite scores resulting from the no-correlation condition was much smaller than either from the actual or the IRT approach context.

Table 2 also shows the overall SEMs and the reliability coefficients of the composite scores of the five tests, which were computed by using Kolen et al.'s formulas (2012). It can be seen that while both the IRT and no-correlation cases yielded similar SEMs, the reliability of .9874 for the IRT modeling was higher than the value of .9465 for the no-correlation condition.

The frequency distributions of the BCTEST composite scores composed of the five test scale scores are illustrated in Figure 2. This figure contains the results for the three different contexts of the actual composite scores, Kolen et al.'s (2012) IRT modeling (as appeared as the IRT modeling in the figure) and the no-correlation. The plots were smoothed by the polynomial curve fitting in a degree of 10. For the actual composite scores, it can be seen that the scores spread over the entire continuum with two apparent modes; one mode occurred at the lower end of the distribution around composite scores below 50 and the other around 200 to scores slightly above 250. For the IRT modeling, the curve showed fairly evenly distributed composite scores along the scale except for one mode at the low end.

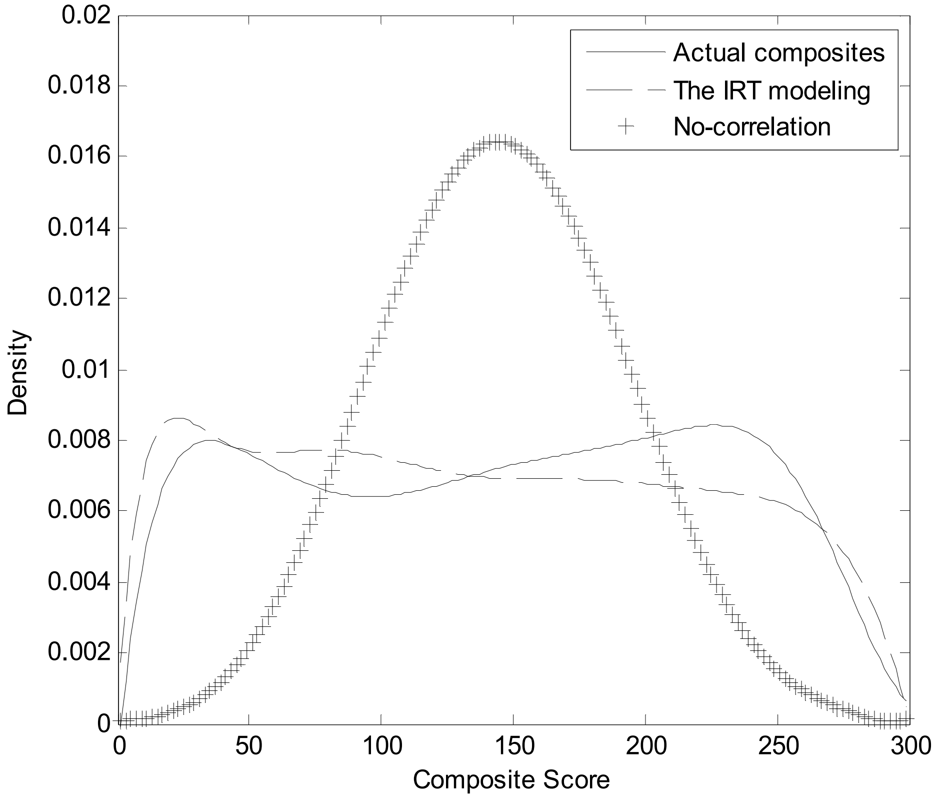


Figure 2 The frequency distributions of the BCTEST composite scores under the various modeling conditions.

The results of the analyses concerning the no-correlation case are also revealed in Figure 2 for more investigation and comparison. While also in the latent multivariate framework but with zero correlations between pairs of the five IRT proficiencies, the distribution under the no-correlation circumstance appeared in a distinctly different manner from both the actual composite scores and the IRT modeling. It can be noticed that there were many more composite scores gathering around the middle part of the scale for the no-correlation than for the IRT modeling distribution. Such a phenomenon might be explainable in that the scale scores of the individual tests for the no-correlation case were not correlated so any combinations of the test scale scores were possible. Composite scores in the middle region of the scale could be formed by either scores of low and high values or scores of average values so the mid-scores would be the mostly established composite scores under the no-correlation context. It follows that there occurred many more mid-composite scores than the composite scores at the two ends. For those test scale scores under the IRT modeling, however, they were correlated so only scores that were close in value were more likely to belong to the same combination. This reduced the chance of the combinations for the middle area of the composite score continuum. As observed in Figure 2, there existed relatively fewer composite scores over the middle part of the scale for the IRT approach than for the no-correlation case.

Table 3 reports the correlation values between pairs of the IRT proficiencies of the five BCTEST tests, which were attained via the EM procedure in step 2 of the process section. The converging status of the iterations in the EM stage is shown in Figure 3. The process appeared to converge well. The entire EM procedure took approximately four computing hours and was completed after 18 iterations. In Table 3, it can be seen that the five IRT proficiencies of the BCTEST were correlated with one another rather strongly; many of the pairs reached the coefficients above .90. The Chinese test had strong correlations with each of the three tests of English, Natural Science, and Social Studies. Both Mathematics and Natural Science and both Natural Science and Social Studies were also highly-correlated pairs. The outcomes in Table 3 reveal strong relationships of the proficiencies among the five BCTEST tests.

Table 3 Correlations Between Pairs of the IRT Proficiencies of the Five BCTEST Tests

	Chinese	English	Mathematics	Natural Science	Social Studies
Chinese	1				
English	0.9127	1			
Mathematics	0.8713	0.8893	1		
Natural Science	0.9081	0.8669	0.9046	1	
Social Studies	0.9498	0.8860	0.8713	0.9312	1

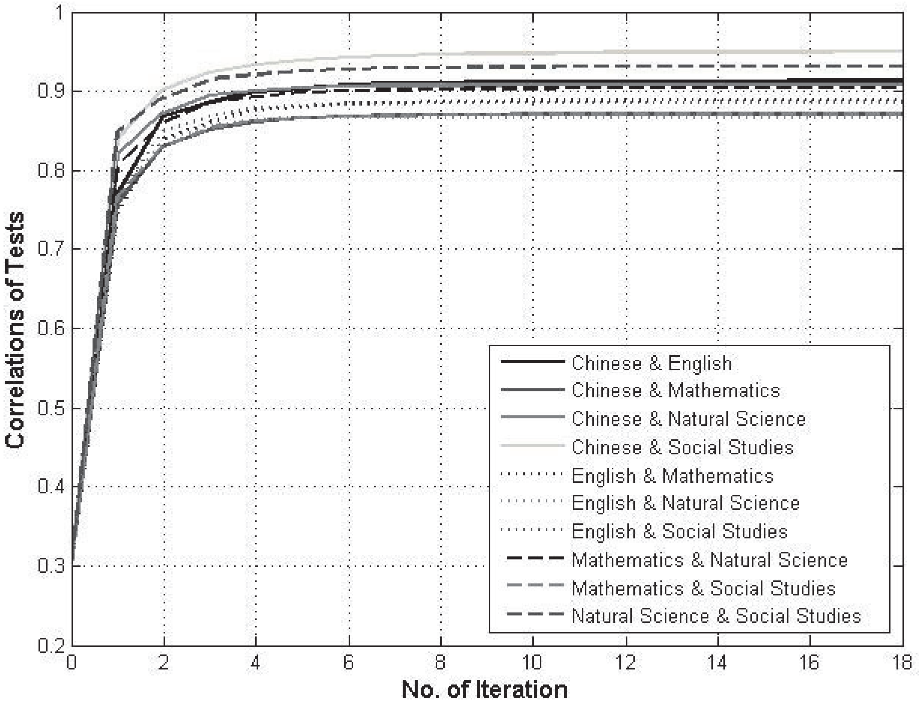
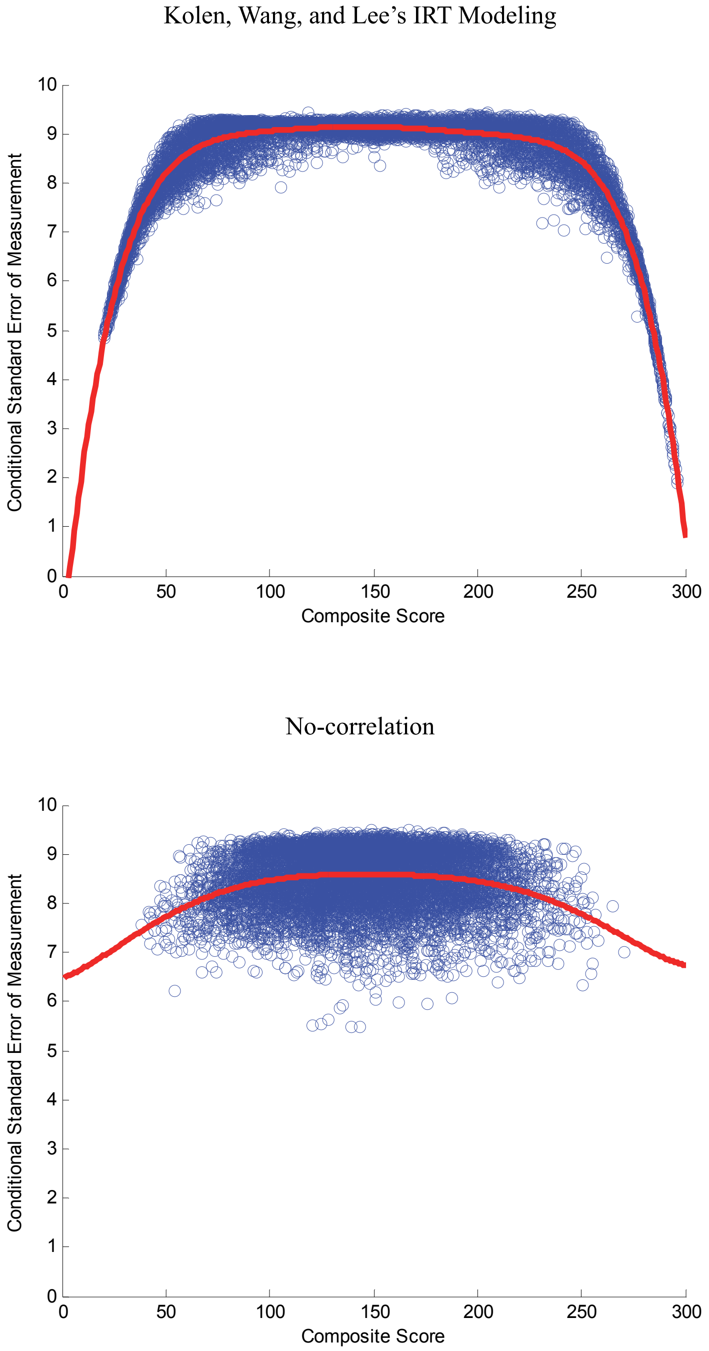


Figure 3 Results of the iterations during the EM process.

The CSEMs were graphed by true, or expected, composite scale score in Figure 4. The upper portion of the figure shows the results for Kolen et al.'s (2012) IRT procedure. As can be observed, there were scatters present over the composite score continuum. The outcome was expected because various combinations of the five proficiencies of the BCTEST could lead to the same composite score, and each combination could possibly result in a different CSEM value. This part of Figure 4 depicts that the magnitudes of the CSEMs were very similar across the true composite score scale. But, there were relatively more scatters around the lower and higher ends of the continuum than in the middle. Also, the values of the CSEMs decreased towards both extreme ends.



Kolen, Wang, and Lee’s IRT Modeling

No-correlation

Figure 4 The CSEMs of the BCTEST composite scores of the five tests, graphed by true composite scale score.

The lower part of Figure 4 portrays the CSEMs over the true composite score scale for the no-correlation setting where the correlations among the five proficiencies of the BCTEST were set to zero. It can be noticed that there seemed to be more scatters here than in the upper portion of this figure for the IRT modeling; clearly, the scatters mostly gathered around the middle. There were almost no scatters existing below the true composite scores around 50, or above the scores about 250. Such a phenomenon may be explained by reviewing the results in Figure 2 for the no-correlation condition. There, the distribution of the composite scores was that most scores occurred in the middle. Again, different combinations of the proficiencies can yield the same composite score, and each combination could potentially lead to a different CSEM. With the distributional plot for the no-correlation in Figure 2 showing most composite scores centering around the middle, there existed many more different score combinations possible over this middle range for the no-correlation than for the IRT modeling. For this reason of being associated with greater potential to have many more different score combinations, more CSEMs of different values were present and hence, more scatters appeared in this middle area of the scale under the no-correlation environment.

Summary and Conclusions

Summary of Results

In this research, IRT was used as the psychometric framework for the development of the procedure. Unidimensionality was assumed for each individual test and examinees' performance on each test can be described by a single IRT underlying trait. The underlying traits were allowed to be correlated.

Based on the random sample of the 2008 operational BCTEST testing, this study proceeded by following Kolen et al.'s (2012) mechanism for creating the IRT framework of multivariate ability distribution for the explorations of the composite scores while adapting to the current testing situation of the employment of the BCTEST. The BCTEST is a test battery including the five test subjects: Chinese, English, Mathematics, Natural Science and Social Studies. This research also incorporated the study of the no-correlation case for more investigation and comparison of the statistical and psychometric characteristics of the resulting composite scores. The no-correlation setting was established where no correlations existed among pairs of the five abilities of the BCTEST.

Overall, the summary statistics of the composite scale scores were obtained for both the actual BCTEST data and Kolen et al.'s (2012) method, and their frequency distributions were plotted. There were also the results of the overall SEM and the reliability of the composite scale scores, as well as the CSEMs conditional on true (or expected) composite scale score. In addition, the outcomes included the intercorrelations among the five IRT proficiencies of the BCTEST underlying Kolen et al.'s multivariate ability environment. The results of the analyses for the no-correlation were also presented for evaluation and comparison.

The report of the summary statistics in Table 2 showed that overall, the descriptive features of the composite scores based on Kolen et al.'s (2012) IRT modeling were closer to the actual BCTEST composite scores than were those of the no-correlation to the actual BCTEST composite scores. Also, the reliability coefficient for the IRT modeling was slightly higher than the no-correlation condition.

In Table 3, the high values of the correlation suggested that there existed strong relationships between pairs of the five IRT proficiencies for the BCTEST. The correlations of the examinees' latent proficiencies reported here were attained under the multivariate distribution via the EM procedure. The satisfactory results of the EM iteration could be found in Figure 3.

The distributional shapes for the actual BCTEST composite scores, the IRT modeling and the no-correlation setting were shown in Figure 2. The plots depicted for the actual BCTEST composite scores a spread-out distribution with each end of the scale associated with one mode. The composite scores estimated from the IRT modeling were fairly evenly distributed across the continuum whereas those under the no-correlation gathered relatively more in the middle than at the two ends of the scale.

The observations in Figure 2 suggested similarity between both the actual composite scores and the IRT modeling situations.

Figure 4 presented the curves of the composite score CSEMs for both the IRT modeling and the no-correlation condition. The horizontal axis is the true, or expected, composite score. The comparison of these two plots revealed that there were more conditional error variances for the no-correlation than for Kolen et al.'s (2012) method. It can be seen that for Kolen et al.'s approach, while there were more scatters around both ends of the scale than in the middle, there were relatively fewer scatters for this modeling mechanism than for the no-correlation. There contained more scatters in the lower portion of Figure 4 under the no-correlation situation, with most of them gathering around the middle part of the continuum. Such a phenomenon could be attributed to the evidence that most composite scores centered over the middle scale range for the no-correlation case (see Figure 2). Various combinations of the five proficiencies could yield the same composite score, and each combination has the potential to be connected with a different CSEM. With the middle part loaded with composite scores and thus more combinations likely, there appeared more scatters in this middle area under the no-correlation framework.

Conclusions

For large-scale testing programs that rely on two or more tests to make high-stakes decisions, the issue of combining individual scores into a composite score is critical. Testing agencies aim to create composite scores that better serve as an indicator of examinees' overall performance on the test battery. Previous research on composite scale scores has been fairly limited due to no methods available for the estimation of the multivariate proficiency distribution of the examinees in which the intercorrelations of the examinees' latent abilities could be accommodated. With the introduction of Kolen et al.'s (2012) IRT-based procedure, a multivariate IRT model can be implemented to provide a representation of the multivariate probability distribution of item responses. The correlations among the separate tests of a battery can be considered and the composite scores of the examinees are more accurately estimated.

The multivariate ability distribution modeled by using Kolen et al.'s (2012) IRT-based method can also be applied to a broad variety of composite score situations that better meet the requirements of the testing programs. As mentioned by Kolen et al., a great number of testing settings can be embraced to address different concerns and issues about the formation of composite scores. For example, testing situations may involve a mixture of item types and responses. Besides the dichotomously scored items, the polytomous types of items can also be incorporated into the multivariate proficiency distribution framework, such as the constructed-response items, essay items and multiple true-false choice items. Or, instead of assuming a single IRT underlying trait to describe the examinees' performance on each test, as in the case of this study, other IRT models (e.g., a multidimensional IRT) can be used with the multivariate context. Various raw-to-scale score transformation methods may also be considered. Except for the arcsine transformation procedure this research employed, both the linear transformation and the normalization methods can be attempted. Composite scores resulting from other forms of combination may also be explored. For the data that are fit to the IRT models, the flexibility of applying the multivariate proficiency distribution shall pave the way to many other alternative test-scoring systems that best suit the needs of testing programs.

However, the implementation of Kolen et al.'s (2012) general IRT-based procedure is one major undertaking. The mechanism of the model itself is rather complex. To assemble the required algorithms and details into a complete, general computational routine was no easy task. In addition, to actually execute the procedure was complicated and time-consuming as well. Besides the iterations proceeded during the EM period, the replications over the random selections of the multivariate random variable were especially computationally intensive. The tedious task of replicating over 10,000 times in step 9 of the process section of this study took a substantial amount of time to complete. It is perceivable that an enormous amount of time would be needed if, for example, there were more scale score points in a test battery. More score combinations would then follow, and greater computational effort would be involved for the entire course of the process.

Nevertheless, despite all the appealing features that Kolen et al.’s (2012) IRT approach might have to offer, when it comes to combining individual test scores, or scoring and weighting, a “not-equally-weighted” model can always be very hard to explain to the examinees and the users of the composite scores. The concern is unavoidable, as it is for the condition under the classical test theory using various weighting schemes. It is very important that the scoring, weighting methods or the way the optimal relative weights are determined be based on both the sound psychometric rationales and the nature of the tests being administered. Clear and useful guidelines should be provided prior to testing to inform about the formation of the composite scores under the IRT setting. Additionally, evaluation of the consequences of implementing Kolen et al.’s scheme must be closely followed.

With Kolen et al.’s (2012) IRT procedure having a strong appeal of providing a representation of the multivariate probability distribution for accommodating a broad variety of composite score related situations, and along with the current research utilizing the empirical real data of the BCTEST possessing unique score characteristics of its own, the results of this study have uncovered more of the psychometric attributes of the composite scores under the multivariate latent composite proficiency context. This study has also helped lay a more solid background for many studies to embark on the estimation of examinees’ ability levels in the multivariate proficiency environment via IRT. A further extension of the use of the multivariate proficiency distribution in IRT can be anticipated, and an abundance of research would flourish and advance the understanding of the many composite score related issues.

References

Chang, S. W. (2006). Methods in scaling the Basic Competence Test. *Educational and Psychological Measurement, 66*(6), 907-929.

Chang, S. W. (2008, March). *Choice of Weighting Scheme in Forming the Composite.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Gulliksen, H. O. (1950). *Theory of mental tests*. New York, NY: John Wiley & Sons.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. 2nd ed. New York, NY: Springer Science+Business Media.

Kolen, M. J., & Hanson, B. A. (1989). Scaling the ACT Assessment. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 35-55). Iowa City, IA: American College Testing Program.

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*(1), 1-20.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings”. *Applied Psychological Measurement, 8*(4), 452-461.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York, NY: American Council on Education, and Macmillan.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20*(1), 16-19.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 4*, 663-705.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International, Inc.

收稿日期：2011 年 06 月 21 日
一稿修訂日期：2012 年 09 月 17 日
二稿修訂日期：2012 年 12 月 20 日
三稿修訂日期：2013 年 02 月 22 日
接受刊登日期：2013 年 02 月 22 日

國立臺灣師範大學教育心理與輔導學系
教育心理學報，2013，45 卷，2 期，261-278 頁

運用試題反應理論計算多向度能力分配中量尺總分之探討*

章舜雯　　滕　欣　　盧家鋒

國立台灣師範大學
教育心理與輔導學系

國立陽明大學
生物醫學影像暨放射科學系

本研究使用實徵資料，探討應用試題反應理論計分模式在多向度能力分配上，對於測驗量尺總分的計分效果。本研究的目的在使用 Kolen、Wang 和 Lee (2012) 的方法建立以試題反應理論模式爲基礎的程式，在測驗組合、各學科所測量能力彼此互為相關的情境下，計算量尺總分，然後再對所得總分不同的特性進行分析與探討。本研究使用國中基測 2008 年第一次測驗 5,000 筆考生的隨機資料；評鑑的標準包含考生在多向度能力分配裡所得量尺總分之描述統計值與分數分配圖形、總誤差量、信度值，以及測量標準誤。研究結果顯示國中基測各學科所測量的能力彼此存在高度的相關，Kolen 等人的試題反應理論計分程序有令人滿意的表現。對於合併使用兩種或兩種測驗以上作影響性決定的大型測驗而言，如何將不同的測驗分數結果整合是相當重要的議題；本研究建立多向度能力分配、使用實徵資料，探討以 Kolen 等人的模式計算量尺總分的效果，相信研究的結果已提供了更多有關量尺總分的測量特性，對於日後使用試題反應理論模式，在多向度能力情境中計算考生能力的相關研究上也提供了更有力的基礎 。

關鍵詞：多向度能力分配、量尺總分、測量誤差、測驗組合、試題反應理論

* 本篇論文通訊作者：章舜雯，通訊方式：shwchang@ntnu.edu.tw。