Bulletin of Educational Psychology, 2009, 40 (3), 489-510
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Choice of Weighting Scheme in Forming the Composite*

Shun-Wen Chang

Department of Educational Psychology and Counseling
National Taiwan Normal University

The present study investigated and compared the results of establishing composite scores based on the five weighting schemes of the equally-weighted model, the reliability weighting model, the standard deviation (or *SD*) weighting model, the error of measurement weighting model, and the effective score point model. The purpose of this study was to seek optimal relative weights in forming the best composite possible, as well as to offer more information about the various weighting schemes considered on the different measurement qualities of the tests. The five tests of the Basic Competence Test (BCTEST) were employed for exploration in this research. A random sample of 5,000 examinees drawn from the data obtained from the 2005 test administration was used. To evaluate the various weighting schemes, this study examined the statistical and psychometric properties of the test scores and the weighted composite scores, the effective contributions of individual tests to the composites, as well as the impact on the admission decisions. The findings indicated that the reliability coefficients of the variously formed composite scores were all very high. However, with regard to the effective contributions, the results were very different among the various weighting schemes. Overall, the *SD* and the error of measurement weighting models seemed to perform better in establishing the composites than the reliability or the effective score point model, although there still remained the issue of inequality of the effective contributions for both the *SD* and the error of measurement models. Results from this study should advance the understanding of weighting issues while combining individual test components into composites.

KEY WORDS: composite score, composite weighting, effective contribution, effective weight, nominal weight

In the testing environment where a test battery is administered and a single overall score is used as a basis for making a decision, the individual test scores of the battery may be summed or averaged to form a composite score. The ACT Assessment provides a classic example of the weighting scheme in which the

* The present article is an extension of a paper presented at the annual meeting of the National Council on Measurement in Education, New York, March 2008.

composite score is the average of the four ACT Assessment scale scores (Kolen & Brennan, 2004). Wang and Stanley (1970) called the weight applied to a test score in forming a composite a nominal weight. For such a weighting rule where the scale scores are summed or averaged to arrive at the composite, the nominal weights for all tests are equal since the scores are all equally weighted in forming the composite. This weighting scheme belongs to what has been called the compensatory model, for which the scores on the separate tests are combined into a single composite and a high score on one test can compensate for a low score on another test. Besides this model, there is also the noncompensatory model where the scores on the separate tests are considered independently and a certain minimum score on each test must be earned for an examinee to be selected. The compensatory and noncompensatory models are the two general approaches to decision making where the decision is to be based on more than one test score (Allen & Yen, 1979; Kane & Case, 2004). But, for most high-stakes testing programs, the compensatory model is usually adopted (Kane & Case).

In a compensatory model where the scores on the separate tests are combined into a single composite and a high score on one test can compensate for a low score on another test, when the reliabilities of the tests are equal, equal nominal weights for the tests are considered the optimal relative weights (Wainer & Thissen, 2001). However, it may be encountered that the tests are not equally reliable and some calculations might be needed in order to determine the best relative weights possible for forming a test battery from different test components. For a test battery consisting of several tests or components, the numbers of items may not be the same, the raw score distributions may vary, and the intercorrelations among the tests may also be different. Also, if raw scores of the tests are converted into scale scores through nonlinear transformation of some forms, the scale score distributions would even differ from those of their original raw scores. These varying score properties would likely lead to differences in reliability among the tests.

Examinees' raw scores are often transformed into scale scores for reporting or equating purposes. Unless the transformation methods are specially designated for converting the score distributions into the same shape, scale score standard deviations (*SD*s)/variances among the tests are likely to occur. The unequal variances among the tests would result in unequal *effective weights*, as defined by Wang and Stanley (1970) as the covariances between the test scores and the composite score for a group of examinees. For tests with larger effective weights, their contributions to the composite scores are greater than those with smaller effective weights. Problems could arise when examinees who perform particularly well on a test of greater scale score variability obtain higher composite scores, whereas examinees of whom on a test of lesser variability also can do particularly well, do not seem to have such an "advantage". Vexing arguments could really center around such a kind of "unfairness" issue. Giving the same nominal weight to tests with different effective weights might not be appropriate.

Petersen, Kolen, and Hoover (1989) commented that in situations where tests have very different *SD*s, intercorrelations, or reliability, the difference between their nominal and effective weights can be substantial. The more apparent the differences between the nominal and effective weights become, the more important the issue of the individual contributions to the composite scores also becomes. But, besides simply weighting each test equally, could there be better weighting models to employ? Allen and Yen (1979) pointed out that assigning the tests the same weight is a user-chosen weighting scheme and there exists no "natural" weighting system that can be adopted to bypass this decision making. Gulliksen (1950) also indicated that as long as an overall score is to be formed from separate test scores, the weighting problem

arises. How could the composite be better established to convey more appropriate meanings about examinees' performance over the various tests? Because a composite score implies a compensatory model, carefully determined weights are necessary. More thought is needed about the formation of the composite scores.

Thorough reviews of test component weighting have been well documented in the literature of the conventional test theory (Gulliksen, 1950; Wang & Stanley, 1970). Some of Gulliksen's (1950) extensive discussion contained weighting specifically in relation to test attributes such as *SD*, intercorrelation, and reliability. In the absence of an external validity criterion to be used in determining the weights, the goal of the approaches Gulliksen described was to maximize the reliability of the composite score. One method was to weight tests according to their reliability by using the equation $r/(1-r)$, the ratio of the test reliability to the error variance of the standard scores. Following this rule, the more reliable tests would be assigned greater weights. While tests of higher reliability may seem to deserve a larger weight in determining the composite, Gulliksen indicated that there is usually no justification for the employment of such particular weights.

Another method of combining tests was to weight inversely to the *SD* of a test, which Gulliksen (1950) stated has also been frequently used. By assigning the score weight of a test as $1/s$ (the reciprocal of the *SD* of a test), a composite will be impacted to a greater extent by tests with larger *SD*s than by tests with smaller *SD*s. However, due to the fact that *SD*s can be influenced by both the true score and error score variances, and the test length, Gulliksen cautioned that weighting based on this rule is best avoided as a routine method.

Still another weighting scheme Gulliksen (1950) discussed was weighting inversely to the error of measurement. Gulliksen pointed out that weighting each test by the factor, $1/[s(1-r)^{1/2}]$, automatically corrects for any arbitrary multiplying factors introduced in the scoring scheme. He emphasized that by utilizing this rule, the weight of a test increases as the true variance or reliability is increased, and the weight of a test decreases as the error variance is increased. Gulliksen credited this weighting scheme for its excellent properties from a common-sense perspective and recommended its use when no criterion is available and when the tests appear indifferent as far as the test content is concerned. Gulliksen considered this approach an arbitrary rule of thumb method for use.

Many other weighting schemes derived from the classical test theory have also been proposed (Ma, Kim, & Walker, 2006; Pei & Maller, 2006; Wang, 1998). One method studied in Ma et al. (2006) also involved using the test component reliability for determining the relative weights in forming the composite. This method embraced the concept of *effective test length* borrowed from Livingston and Lewis (1995); it assigned equal importance or weight to each "effective score point" on the component (Ma et al.). Livingston and Lewis used *effective length* as a property of the test that is closely linked with the score precision and defined it as follows: "The effective test length corresponding to a test score is the number of discrete, dichotomously scored, locally independent, equally difficult items required to produce a total score of the same reliability" (p. 186). The idea in Ma et al. was that the more reliable the component, the greater the effective test length. More weight was also assigned to components with higher reliability, in addition to components with more score points.

Gulliksen (1950) indicated that if many tests with high intercorrelations are to be combined, different weighting systems will not make too much difference in the resulting composites; however, if few tests are

to be combined and their correlations are low, weighting can have an important effect on the composites. Allen and Yen (1979) also made similar points using predictors in a multiple regression framework as an illustration. In the case where few predictors are used in a prediction equation and the predictors are uncorrelated, weighting the various tests will make an appreciable difference, and the choice of an optimal set of weights is essential. Pei and Maller (2006) also provided evidence to show that factors such as the number of components and their psychometric properties could affect the results of the composite reliability and validity. In addition, Wainer and Thissen (2001) went so far as to state that doing complex computations to determine optimal relative weights for the components is not necessary when the two components to be combined have equal reliability; that is, different solutions to the optimal weight problem would make no difference.

However, it may not be clear in any given instance how high the intercorrelations would need to be for different weighting models to make little difference, or how many components would be considered too few so that different weighting models would impact the weighting results. While there has been a great deal of research on this component weighting issue (Carlson, 2006; Feldt, 2004; Kane & Case, 2004; Lord, 1980; Ma et al., 2006; Pei & Maller, 2006; Rudner, 2001; Wang, 1998; Yen & Candell, 1991), many of the studies were conducted based on simulated data where the data could be generated to fit the study designs perfectly. For empirically realistic data to possibly possess more unique, distinct score features of their own (both statistical and psychometric properties), how well would the comments regarding the strengths and weaknesses which were provided in the previous studies hold? Especially in situations where the test components are not scaled to have similar score distributions and their characteristics vary to some great extent, would the measurement effectiveness still be improved by using the various weighting schemes? Having recognized the importance of assigning optimal relative weights to the tests in a battery, a closer look at these weighting schemes with empirical data could help to testify to the effects of the alternate sets of weights and also, to justify the choice and uses of a weighting scheme. Revisiting the different rationales for deriving the weights using empirically realistic data could well prove to be a worthwhile endeavor.

The Purpose

This study was designed to explore various weighting schemes for forming the composite score using empirical real data. An unweighted summation over all the individual test components was compared with weighting methods introduced in the classical test theory framework. The purpose was to seek optimal relative weights to be applied to the individual tests in order to establish the best composite possible. Specifically, this study attempted to achieve the following objectives:

1. to explore the equally-weighted model with weighting approaches derived within the classical testing framework for which the test score properties of *SD*, reliability, and/or error of measurement were taken into account; and

2. to evaluate and compare both the effects of employing the various weighting schemes on the psychometric properties of the composite scores and the impact on the admission decisions.

Even though the present study has focused mainly on a particular assessment and though the findings for the relative weights might only be optimal for the assessment being studied, the ideas and insights should also hold for most large-scale testing settings in which similar issues are encountered.

Method

The Data

This study employed the Basic Competence Test (or BCTEST) data to explore the effectiveness of various weighting schemes. The BCTEST is a national standardized test that measures educational achievement in Chinese, English, Mathematics, Natural Science, and Social Studies in the context of the junior high school curriculum in Taiwan (Chang, 2006). All tests are comprised of multiple-choice items. The raw scores of each test are converted into scale scores of 1 to 60 using the arcsine transformation procedure (Kolen & Hanson, 1989; Petersen et al., 1989) and the scale scores of the various tests vary to some extent in their score distributions. The current weighting scheme for the BCTEST composite is an equally weighted combination of the scores of all five tests. A random sample of 5,000 examinees drawn from the data obtained from the 2005 test administration was used.

The Process

Based on the random sample of the 2005 operational BCTEST testing, this research proceeded by following the steps laid out below.

Step 1. The raw scores of each test were transformed into the scale scores using the arcsine function of $c(i)=\frac{1}{2}\left\{\sin^{-1}\sqrt{\frac{i}{K+1}}+\sin^{-1}\sqrt{\frac{i+1}{K+1}}\right\}$,

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, c (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. To prevent rounding errors, the transformed scale scores were preserved with decimal points while the computation was in progress.

Step 2. Based on the scale scores obtained in step 1, the relative nominal weights for the individual tests were determined according to the various weighting schemes specified below.

A. The Equally-Weighted Model (the current unweighted model that the BCTEST adopts).

Each test was assigned a weight of one.

B. The Reliability Weighting Model (weighting according to test reliability or inversely to the error variance).

The weights were calculated through the factor of $r_{XX}/(1-r_{XX})$, where r_{XX} is the reliability of test X.

C. The Standard Deviation (*SD*) Weighting Model (weighting inversely to the *SD*).

The weights were computed using the factor of $1/s_X$, where s_X is the scale score *SD*.

D. The Error of Measurement Weighting Model (weighting inversely to the error of measurement).

The weights were obtained via the factor of $1/\left(s_X\sqrt{1-r_{XX}}\right)$.

E. The Effective Score Point Model.

First, the effective test length was obtained using Equation (15) in Livingston and Lewis (1995).

Then, the effective score points were computed for the respective test components using Equations (3) and (1) in Ma et al. (2006).

The formula for computing the effective test length (Livingston & Lewis, 1995, p. 187) is as follows.

Equation (15) $n = \frac{(\mu_X - X_{\min})(X_{\max} - \mu_X) - r\sigma_X^2}{\sigma_X^2(1-r)}$, where *r* is the reliability coefficient.

$X_{\min}$ and $X_{\max}$ are the lowest and highest possible scale scores.

Equations (3) and (1) in Ma et al. are as follows.

(3) $w_{eff} = \frac{n_{eff}}{X_{\max}}$, (1) $p_i^* = \frac{w_i s_i}{\sum_i w_i s_i}$.

Hereafter, the five weighting models under step 2 shall be represented as the A, B, C, D, and E models in the order of their presentation above.

Step 3. The BCTEST composites were formed by summing over the scale scores of all five tests, with the tests being multiplied by the set of weights obtained from step 2 for the respective weighting models. The final composite scores obtained here were then rounded into integer values. Prior to this step, the scale scores remained in decimal points to preserve the best computational accuracy possible.

Step 4. Using Equation (10) provided in Kolen (2006), the reliability of the weighted composite scores formed by the various weighting methods were computed, respectively. Also, using Equation (71) described in Gulliksen (1950), the correlations between each test and the composite scores were calculated. The two formulas are presented below; for more detailed information about the calculations, see Kolen and Gulliksen.

Equation (10) in Kolen (2006, p.161) is represented as:

$\rho_C = 1 - \frac{\sum_j u_j^2 \sigma_j^2 (1 - \rho_{jj'})}{\sum_j \left[u_j^2 \sigma_j^2 + u_j \sum_{k \neq j} u_k \sigma_{jk} \right]}$, the reliability of weighted composite scores,

where u_j is the weight applied to the raw scores on each item type, σ_j^2 is the variance, $\rho_{jj'}$ is the reliability, and σ_{jk} is the covariance between test *j* and *k*.

Equation (71) in Gulliksen (1950, p. 357) is represented as:

$r_{gC} = \frac{W_g s_g + \sum_{h=1}^{K} W_h r_{gh} s_h}{s_C} \quad (g \neq h)$, the correlation of any part (*g*) with the composite (*C*),

where W_g (or W_h) is the weight assigned to any part (*g* or *h*), s_g (or s_h) is the *SD* of that part, r_{gh} is the correlation between two parts (*g* and *h*), and s_C is the *SD* of the composite.

Analyses

This research limited its scope to the case where no external criterion was available and has followed the suggestion made by Gulliksen (1950) by using the composite reliability as an indication of the measurement quality in determining the optimal weights. Specifically, Equation (10) in Kolen (2006) for computing the reliability of the weighted composite scores as well as Equation (71) described in Gulliksen for calculating the correlation between each test and the composite were considered. Overall, the

descriptive statistics of the four moments (i.e., mean, SD, skewness and kurtosis) of the raw scores, the test scale scores and the various weighted composite scores were obtained, and their frequency distributions were plotted. Both the correlation and the variance/covariance matrices of the test component scores were computed and the effective contributions of the individual tests to the composites were compared. Also, for evaluating the impact of the different weighting methods on the admission decisions, the proportions of examinees who had received composite scores of a particular point and below were calculated at each composite scale score point. The differences of the cumulative probabilities of the various weighting methods against the equally-weighted scoring model were investigated.

Results and Discussion

The Characteristics of the Score Distributions

Table 1 presents 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 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. Especially notice that the *SD* values varied among the tests, with the English test showing the greatest variability and the Social Studies test the least. In addition to the four moments of the test raw scores, the Kuder Richardson 20 (or the KR20) coefficients were reported in the table for the various tests. The English test had the highest KR20, followed by Natural Science and Social Studies. Both the Chinese and Mathematics tests possessed slightly lower KR20s.

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

	No. of items	Mean[a]	*SD*[b]	Skewness	Kurtosis	KR20
Chinese	48	32.196	10.632	-0.488	2.150	0.931
		(0.671)	(0.222)			
English	45	29.019	13.094	-0.249	1.466	0.964
		(0.645)	(0.291)			
Mathematics	33	21.238	8.340	-0.293	1.790	0.928
		(0.644)	(0.253)			
Natural Science	58	35.596	12.586	-0.011	1.825	0.939
		(0.614)	(0.217)			
Social Studies	61	42.877	12.073	-0.536	2.370	0.936
		(0.703)	(0.198)			

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. The Chinese and Social Studies test scores were negatively skewed whereas the Natural Science test scores were fairly spread out over most of the score range. The Mathematics test displayed a clear mode prior to its mean, but then showed scores ascending towards the higher side of the scale. The distribution of the English test indicated two rather distinctive modes; the scores accumulated near both ends of the scale with the higher

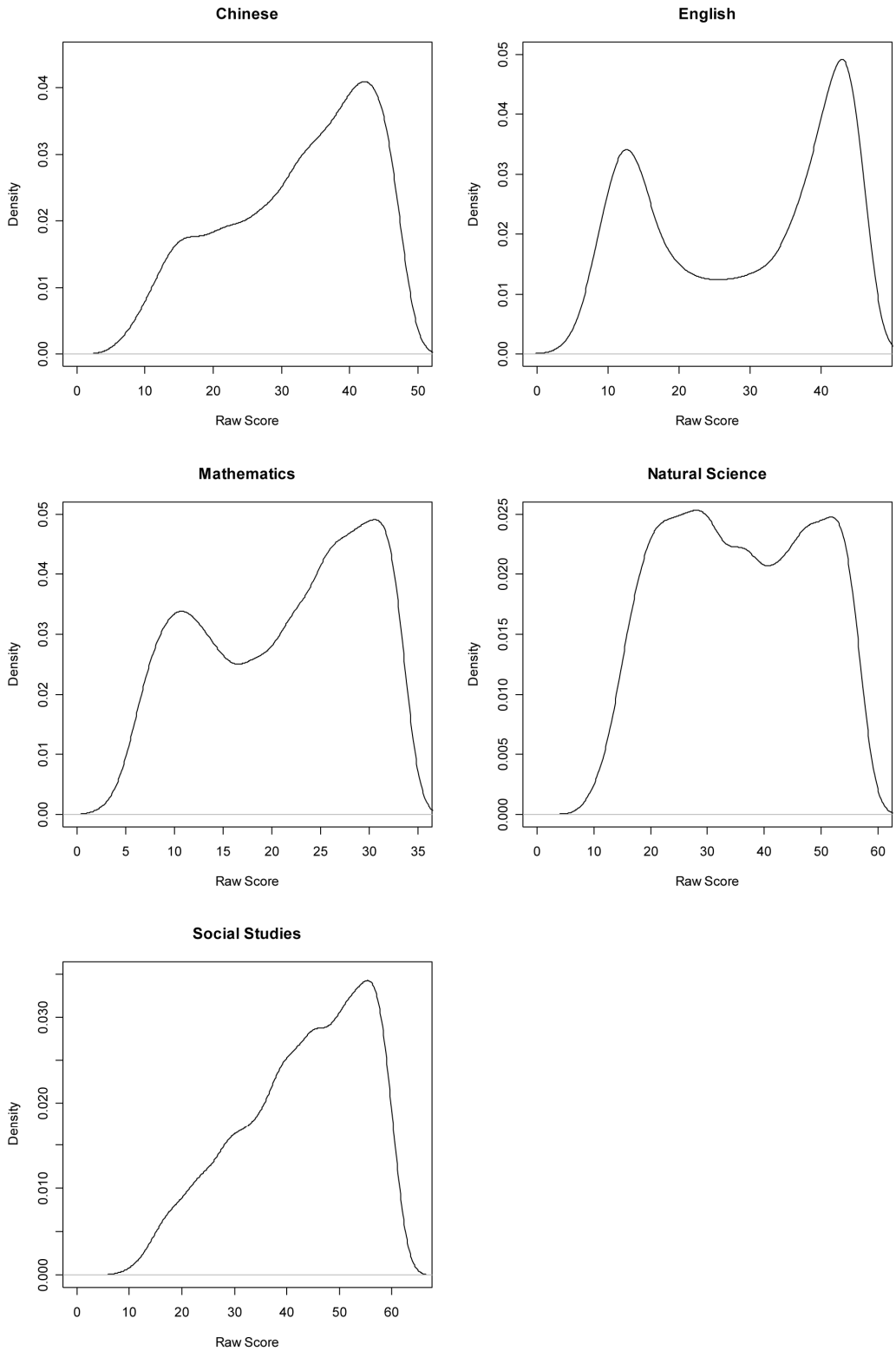

Figure 1 BCTEST raw score distributions for the various tests.

peak occurring at the upper part.

The raw scores were converted into the scale scores for each test through the arcsine transformation (Kolen & Hanson, 1989; Petersen et al., 1989). Transformed scale scores falling into the negative region (i.e., below the lower limit of one) and/or exceeding the designated value of 60 were truncated and adjusted in order to stay within the 1 to 60 score range. The arcsine transformation of the raw-to-scale scores seemed to alter the raw score distributions only slightly. Both Table 2 and Figure 2 contain the summary statistics and the distribution display for the scale scores. It can be seen that the scale score distributions varied while the *SD*s of the test scale scores were different from one another. The order in score variability still remained the same, except for the two tests with the least extent of variation.

Table 2 The BCTEST Scale Score Summary Statistics for the Various Tests

	No. of items	Mean	*SD*	Skewness	Kurtosis
Chinese	48	30.038	14.351	-0.180	2.160
English	45	30.070	19.522	-0.013	1.560
Mathematics	33	30.029	16.340	0.015	1.960
Natural Science	58	30.001	12.251	0.205	2.046
Social Studies	61	30.038	13.853	-0.159	2.256

Note. Scale scores range from 1 to 60 for each test.

The Results of Employing Weighting Schemes

Table 3 reports the results of the nominal weights of the tests for the various weighting schemes, along with their corresponding relative proportional values presented in parentheses. It can be seen in the table that employing different methods for weighting produced very different sets of relative weights among the tests. The equally-weighted model, or the A model, weighted each test equally by assigning each test a weight of one. When the reliability weighting model, or the B model, was used to calculate the weights utilizing the magnitude of reliability, the English test was connected with the highest weighting and the Mathematics test with the lowest weighting. However, the *SD* model, or the C model, employed the scale score *SD* values and produced the lowest weighting for the English test and resulted in the Natural Science test being given the greatest weighting. The error of measurement weighting model, or the D model, derived the weights via the error of measurement and seemed to have somewhat lessened the phenomenon of inequality across tests. Among the relative nominal weights, the Mathematics test was assigned the smallest weight of all. The effective score point model (i.e., the E model) employed the effective test length concept and produced the most dominating weighting for the Natural Science test, followed next by the Social Studies test. As can be seen in Table 3, the nominal weights were 1.57394 and 1.10873 for Natural Science and Social Studies, respectively, for this E model. Such results may not be surprising since the E model took into account the effective test length in computing the nominal weights for the tests. For both English and Mathematics tests of comparatively shorter lengths, the nominal weights attached to them were accordingly smaller, particularly that attached to the Mathematics test. In fact, except for the equally-weighted and the scale score *SD* weighting models, the Mathematics test appeared to be the least favored and was placed at the lowest rank in weighting among all the five tests.

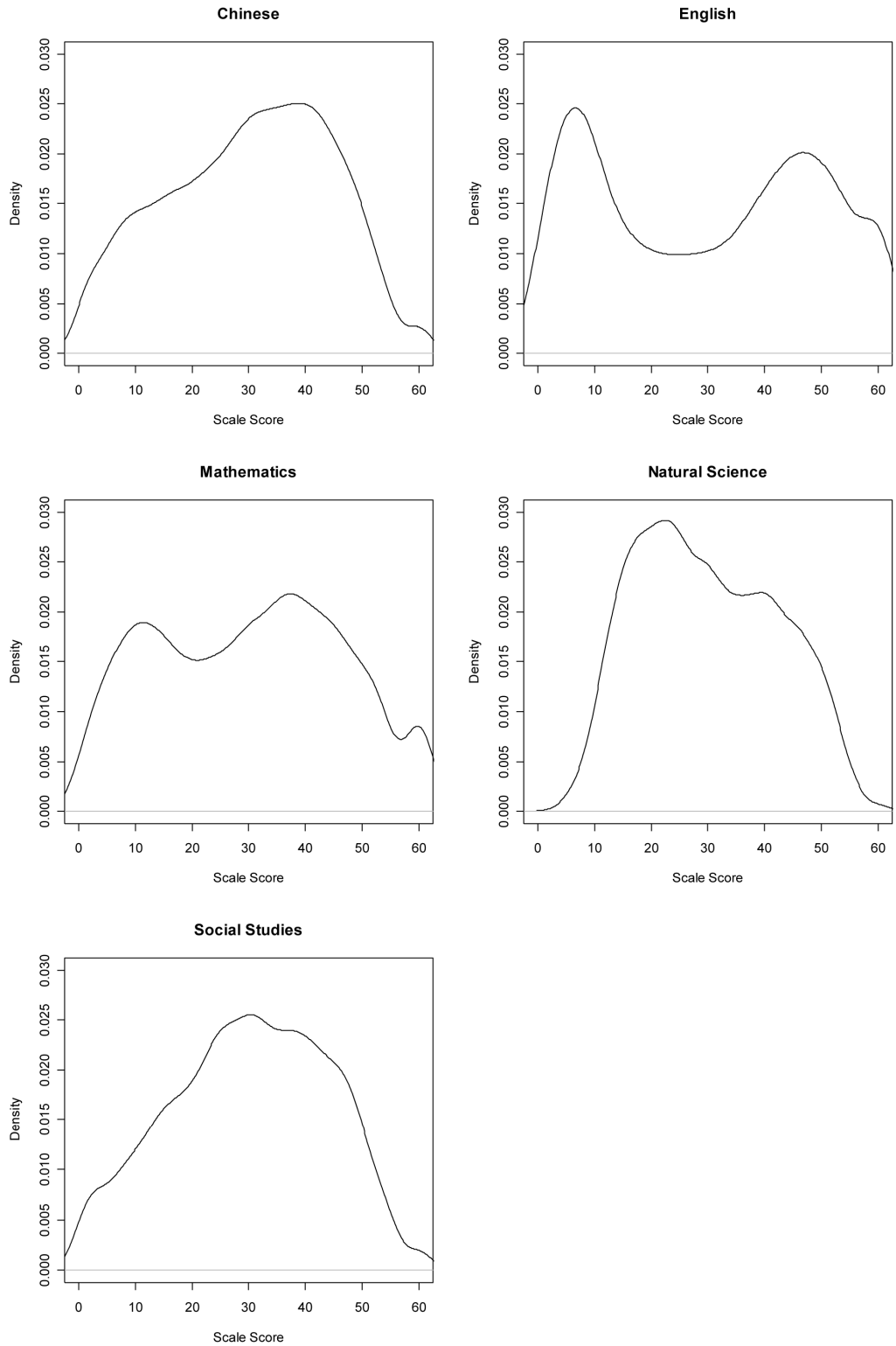

Figure 2　BCTEST scale score distributions for the various tests.

Table 3 Results of the Relative Nominal Weights of the Tests for the Various Weighting Schemes

Weighting Scheme	Chinese	English	Mathematics	Natural Science	Social Studies
A	1	1	1	1	1
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
B	0.81854	1.60133	0.77926	0.92451	0.87636
	(0.16371)	(0.32027)	(0.15585)	(0.18490)	(0.17527)
C	1.03721	0.76247	0.91093	1.21496	1.07444
	(0.20744)	(0.15249)	(0.18219)	(0.24299)	(0.21489)
D	0.96551	0.97593	0.82879	1.19722	1.03254
	(0.19310)	(0.19519)	(0.16576)	(0.23944)	(0.20651)
E	0.95226	0.72134	0.64373	1.57394	1.10873
	(0.19045)	(0.14427)	(0.12875)	(0.31479)	(0.22175)

Note. The values in parentheses are the relative proportional nominal weights.
A=the equally-weighted model; B=the reliability weighting model;
C=the *SD* weighting model; D=the error of measurement weighting model;
E=the effective score point model.

The composite scores were formed by applying to the test scale scores those relative nominal weights developed by the various weighting schemes. The final combined scores were rounded up if the value was half-way between two integers (e.g., a 150.5 would be rounded to a 151). Table 4 contains the four moments of the weighted composite scores as well as their reliability coefficients. The composite score reliability value was computed by using Equation (10) in Kolen (2006). Accompanying this table and for showing the results of the composite scores is Figure 3, in which the distributions are displayed for the various weighting schemes. In general, the outcomes of the variously formed composite scores were fairly similar, except for that of the B model (or the reliability weighting model) showing a little lower density in the middle of the scale and becoming more different from the others (see Figure 3). The similarity in the distributions indicated that the overall composite score distributions were not considerably affected by using the different weighting mechanics, in spite of the fact that fairly distinct values for the relative weights of the tests were derived via the varying weighting approaches. As for the composite score reliability, all the weighting methods led to very similar, high coefficients for the composite scores (see Table 4). None of the values were less reliable than any of the individual tests.

Table 4 The BCTEST Composite Score Summary Statistics for the Various Weighting Schemes

Weighting Scheme	Mean	*SD*	Skewness	Kurtosis	Kolen's Reliability
A	150.174	70.677	-0.014	1.862	0.98624
B	150.197	73.481	0.000	1.788	0.98630
C	150.173	69.041	-0.013	1.895	0.98593
D	150.170	69.902	-0.007	1.868	0.98651
E	150.145	67.794	0.003	1.906	0.98565

Note. The composite scores range from 5 to 300.

The correlations among the various tests and the composites are shown in Table 5. In this table, the composite scores were designated with whichever specific models the scores had been formed upon. For example, composite A means that the composite was established by employing the A model. Also, in obtaining the current correlation matrix, the results were not affected by whether the test scale scores were weighted or not. As is indicated in the table, all the individual tests were highly correlated with one another;

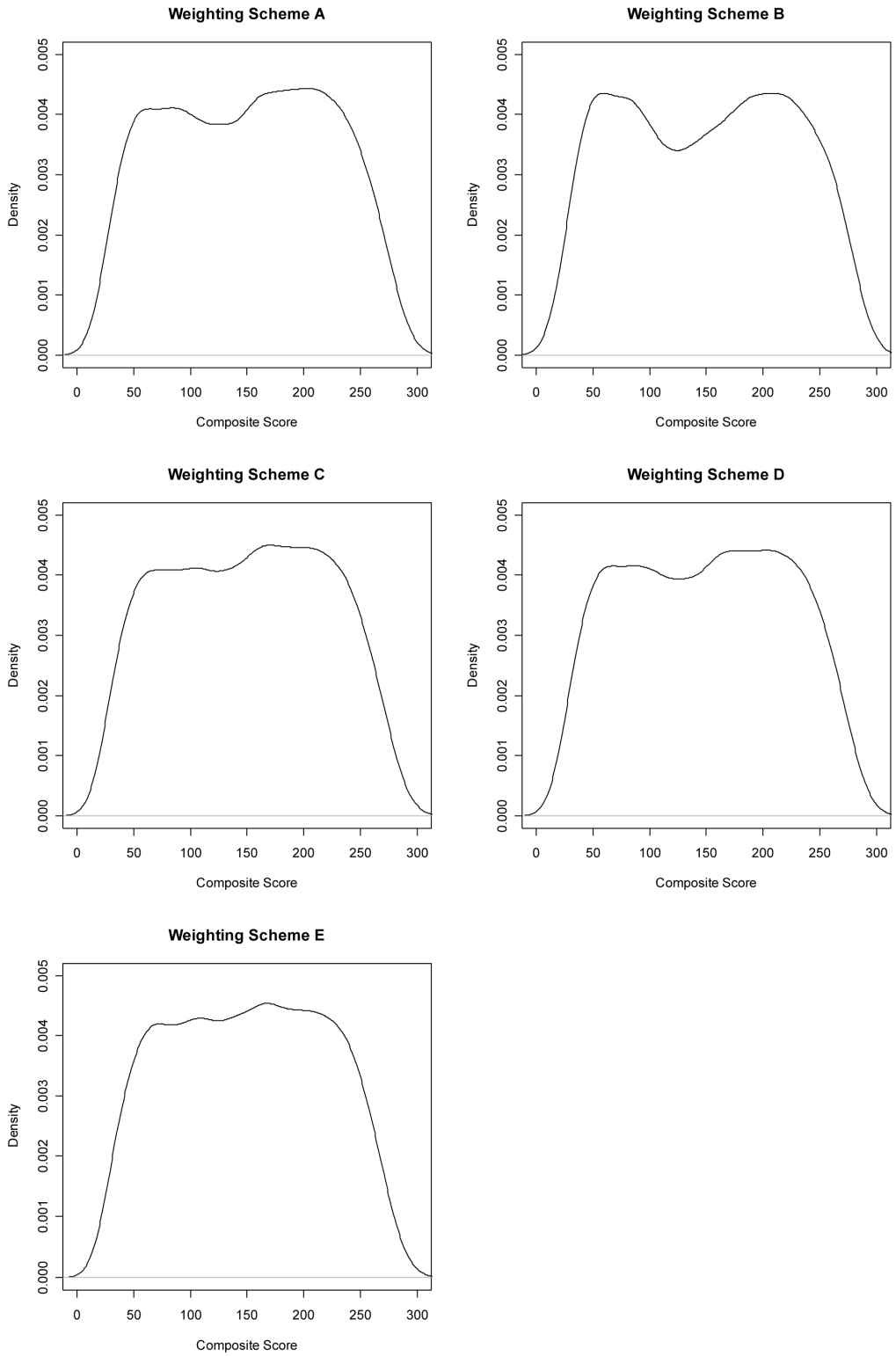

Figure 3　BCTEST composite score distributions for the various weighting schemes.

especially high were the correlation coefficients of the Chinese, Natural Science and Social Studies, of the Mathematics and Natural Science, and of the Natural Science and Social Studies tests. Regarding the correlations with the composites, the Natural Science test seemed to have slightly higher correlations with the various composites whereas both the English and Mathematics tests appeared to have slightly lower correlations, except that the English test was highly correlated with the composite B scores. In addition, the correlations between each test and the composite scores were computed based on Gulliksen's Equation (71); the outcomes are listed in Table 6. Comparing both Tables 5 and 6 for these correlation values, it is obvious that the results were almost the same, with their values agreeing by at least two decimal places.

Table 5 Results of the Correlations among the Various Tests and the Composites

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite A	Composite B	Composite C	Composite D	Composite E
No. of items	48	45	33	58	61	245	245	245	245	245
Prop. of total items	0.2	0.18	0.13	0.24	0.25					
Chinese	1.00000									
English	0.80680	1.00000								
Mathematics	0.80267	0.79924	1.00000							
Natural Science	0.84022	0.80328	0.86513	1.00000						
Social Studies	0.86738	0.79000	0.80612	0.87565	1.00000					
Composite A	0.92716	0.91893	0.92294	0.93750	0.92853	1.00000				
Composite B	0.91487	0.94710	0.90803	0.92465	0.91450	0.99686	1.00000			
Composite C	0.93076	0.90537	0.92130	0.94525	0.93552	0.99932	0.99374	1.00000		
Composite D	0.92744	0.91755	0.91719	0.94213	0.93182	0.99979	0.99658	0.99955	1.00000	
Composite E	0.92915	0.90177	0.91163	0.95322	0.93959	0.99794	0.99217	0.99939	0.99888	1.00000

Note. Results were not affected by whether the scale scores of the tests were weighted or not.

Table 6 Results of the Gulliksen's Correlations of the Test Scale Scores with the Composites of the Various Weighting Schemes

	Chinese	English	Mathematics	Natural Science	Social Studies
Composite A	0.92707	0.91884	0.92284	0.93741	0.92844
Composite B	0.91478	0.94701	0.90794	0.92456	0.91441
Composite C	0.93067	0.90528	0.92120	0.94516	0.93543
Composite D	0.92735	0.91746	0.91709	0.94204	0.93172
Composite E	0.92906	0.90168	0.91154	0.95313	0.93950

Note. The correlations were computed based on Gulliksen's equation (71).

As for the variances and covariances, Tables 7 to 11 respectively present the results for the five tests with the variously formed composites. The test scale scores employed for the computation of the variance/covariance matrices here were those weighted according to the relative weights derived via the respective weighting mechanics. Tables 7 to 11, again, also include the relative proportional nominal weights given by the various models, which can be found in the Nominal Weight row. Finally, the effective weights are shown in the last row of each table, which were derived by finding a sum of all of the elements in one row (or column) of the variance/covariance matrix, using this sum as the numerator and then dividing it by the sum of all the numerator values for all of the five test scores in the matrix. The effective weights for the five tests were standardized so that they would sum up to one. Based on the effective weight criterion (Wang & Stanley, 1970), the greater the covariance between the test and the composite, the larger the effective weight of the test; for tests with larger effective weights, their contributions to the composite scores were greater than those with smaller effective weights.

Table 7 Results of the Variances, Covariances, and Effective Weights for the Various Tests and Composite A

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite A
Nominal Weight	0.2	0.2	0.2	0.2	0.2	
Chinese	205.94	226.02	188.22	147.72	172.44	940.35
English	226.02	381.10	254.95	192.11	213.65	1267.83
Mathematics	188.22	254.95	267.00	173.19	182.48	1065.83
Natural Science	147.72	192.11	173.19	150.09	148.62	811.73
Social Studies	172.44	213.65	182.48	148.62	191.92	909.10
Composite A	940.35	1267.83	1065.83	811.73	909.10	4994.85
Effective Weight	0.18826	0.25383	0.21339	0.16251	0.18201	

Note. The computations were based on the weighted scale scores of the tests.

Table 8 Results of the Variances, Covariances, and Effective Weights for the Various Tests and Composite B

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite B
Nominal Weight	0.16371	0.32027	0.15585	0.18490	0.17527	
Chinese	137.98	296.26	120.06	111.79	123.70	789.79
English	296.26	977.24	318.14	284.41	299.82	2175.87
Mathematics	120.06	318.14	162.13	124.77	124.62	849.71
Natural Science	111.79	284.41	124.77	128.28	120.41	769.67
Social Studies	123.70	299.82	124.62	120.41	147.39	815.94
Composite B	789.79	2175.87	849.71	769.67	815.94	5400.99
Effective Weight	0.14623	0.40287	0.15733	0.1425	0.15107	

Note. The computations were based on the weighted scale scores of the tests.

Table 9 Results of the Variances, Covariances, and Effective Weights for the Various Tests and Composite C

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite C
Nominal Weight	0.20744	0.15249	0.18219	0.24299	0.21489	
Chinese	221.55	178.75	177.83	186.15	192.17	956.46
English	178.75	221.55	177.07	177.97	175.03	930.37
Mathematics	177.83	177.07	221.55	191.67	178.60	946.73
Natural Science	186.15	177.97	191.67	221.55	194.00	971.35
Social Studies	192.17	175.03	178.60	194.00	221.55	961.35
Composite C	956.46	930.37	946.73	971.35	961.35	4766.27
Effective Weight	0.20067	0.1952	0.19863	0.2038	0.2017	

Note. The computations were based on the weighted scale scores of the tests.

Table 10 Results of the Variances, Covariances, and Effective Weights for the Various Tests and Composite D

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite D
Nominal Weight	0.19310	0.19519	0.16576	0.23944	0.20651	
Chinese	191.98	212.98	150.62	170.76	171.91	898.24
English	212.98	362.97	206.21	224.47	215.29	1221.92
Mathematics	150.62	206.21	183.40	171.84	156.16	868.23
Natural Science	170.76	224.47	171.84	215.13	183.72	965.91
Social Studies	171.91	215.29	156.16	183.72	204.61	931.69
Composite D	898.24	1221.92	868.23	965.91	931.69	4886.01
Effective Weight	0.18384	0.25009	0.1777	0.19769	0.19069	

Note. The computations were based on the weighted scale scores of the tests.

Table 11 Results of the Variances, Covariances, and Effective Weights for the Various Tests and Composite E

	Chinese	English	Mathematics	Natural Science	Social Studies	Composite E
Nominal Weight	0.19045	0.14427	0.12875	0.31479	0.22175	
Chinese	186.75	155.26	115.38	221.40	182.06	860.85
English	155.26	198.30	118.38	218.12	170.87	860.92
Mathematics	115.38	118.38	110.64	175.47	130.24	650.12
Natural Science	221.40	218.12	175.47	371.82	259.35	1246.16
Social Studies	182.06	170.87	130.24	259.35	235.92	978.44
Composite E	860.85	860.92	650.12	1246.16	978.44	4596.49
Effective Weight	0.18728	0.1873	0.14144	0.27111	0.21287	

Note. The computations were based on the weighted scale scores of the tests.

For the A model, the proportional nominal weights were all equal to 0.2 but the effective weights were not the same. As can be seen in Table 7, the English test contributed the most to the composite A among the five tests. The covariance between English and the composite A was greater than those between the other tests and the composite A, because English had the largest scale score variance (as of 381.10 reported in Table 7; see also Table 2 for the *SD*s of the scale scores of the various tests) and its covariances with the other tests tended to be the highest. The Mathematics test had the next highest effective weight; the Natural Science test possessed the lowest effective weight for this composite A.

For the B model, or the reliability weighting model, the weights were found through the factor of $r/(1-r)$ and the results were that the English test was associated with considerably greater weighting than the other tests (see Table 8). By utilizing the test reliability information, the English test not only outweighed the other tests in terms of the amount of contribution, but also proceeded in a rather strong fashion. It seems that employing this weighting scheme would only make the contributions of the individual tests even more unequal, which might not appeal to testing programs as it would likely invite more controversies as well as vexing discussions among the public.

On the other hand, by weighting with the inverse of the *SD* (i.e., using the factor of $1/s$), the C weighting model assigned the English test the lowest nominal weighting of all the five tests (see Table 9). However, the effective weights resulting from the employment of the *SD* weighting seemed to be the most equal of all the five models; that is, the contributions of the individual tests to the composite C were the most alike among the five weighting methods.

For the error of measurement weighting model (or the D model) through the factor of $1\Big/\left(s_X\sqrt{1-r_{XX}}\right)$, the results of the nominal weights obtained for the various tests were the least varied, except for the equal weighting from the A approach (detailed information on the extent of variation will be provided in the next section). For the effective weights, this D model placed the English test in the most powerful position, in the same way that the A model functioned for English. But, unlike the A model which assigned the Mathematics test the second strongest weighting, the D model assigned this test the weakest weighting. Aside from the heavy weight of the English test, the results of implementing the D model seemed also appealing; the amounts of individual effective contribution were also fairly similar across the tests. More details on the D model can be found in Table 10.

As discussed earlier, with the incorporation of the idea of the effective test length, the E model yielded rather unequal nominal weights among the tests; the Natural Science test possessed the highest apparent nominal weight of all, followed by Social Studies (see Table 3). In fact, the nominal weights for the E

method were the most unequal among the five models (again, the details for the variation are contained in the following section). As for the effective weights here (see Table 11), the results did not seem intriguing either. Both the Natural Science and Social Studies tests were provided with rather dominating weighting, while the Mathematics test was given a rather minor place contributing only a relatively small amount of weight.

To study the effects on admission decisions, the proportions of examinees earning composite scores of a particular point and below were calculated at each score point for the various weighting schemes. Figure 4 presents the results of the cumulative probability differences among the weighting procedures. In this plot, the points labeled "B-A" refer to the differences in cumulative probability between the B and A models, with the values of the A being subtracted from those of the B model. Similar meanings were interpreted with the triangles labeled "C-A", the circles labeled "D-A", and the 'x' symbols labeled "E-A". These differences show how the B, C, D, or E scheme performed relative to the A scheme in accumulating examinees up to the current scale score point along the composite score continuum. That is, the differences reveal the discrepancy in the proportion of examinees obtaining composite scores of a particular point and below when the A versus the B, C, D or E method was applied, giving an indication of how the selection decisions might be affected if a certain cut-off point were to be specified for admission using the B, C, D, or E model instead.

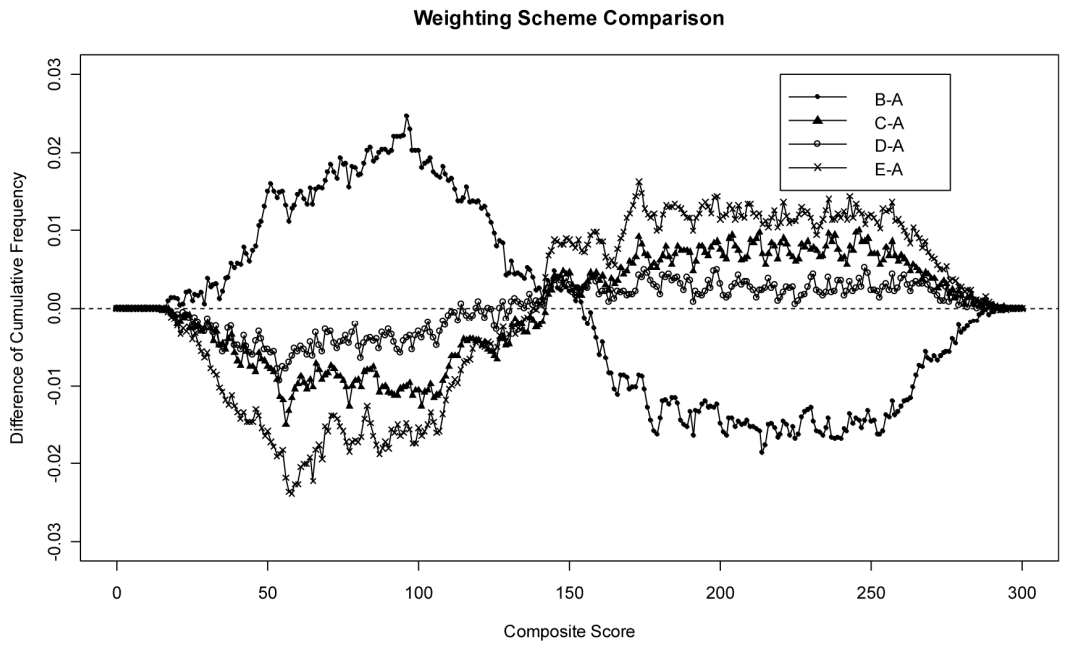

Figure 4 The cumulative probability differences among the weighting schemes.

An inspection of Figure 4 reveals that employing the B model would result in accumulating examinees more rapidly than the A model for composite scores below the middle part of the scale but more slowly on the other side of the continuum. All of the C, D and E methods functioned in the opposite way by cumulating examinees more slowly prior to the mid-composite scale score points but with a faster

accumulation from the middle towards the upper part of the scale. The extent to which these three models deviated from the A model was the greatest with the E weighting model, followed by the C and then the D schemes. Also, exercising the E scheme for the computation of the composite scores seemed to have a stronger impact at the lower part of the scale than at the upper part. Overall, both the B and E methods appeared to perform more differently from the A model than the two methods of C and D. That is, with respect to the cumulative probability results, employing the B or E model in establishing the composite scores would create more differences from the A model than would the C or D model. Replacing the A model with the B or E weighting scheme would probably result in a greater influence on the admission decisions than with the C or D scheme.

Summary and Conclusions

Summary of Results

The present study investigated and compared the results of establishing the composite scores based on the five weighting schemes of the equally-weighted model, the reliability weighting model, the *SD* weighting model, the error of measurement weighting model, and the effective score point model (represented as the A, B, C, D, and E models, respectively). The five test components of the BCTEST, Chinese, English, Mathematics, Natural Science, and Social Studies, were formed into single composite scores according to the rationales of the respective weighting models. The equally-weighted model, the unweighted model that the BCTEST currently adopts, served as a baseline for the comparison. A random sample of 5,000 examinees drawn from the data obtained from the 2005 BCTEST test administration was used in this study.

The effectiveness of the various weighting schemes was investigated via the descriptive statistics and the frequency distributions of the raw scores, the test scale scores and the resulting weighted composite scores. The composite score reliability was computed based on Equation (10) provided in Kolen (2006) as an indication of the overall measurement quality of the variously formed composites. Also, both the correlation and the variance/covariance matrices of the test component scores were computed and the effective contributions of the individual tests to the composites were compared. The impact of the different weighting methods on the admission decisions was evaluated as well.

As shown in Table 3 for the nominal weights, overall, both the Chinese and Social Studies tests seemed to be treated in a fairer manner by the various weighting schemes; the amounts of relative nominal weighting developed for these two tests were more similar across schemes, which were close to 1.0. The Natural Science test tended to have greater nominal weights than the others. Specifically, this Natural Science test had the most nominal weights of all tests with the employment of the C, D and E methods. The Mathematics test tended to have lesser relative nominal weights. As for the English test, the weights varied, depending on the weighting scheme exercised.

Regarding the effective weights obtained by applying the various weighting mechanics, the results varied appreciably both within and across models. The outcomes of the effective weights were not in accordance with those of the nominal weights in terms of the relative standings. Table 12 reveals the degree of variation of the respective sets of the nominal weights, the effective weights and the Gulliksen's

correlations for the various models by finding the *SD* values of the respective sets of the relative weights or correlations. Their rankings in terms of variation are included in parentheses for the respective sets in a column, with the value of "1" indicating the most variable and "5" meaning the least variable. As can be observed in Table 12, the relative rankings for the nominal weights and the effective weights did not agree with each other. It can also be detected from the table that both the B and E schemes were the two approaches leading to the greatest variation for the nominal and the effective weights. Not only did these two methods produce the apparent nominal weights that differed widely among the tests, but they also made the effective contributions of the tests vary considerably as well.

Table 12 *SD* Values of the Nominal Weights, Effective Weights and Gulliksen's Correlations for the Various Weighting Schemes

Weighting Scheme	Nominal Weight	Effective Weight	Gulliksen's Correlation
A	0 (5)	0.03516 (3)	0.00697 (5)
B	0.06814 (2)	0.11354 (1)	0.01532 (2)
C	0.03428 (3)	0.00327 (5)	0.01515 (3)
D	0.02665 (4)	0.02898 (4)	0.01046 (4)
E	0.07403 (1)	0.04737 (2)	0.02076 (1)

Note. The numbers in parentheses are the rankings in variation of the cell values in the respective columns, with "1" indicating the most variable and "5" meaning the least variable.

The variation in the relative weights based on the applications of the C and D models appeared more enticing than that of the B and E models (see Table 12). Concerning the criterion of approximately equal amounts of individual contribution to the composite, the C weighting model would be the most favored of all. The D model would also seem like a favorable choice if the dominating effective weighting from the English test were neglected (see Table 10 for the values). It was noted earlier that Gulliksen (1950) pointed out the frequent use of the *SD* weighting model (represented as the C model here), but he particularly recommended the use of the error of measurement weighting model, the D model in this study. The outcomes of this study seemed to be consistent with Gulliksen's suggestions, disregarding the troubled weighting of the English test with the D model.

Inspecting the rankings for the Gulliksen's correlations in Table 12, both the B and E schemes remained as the two models with the greatest variation in their resulting correlations. Nevertheless, the values of the Gulliksen's correlations of the test scale scores and the composite scores were all very high (see Table 6).

When it came to the effects on the admission decisions, adopting either the B or the E model might also be of more concern to the testing programs than would either the C or the D model. Because utilizing either the B or the E model would yield more discrepancies in cumulative probability from the A model, a bigger impact on the admission decisions could be expected to follow. Perhaps more discussions would then be provoked as the examinees and test score users compare the results of the new weighting models with the existing, conventional equal weighting system.

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 single total score is critical. For data possessing unique, distinct score

attributes of their own, some calculations might be needed in order to determine the best nominal weights possible. Especially when the score properties vary considerably among the tests, employing different sets of weights might be deemed necessary for forming the composite scores. Ignoring the differences in variation across tests could yield effective weights being very different from one another and thus lead to disparate effective contributions among the individual tests. Equal nominal weights for the tests may not be appropriate. The choice of a weighting model aims to derive composite scores that better serve as an indicator of examinees' overall performance on the test battery.

This study examined the various weighting schemes for combining five tests into composite scores using empirical real data. The tests employed in this study varied in their score distributions and measurement properties. An unweighted summation over all the individual components was compared with the weighting methods introduced in the classical test theory framework in Gulliksen (1950). The purpose was to seek optimal relative weights in forming the best composite scores possible, as well as to offer more information about the various weighting schemes that made use of the different statistical and measurement qualities of the tests.

The findings of this study indicated that for the variously formed composites, both of their reliability coefficients and the Gulliksen's correlations were all very high. Choosing any weighting scheme would not seem like a big issue as far as these two criteria were concerned. On the other hand, employing a model to actually execute the differential weighting might not be necessary since the overall measurement precision would not be practically improved. However, when it came to the effective contributions of the individual tests to the composites, the various weighting schemes performed very differently from one another. No one scheme was found to lead to equal amounts of contribution for the five tests and therefore to fully accomplish the goal of assisting in establishing the optimal composite scores. Despite the various weighting methods taking account of the varying score characteristics while developing the nominal weights, the relative effective weights of the tests could still differ widely. Some tests appeared especially more dominating than others in their contributions to the variances of the composites.

From another perspective, given the distinctive features of each individual test of the BCTEST, combining all the components into a single composite score with equal amounts of effective weighting might be very hard, if not impossible. The test score characteristics of the BCTEST considered in this study were that the numbers of the items were not the same, the score distributions were not similar, and the intercorrelations among the tests were also different. The findings of this research seemed to imply that it might be worth paying more attention to the English test while including it as one component of the BCTEST. With the English test showing rather peculiar score distributions in this study, its role in the formation of the composite scores can be very influential, due to both its particular test nature and its possible interactions with the other tests. More thoughts on the construction of the English test might be worthwhile. Striving to reach optimal composites by implementing different weighting schemes afterwards might not enhance the results to a large extent. Expectations for the weighting models to fulfill the goal might be too high to realize.

This research investigated the various weighting methods based on tests utilizing the same type of multiple-choice items. The appropriateness and effectiveness of the current schemes are unknown for test components comprised of more than one item type. For high-stakes testing programs that depend on multiple assessments to make a decision, it is conceivable that the issue of the weights that multiple

assessments exert would be much more complicated. For example, combining scores of tests using multiple-choice and constructed-response formats into one single score can be a very demanding task as their test attributes, such as the numbers of items and the reliability coefficients, might be distinctly different.

Besides the conventional approaches employed in this study, the IRT pattern scoring (Carlson, 2006; Lord, 1980; Rudner, 2001; Yen & Candell, 1991) should also be attempted. The IRT environment has offered theoretically and conceptually appealing procedures for determining relative component weights and weighting via IRT is common. However, despite the large number of studies based on the IRT framework exploring this component weighting issue, studies utilizing actual data are still relatively very few. In the IRT realm of research, data are often generated to fit the study designs perfectly and the conditions are fixed to the specifications, leaving the possibilities that the findings might be greatly altered if the assumptions underlying the analyses are not satisfactorily met or the data do not fit the model well enough. It would be a worthwhile endeavor to apply the IRT scoring schemes to the real data of the BCTEST. The results should be valuable for further justifying the performance of the IRT approaches. For tests incorporating a mixture of item types, investigations via IRT strategies should be helpful as well. Also, since the five tests of the BCTEST assess different subjects from different domains, the IRT weighting would proceed with designs under a multivariate score distribution. It would be beneficial for more studies to focus on the estimation of examinees' ability levels in the IRT context for the formation of composite scores.

While the emphasis is on the selection of a proper weighting model, a "not-equally-weighted" model can be very hard to explain to the examinees and the users of the composite scores. It is very important that 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 help understand the formation of the composite scores. Additionally, evaluation of the consequences of implementing different weighting schemes must be carefully conducted.

Although this study has confined its explorations to the particular BCTEST assessment, the ideas and issues also carry over to most large-scale testing settings where similar situations are encountered. However, the results might be still rather specific to the given year 2005. Further studies could also continue with test data obtained from different testing years. Adopting an appropriate weighting scheme is a testing program's choice, but the decision can be very hard to make. As Gulliksen (1950) indicated, the weighting problem arises whenever a single overall score is to be established from the separate test scores. Solving such a problem can be very challenging as it requires many detailed considerations of the test attributes as well as the score uses. Results from this research should help psychometric researchers and test practitioners gain more insights into the impact that individual tests have on the formation of composite scores and also, advance their understanding of the weighting issues while combining test components into a test battery that best meets the testing program's needs.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Carlson, J. E. (2006, April). *Issues in differential weighting of items in IRT scoring*. Paper presented at the

annual meeting of the National Council on Measurement in Education, San Francisco.

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

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in Counseling and Development, 37*(3), 184-190.

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

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221-240.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). CT: American Council on Education and Praeger Publishers.

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

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.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

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

Ma, X., Kim, S., & Walker, M. E. (2006, April). *Optimal weighting of section scores and forming a composite score*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Pei, L. K., & Maller, S. J. (2006, April). *Monte carlo simulation study of differential weights on composite reliability and validity*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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: American Council on Education, and Macmillan.

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

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 23-72). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Wang, T. (1998). Weights that maximize reliability under a congeneric model. *Applied Psychological Measurement, 22*(2), 179-187.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4*(3), 209-228.

收 稿 日 期：2008 年 04 月 07 日
一稿修訂日期：2008 年 07 月 10 日
接受刊登日期：2008 年 08 月 21 日

國 立 臺 灣 師 範 大 學 教 育 心 理 與 輔 導 學 系
教育心理學報，民 98，40 卷，3 期，489－510 頁

量尺總分加權機制之探討

章 舜 雯
國立台灣師範大學
教育心理與輔導學系

本研究探討與比較使用「等比重加權」、「信度加權」、「標準差加權」、「測量誤差加權」，以及「有效分數加權」這五種不同的加權機制模式建立量尺總分的效果。 研究的目的在探索適合各測驗學科的最佳名義權重，以致能形成最合適的量尺總分，同時也提供將測驗學科的測量特性與分數分配的特徵列入計算各科權重的過程後，更多關於這些不同模式的有用訊息。本研究使用國民中學學生基本學力測驗的五科測驗進行，樣本採自民國 94 年考生分數 5,000 筆的隨機資料。研究評鑑各加權機制模式效果的準則包含，各學科分數與加權後總分的統計與測量方面的特性、各學科對總分的有效貢獻量，以及之於高中入學選擇決定的影響。研究結果指出，經由不同加權機制模式所形成的量尺總分之信度係數都很高。然而，當一一檢視各學科對總分的有效貢獻量時，不同模式的效果卻有極大的差異。雖然使用「標準差加權」與「測量誤差加權」模式仍無法使每個測驗學科在對總分的有效貢獻量上達到大致相當的最佳目標，但整體而言，「標準差加權」與「測量誤差加權」這兩種加權機制模式的表現仍比「信度加權」或「有效分數加權」的模式來得好。本研究的結果與建議，探討如何將各科測驗分數作最合適的組合以及有關量尺總分的相關議題，對於測驗的研究或實務方面都可提供相當的助益。

關鍵詞：有效貢獻量、有效權重、名義權重、總分、總分加權