Article

Applied Psychological Measurement
2015, Vol. 39(7) 539–550

DOI: 10.1177/0146621615584703
apm.sagepub.com

Weighting Components of a Composite Score Using Naïve Expert Judgments About Their Relative Importance

Peter Baldwin[1]

Abstract

A common problem that arises in testing—as well as other contexts such as candidate selection—is how to combine various scores into a weighted composite that reflects expert judgments about each component's relative importance. For experts to provide nominal weights explicitly, they must fully account for the variances of the components, the covariances among components, and the reliability of each component. This task can be challenging, and in many cases, experts may have greater success making simple judgments about component importance without regard for the variances, covariances, and reliabilities. In this article, it is shown how to estimate the requisite nominal weights when only these kinds of naïve judgments are available, and the analytical solution is demonstrated with a small simulation study. Results from the simulation suggest that the proposed estimators could yield more valid composite scores in practice.

Keywords

composite scores, weighted composites, test batteries

Introduction

The need to combine different measures to produce a single composite arises often. In testing, for example, it may be necessary to combine scores from complex performance assessments with scores from conventional multiple-choice tests into a single summary score. The U.S. Medical Licensure Examination and the Uniform Certified Public Accountant Examination are two well-known licensure exams that follow this practice (American Institute of Certified Public Accountants, 2011; Clauser, Harik, & Margolis, 2006). In other cases, it may be desirable to combine measures—which may be remarkably diverse with respect to both construct and reliability—into a single composite for the purpose of making selection or other decisions. For example, in the United States and Canada, medical school admissions decisions are informed by combining undergraduate grade point average, test scores, application essay quality, admission interview, and various other heterogeneous measures into a single composite intended to summarize each applicant (Eva & Reiter, 2004; Kreiter, Yin, Solow, & Brennan, 2004).

[1]National Board of Medical Examiners, Philadelphia, PA, USA

Corresponding Author:
Peter Baldwin, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104-3102, USA.
Email: pbaldwin@nbme.org

Usually composite scores can be expressed in the form of a weighted sum and because these scores are often used to make important decisions, methods for determining the associated weights have been of long-standing interest to researchers (e.g., Corey, 1930; Edgerton & Kolbe, 1936; Govindarajulu, 1988; Horst, 1936; Hotelling, 1935; Kaiser, 1967; Kaiser & Caffrey, 1965; Peel, 1947, 1948; Perloff & Persons, 1988; Ryans, 1954; Thomson, 1940; Wainer, 1976; Wang and Stanley, 1970; Wilks, 1938; Woodbury & Lord, 1956). When external validity criteria are available, weights are typically selected that yield composite scores with some specified relationship (e.g., least squares) with the criteria. In contrast, when external criteria are unavailable, weights are chosen that either optimize certain statistical properties of the composite scores (e.g., reliability) or reflect expert judgments of one kind or another (Wang & Stanley, 1970).[1] In this article, the concern is with the use of weights based on expert judgments. These kinds of weights are typically called *a priori* or *subjective weights*.

A priori weights allow each component score to contribute differently to a composite in a manner that corresponds with experts' or policy makers' beliefs about its relative importance (Burt, 1950).[2] Ideally, when experts are consulted about the relative importance of each component, they would make recommendations about the needed weights directly, which could then be used to produce composite scores; however, for such weights to capture experts' views about relative importance, judges must fully account for the variances of the components, the covariances among components, and the reliability of each component. Making such sophisticated judgments is challenging, and in many cases, experts may have greater success making simple judgments about component importance *without* regard for the variances, covariances, and reliabilities of each component. Conceptually, such naïve judgments correspond to the contribution of each component to the variance of the resultant *true score* composite. In this article, it is shown how weights can be estimated that minimize the mean square differences between the observed score composite and the (unknown) true score composite specified by these kinds of judgments. That is, a solution is given to the important practical problem: *When experts provide naïve judgments about component importance, what weights should be used in practice?*

Background: Observed Score Nominal Weights, True Score Nominal Weights, and True Score Effective Weights

Let x be a weighted composite of S component scores such that

$$x = \sum_{s=1}^{S} w_s x_s, \tag{1}$$

where w_s is the *nominal* weight and x_s is the observed score for component s. An analogous relationship may also be asserted for true scores: Let τ be a composite of S true component scores such that

$$\tau = \sum_{s=1}^{S} \omega_s \tau_s, \tag{2}$$

where ω_s is the *true score nominal weight* for component s and τ_s is the true score on component s. It follows that the variance of the composite true score, $\sigma^2(\tau)$, is given by

$$\sigma^2(\tau) = \sum_{s=1}^{S} \omega_s^2 \sigma^2(\tau_s) + \sum_{s=1}^{S} \sum_{\substack{j=1 \\ j \neq s}}^{S} \omega_s \omega_j \sigma(\tau_s, \tau_j), \tag{3}$$

where $\sigma(\tau_s, \tau_j)$ is the covariance between component s and component j true scores. Disaggregating Equation 3, it can be seen that the contribution of each component s to the true composite score variance is simply

$$\kappa_s = \omega_s^2 \sigma^2(\tau_s) + \sum_{\substack{j=1 \\ j \neq s}}^{S} \omega_s \omega_j \sigma(\tau_s, \tau_j), \tag{4}$$

where the contribution of component s, κ_s, is referred to here as the *true score effective weight*.[3] Note that true score effective weights correspond to the naïve judgments we might expect experts to make most easily about the relative importance of each component. Of course, as a practical matter, the *observed score nominal weights*, $\mathbf{w}$, are the weights needed to proceed with the business of producing scores—and, in contrast to true score effective weights, these nominal weights must indeed account for the variances, covariances, and reliabilities of the components. In this article, a method for estimating the nominal weights, $\mathbf{w}$, given the true score effective weights, $\boldsymbol{\kappa}$, is introduced.

Estimating w Given κ

Regression provides a convenient framework for estimating the nominal weights that minimize the mean squared error between the observed composite scores, $\mathbf{x}$, and true composite scores, $\boldsymbol{\tau}$. Specifically, after converting the scores for each component into deviation scores (which can be done without loss of generality), $\boldsymbol{\tau}$ can be regressed onto the observed weighted sum:

$$\boldsymbol{\tau} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\mathbf{X}$ is the person-by-component score matrix. It follows that the method of ordinary least squares can be used to estimate the nominal weights, $\mathbf{w}$:

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tau, \tag{6}$$

where $\mathbf{X}'$ is the matrix $\mathbf{X}$ transpose. Given deviation scores, $(\mathbf{X}'\mathbf{X})^{-1}$ will be equal to the inverse of the product of S and the variance–covariance matrix for the observed component scores, $(S\boldsymbol{\Sigma})^{-1}$. Likewise, $\mathbf{X}'\tau$ will be equal to the product of S and the vector of covariances between the true composite scores and each component's observed scores, $S\boldsymbol{\sigma}$. In this way, the matrix formulation for the computation of the regression coefficients may be rewritten more compactly as

$$\hat{\mathbf{w}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}, \tag{7}$$

where again $\boldsymbol{\Sigma}$ is the variance–covariance matrix for the observed component (deviation) scores and $\boldsymbol{\sigma}$ is the vector of covariances between the true composite scores and each component's observed scores. Obviously, $\boldsymbol{\Sigma}$ may be estimated directly from the observed data; however, estimating $\boldsymbol{\sigma}$ requires somewhat more effort.

The covariance between a single random variable and a weighted sum of random variables is simply the weighted sum of component covariances. Therefore, a given covariance element, $\sigma(\tau, x_s)$, from the vector $\boldsymbol{\sigma}$ may be expressed as

$$\sigma(\tau, x_s) = \sum_{j=1}^{S} \omega_j \sigma(\tau_j, x_s), \tag{8}$$

where ω_j is the true score nominal weight for component j, and $\sigma(\tau_j, x_s)$ is the covariance between the observed scores for component s and true scores for component j. Setting aside the true score nominal weights for a moment, note that because errors are uncorrelated with true scores, it follows that $\sigma(\tau_j, x_s) = \sigma(x_j, x_s)$ for all components $j \neq s$. Elsewhere, when $j = s$, the covariance is given by

$$\sigma(\tau_s, x_s) = \rho_{x_s \tau_s} \sigma(x_s) \sigma(\tau_s), \tag{9}$$

which may be written more compactly as

$$\sigma(\tau_s, x_s) = \rho_{x_s x'_s} \sigma^2(x_s), \tag{10}$$

where $\rho_{x_s x'_s}$ is the reliability of component s and $\sigma^2(x_s)$ is the variance of the observed scores on component s, both of which can be estimated from observed data. Thus, the only remaining quantities needed to solve Equation 8 are the true score nominal weights, $\boldsymbol{\omega}$, which are estimable given the true score effective weights, $\boldsymbol{\kappa}$ (i.e., naïve expert judgments), and the true score variances and covariances, which can be estimated from the observed scores.

In the case of only two components, an analytical solution to the problem of estimating $\boldsymbol{\omega}$ exists, which is given in the appendix[4]; for the general case, in the absence of any known analytical solution, iterative solutions have been proposed by Dunnette and Hoggatt (1957) and Elster (1972). Readers are referred to the attached appendix or to these other articles for specific details on each solution; here, the method by which $\boldsymbol{\omega}$ is estimated is not of particular interest—it is simply assumed that estimates are obtained one way or another.

After estimating the true score nominal weights, $\boldsymbol{\omega}$, Equation 8 and Equation 10 can be combined to estimate each element in the vector $\boldsymbol{\sigma}$:

$$\sigma(\tau, x_s) = \omega_s \rho_{x_s x'_s} \sigma^2(x_s) + \sum_{\substack{j=1 \\ j \neq s}}^{S} \omega_j \sigma(x_j, x_s). \tag{11}$$

Recalling Equation 7, it can be seen that Equation 11 along with the variance–covariance matrix, $\boldsymbol{\Sigma}$, is all that is needed to estimate the regression weights—that is, the nominal weights, $\mathbf{w}$, that minimize the mean squared error between the observed composite scores, $\mathbf{x}$, and the true composite scores, $\boldsymbol{\tau}$, given the true score effective weights, $\boldsymbol{\kappa}$.

One final quantity of interest will be the reliability of the composite x, which can be estimated by calculating the ratio of the true score composite variance to the observed score composite variance. This is given by

$$\rho_{xx'} = \rho^2_{x\tau} = \frac{\mathbf{w}'\boldsymbol{\Sigma}^*\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}, \tag{12}$$

where $\boldsymbol{\Sigma}^*$ is equal to the variance–covariance matrix $\boldsymbol{\Sigma}$ except that the diagonal is equal to 1.

Simulation Study

Some readers may find it valuable if an application of the proposed method is provided for illustrative purposes. To this end, a small simulation study was undertaken, which—aside from any pedagogical value—also creates an opportunity to verify the analytical result presented above and to show how the proposed weights compare with other approaches.

Step-by-Step Application of the Proposed Method

Suppose policy makers wished to combine scores from two measures, A and B, into a weighted composite. Experts are consulted, and their naïve judgments about the relative importance of each component indicate that Component B is 50% more important than Component A. Given true composite scores with unit variance, these judgments correspond to true score effective weights of $\kappa_A = .4$ and $\kappa_B = .6$. Furthermore, suppose the true score correlation between these two components was $\rho_{\tau_A\tau_B} = .50$, that the component reliabilities were $\rho_{x_A x'_A} = .80$ and $\rho_{x_B x'_B} = .45$, respectively, and, for convenience (and without loss of generality), true score means and variances were scaled such that $\mu_{\tau_A} = \mu_{\tau_B} = 0.0$ and $\sigma^2(\tau_A) = \sigma^2(\tau_B) = 1.0$. Finally, let the components' true and observed scores follow a multivariate normal distribution, which—given the various characteristics just described—yields the following mean vector and variance–covariance matrix:

$$\mathbf{G} \sim N\left(\begin{pmatrix}\mu_{\tau_A}\\ \mu_{\tau_B}\\ \mu_{\tau_A}\\ \mu_{\tau_B}\end{pmatrix}, \begin{pmatrix}\sigma^2(\tau_A) & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_A) & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B)\\ \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_B) & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_B)\\ \sigma^2(\tau_A) & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_A)\rho^{-1}_{x_A x'_A} & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B)\\ \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_B) & \rho_{\tau_A\tau_B}\sigma(\tau_A)\sigma(\tau_B) & \sigma^2(\tau_B)\rho^{-1}_{x_B x'_B}\end{pmatrix}\right)$$

$$\mathbf{G} \sim N\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix}1.0 & 0.5 & 1.0 & 0.5\\ 0.5 & 1.0 & 0.5 & 1.0\\ 1.0 & 0.5 & 1.25 & 0.5\\ 0.5 & 1.0 & 0.5 & 2.22\end{pmatrix}\right). \tag{13}$$

Scores for 10,000 simulees were sampled from the multivariate normal distribution **G**. These scores were then rescaled such that the (estimated) true score means and variances were 0.0 and 1.0, respectively, yielding deviation scores (e.g., for Component A, this was accomplished by applying a linear transformation with a $slope_{x_A} = \left[\sigma(x_A)\sqrt{\rho_{x_A x'_A}}\right]^{-1}$ and an interecpt$_{x_A} = 0.0 - slope_{x_A}\mu_{x_A}$). The resultant mean vector and variance–covariance matrix for the *rescaled* sampled scores were

$$\begin{pmatrix}\mu_{\tau_A}\\ \mu_{\tau_B}\\ \mu_{x_A}\\ \mu_{x_B}\end{pmatrix} = \begin{pmatrix}.01\\ .00\\ .00\\ .00\end{pmatrix} \tag{14}$$

and

$$\begin{pmatrix}\sigma^2_{\tau_A}(\tau_A) & \sigma(\tau_A,\tau_B) & \sigma(\tau_A,x_A) & \sigma(\tau_A,x_B)\\ \sigma(\tau_A,\tau_B) & \sigma^2(\tau_B) & \sigma(\tau_B,x_A) & \sigma(\tau_B,x_B)\\ \sigma(\tau_A,x_A) & \sigma(\tau_B,x_A) & \sigma^2(x_A) & \sigma(x_A,x_B)\\ \sigma(\tau_A,x_B) & \sigma(\tau_B,x_B) & \sigma(x_A,x_B) & \sigma^2(x_B)\end{pmatrix} = \begin{pmatrix}1.00 & 0.51 & 1.00 & 0.51\\ 0.51 & 1.00 & 0.50 & 1.00\\ 1.00 & 0.50 & 1.25 & 0.51\\ 0.51 & 1.00 & 0.51 & 2.22\end{pmatrix}, \tag{15}$$

which are very close to the population values shown in Equation 13.

Note that in practice only quantities for observed scores will be known; these values are reported in the four cells in the lower rightmost corner of the matrix shown in Equation 15 and correspond to the observed score variance–covariance matrix used in Equation 7 above:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.25 & 0.51 \\ 0.51 & 2.22 \end{pmatrix}. \tag{16}$$

Using this variance–covariance matrix along with the component reliabilities, the variance–covariance matrix for the true scores can be estimated:

$$\hat{\boldsymbol{\Sigma}}_{\tau} = \begin{pmatrix} 1.00 & 0.51 \\ 0.51 & 1.00 \end{pmatrix}. \tag{17}$$

Note that while similar, these values will *not* correspond exactly to those shown in the upper left of the matrix from Equation 15, which describe the simulated true scores; rather, Equation 17 shows the *estimated* true values based on the observed score variance–covariance matrix and each component's reliability—that is, these are the estimated values that could be obtained in practice.

The estimated true score variance–covariance matrix shown in Equation 17 along with the true score effective weights $\kappa_A = .4$ and $\kappa_B = .6$ can then be used to estimate the true score nominal weights, ω_A and ω_B, using the method shown in the appendix:

$$\hat{\omega}_A = \sqrt{\frac{2\kappa_A^2\hat{\sigma}^2(\tau_B)}{2\kappa_A\hat{\sigma}^2(\tau_A)\hat{\sigma}^2(\tau_B) + \hat{\sigma}(\tau_A,\tau_B)\sqrt{\hat{\sigma}^2(\tau_A,\tau_B)(\kappa_A-\kappa_B)^2 + 4\kappa_A\kappa_B\hat{\sigma}^2(\tau_A)\hat{\sigma}^2(\tau_B)} - \hat{\sigma}^2(\tau_A,\tau_B)(\kappa_A-\kappa_B)}}$$

$$\hat{\omega}_A = .49 \tag{18}$$

and

$$\hat{\omega}_B = \omega_A \frac{\sqrt{\hat{\sigma}^2(\tau_A,\tau_B)(\kappa_A-\kappa_B)^2 + 4\kappa_A\kappa_B\hat{\sigma}^2(\tau_A)\hat{\sigma}^2(\tau_B)} - \hat{\sigma}(\tau_A,\tau_B)(\kappa_A-\kappa_B)}{2\kappa_A\hat{\sigma}^2(\tau_B)} = .66. \tag{19}$$

Note that the true score variances and covariances in Equation 18 and Equation 19 are the estimated values from Equation 17 that could be obtained in practice.

Recalling Equation 7 above, only two things are needed to estimate the nominal weights for the observed scores: the variance–covariance matrix for the observed component scores, $\boldsymbol{\Sigma}$, which for these particular data is shown in Equation 16, and the vector of covariances between the true composite scores and the observed component scores, $\boldsymbol{\sigma}$, which—now that the true score nominal weights, ω_A and ω_B, have been estimated—can be estimated using Equation 11 above:

$$\hat{\boldsymbol{\sigma}} = \begin{bmatrix} \hat{\sigma}(\tau, x_A) \\ \hat{\sigma}(\tau, x_B) \end{bmatrix} = \begin{bmatrix} \hat{\omega}_A\rho_{x_A x'_A}\sigma^2(x_A) + \omega_B\sigma(x_A,x_B) \\ \hat{\omega}_B\rho_{x_B x'_B}\sigma^2(x_B) + \omega_A\sigma(x_A,x_B) \end{bmatrix} = \begin{bmatrix} 0.49\times0.80\times1.25 + 0.66\times0.51 \\ 0.66\times0.45\times2.22 + 0.49\times0.51 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 0.91 \end{bmatrix}. \tag{20}$$

Table 1. Reliability and Validity for Three Sets of Nominal Weights.

	Reliability ($\rho_{xx'}$)	Validity ($\rho_{x\tau}$)	w_A	w_B
$\mathbf{w} = \boldsymbol{\kappa}$	.62	.78	.40	.60
$\mathbf{w} = \mathbf{w}_{maximum\ reliabilty}$	.81	.78	.89	.11
$\mathbf{w} = \mathbf{w}_{proposed}$	.76	.84	.66	.34

Note. To simplify comparisons, nominal weights were rescaled such that $w_A + w_B = 1$.

The product of this vector of covariances and the inverse of the observed score variance–covariance matrix from Equation 16 can then be used to estimate the nominal weights using Equation 7:

$$\hat{\mathbf{w}} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\sigma}} = \begin{bmatrix} 1.25 & 0.51 \\ 0.51 & 2.22 \end{bmatrix}^{-1} \begin{bmatrix} 0.82 \\ 0.91 \end{bmatrix} = \begin{bmatrix} 0.88 & -0.20 \\ -0.20 & 0.50 \end{bmatrix} \begin{bmatrix} 0.82 \\ 0.91 \end{bmatrix} = \begin{bmatrix} 0.54 \\ 0.28 \end{bmatrix}. \tag{21}$$

That is, the nominal weight for Component A is 0.54, and the nominal weight for Component B is 0.28.

Comparing the Proposed Method With Two Alternative Methods

Now that the steps needed for its implementation have been shown, results from the proposed method may be compared with those obtained using two alternative approaches. Thus, nominal weights, $\mathbf{w}$, were produced for three methods altogether: (a) nominal weights equal to the true score effective weights, $\mathbf{w} = \boldsymbol{\kappa}$; (b) nominal weights that maximize reliability, $\mathbf{w} = \mathbf{w}_{maximum\ reliabilty}$; and (c) nominal weights equal to the regression coefficients obtained using the method proposed above, $\mathbf{w} = \mathbf{w}_{proposed}$. Note that Method (a), wherein the (true score effective) weights provided by experts are used as nominal weights directly, is typically the approach taken when experts are consulted (Wang & Stanley, 1970).

Table 1 reports the reliability, validity, and nominal weights for each method. Because data were simulated, the true component scores, τ_A and τ_B, and the true score nominal weights, ω_A and ω_B, that correspond with $\kappa_A = .4$ and $\kappa_B = .6$ were known without error. It follows that the true score composite based on these *specific* true score nominal weights may be used as a criterion to evaluate the three methods' respective observed score composites. To this end, *validity* is defined as the correlation between the observed score composite and the criterion true score composite and is denoted by $\rho_{x\tau}$. (Care should be exercised not to confuse the reported correlations, $\rho_{x\tau}$, with reliability indices.)

Table 1 shows that reliability and validity are lowest when true score effective weights are used as nominal weights ($\mathbf{w} = \boldsymbol{\kappa}$). When nominal weights are selected that maximize reliability ($\mathbf{w} = \mathbf{w}_{maximum\ reliabilty}$), reliability increases substantially to .81; however, validity remains unchanged at .78. Reliability also increases substantially when the proposed method ($\mathbf{w} = \mathbf{w}_{proposed}$) is used—although (not surprisingly) the proposed method yields less reliable scores than the *maximum* reliability condition (.76 vs. .81). Most importantly, the proposed method yields the highest validity (.84).

Verifying the Analytical Results

The proposed weights, $\hat{w}_A$ and $\hat{w}_B$, are intended to minimize the mean squared difference between the observed score composite and the true score composite. The extent to which this is

Table 2. On the Quality of the Proposed Nominal Weights, $\hat{w}_A$ and $\hat{w}_B$.

Description	Definition	Observed value based on 1,000 replications
Mean error	$\bar{e}_A = E(\hat{w}_A - w_A)$	0.0000
	$\bar{e}_B = E(\hat{w}_B - w_B)$	0.0000
Root mean square error	$\sqrt{E(e_A^2)} = \sqrt{E\left((\hat{w}_A - w_A)^2\right)}$	0.0030
	$\sqrt{E(e_B^2)} = \sqrt{E\left((\hat{w}_B - w_B)^2\right)}$	0.0030

accomplished is difficult to judge in practice because the true scores are never known; however, because data were simulated, here the estimates, $\hat{w}_A$ and $\hat{w}_B$, may be compared directly with their respective parameter values, w_A and w_B. To this end, let e_A and e_B denote the estimation error in $\hat{w}_A$ and $\hat{w}_B$, where $e_A = \hat{w}_A - w_A$ and $e_B = \hat{w}_B - w_B$.[5]

The simulation described above was replicated 1,000 times, and the properties of e_A and e_B are summarized in Table 2. For both Components A and B, it can be seen that the mean error and the root mean square error across replications are very close to zero. In this way, the analytical result conforms with expectation.

Discussion

Given unsophisticated expert judgments about component importance, it was shown how to estimate the nominal weights needed in practice. Using simulated data, it was found that the correlation between the observed score composites and the true score composites was higher when nominal weights were estimated using the proposed estimators rather than those obtained from two other methods: (a) using the expert judgments as nominal weights directly and (b) selecting nominal weights that maximized the reliability of the observed score composite. More generally, it was also found that the proposed method performed as expected, yielding estimates of $\hat{w}_A$ and $\hat{w}_B$ that were extremely close to their respective parameter values, w_A and w_B. These findings should be of considerable interest to practitioners who use expert judgments to determine or inform their choice of nominal weights.

Although these results are encouraging, practitioners should remain cognizant of two sources of variability that may arise when estimating nominal weights given expert judgments (these limitations are not unique to the proposed estimators, but they nevertheless deserve comment). First, experts will have different views about the relative importance of various components and thus, when judgments are aggregated, the result will be affected by the size and composition of the sample of experts consulted. It is worth noting that this variability would arise even in the absence of any errors in the judgments—or, in other words, we expect judges to have different opinions.

Second, note that the proposed estimators are indeed estimators—the quality of the estimates will vary depending on the number and quality of observations on which they are based. Thus, it is reasonable to expect better results when the reliabilities and the variance–covariance matrix are well-estimated. For the simulation described above, for example, standard errors in the resultant nominal weights can be estimated by bootstrapping. When this is done, 10,000 simulees yield standard errors of .0028; but for n = 1,000, standard errors increase to .0089; and for n = 100, they increase further still to .0336.[6] Of course, any method that relies on estimated values (e.g., maximizing reliability) will encounter the same issues.

In closing, it is worth commenting on the judgment task the proposed method imposes on experts. As explained above, the proposed method arose out of concern that the task expected of experts was too complex—that accounting for the variances, covariances, and reliabilities of the components was onerous and impractical. Instead, it was suggested that identifying the relative importance of the components *without* regard for these characteristics could be a more manageable task. Nevertheless, while it is easy to imagine judges having greater success accomplishing this simplified task (particularly if they are given adequate instruction), it may also present its own set of challenges. Consider two.

First, when judges are asked about the relative importance of *what* is measured, they may find it difficult to set aside knowledge about *how* it is measured. For example, suppose it is necessary to create a composite from a reading score and mathematics score and that a given expert believes that reading is twice as important as mathematics. If the two exams are equal length, the expert may have no difficulty accomplishing the judgment task correctly; however, if the reading exam is merely an hour long but the mathematics exam is a full day long, the expert's beliefs may be influenced by the relative importance implied by the different exam lengths.

Second, experts may find it hard to judge the importance of a given component if the construct being measured differs significantly from the component's intended construct. For example, suppose an exam is meant to measure some area of mathematics but that the reading load is such that examinee performance is affected by their reading proficiency in non-trivial ways. In this case, judges would be called upon to determine the relative importance of the aggregate construct—that is, the combination of the mathematics dimension and the (construct-irrelevant) reading dimension. Accounting for such (non-random) sources of construct-irrelevant variance could make the judgment task more difficult. That said, the validity of the components will be an important factor to consider, no matter how weights are determined. Given this, it follows that the proposed method is in some sense advantageous in that it at least creates the opportunity to account for construct-irrelevant factors of this kind.

Appendix

Determining Nominal Weights Given Effective Weights: An Analytical Solution for Two Components

To begin, it is assumed that the covariances among component scores are positive (or zero) and moreover that the nominal and effective weights for each component are positive. Typically, component scores will conform to these assumptions without requiring any further adjustment; however, when this is not so, data may be rescaled in this manner without loss of generality.

For the case of only two components, Equation 4 from above yields but two equations:

$$\kappa_1 = \omega_1^2\sigma^2(\tau_1) + \omega_1\omega_2\sigma(\tau_1, \tau_2). \tag{A1}$$

$$\kappa_2 = \omega_2^2\sigma^2(\tau_2) + \omega_1\omega_2\sigma(\tau_1, \tau_2). \tag{A2}$$

Together, Equation A1 and Equation A2 make a system of two equations with two unknowns, which yields a tractable quartic polynomial; however, first solving for the ratio $\omega' = \omega_2/\omega_1$ makes the roots of interest more readily apparent. To this end, both sides of Equations A1 and A2 can be multiplied by ω_1^{-2} and ω' can be substituted for ω_2/ω_1:

$$\frac{\kappa_1}{\omega_1^2} = \sigma^2(\tau_1) + \omega'\sigma(\tau_1, \tau_2). \tag{A3}$$

$$\frac{\kappa_2}{\omega_1^2} = \omega'^2 \sigma^2(\tau_2) + \omega' \sigma(\tau_1, \tau_2). \tag{A4}$$

These equations can then be written as the ratio

$$\frac{\kappa_1 \omega_1^2}{\kappa_2 \omega_1^2} = \frac{\sigma^2(\tau_1) + \omega' \sigma(\tau_1, \tau_2)}{\omega'^2 \sigma^2(\tau_2) + \omega' \sigma(\tau_1, \tau_2)}, \tag{A5}$$

or more compactly as

$$\frac{\kappa_1}{\kappa_2} = \frac{\sigma^2(\tau_1) + \omega' \sigma(\tau_1, \tau_2)}{\omega'^2 \sigma^2(\tau_2) + \omega' \sigma(\tau_1, \tau_2)}. \tag{A6}$$

Now, it is a simple matter of rearranging Equation A6 into the quadratic form:

$$\omega'^2 \left(\kappa_1 \sigma^2(\tau_2)\right) + \omega' (\kappa_1 \sigma(\tau_1, \tau_2) - \kappa_2 \sigma(\tau_1, \tau_2)) - \kappa_2 \sigma^2(\tau_1) = 0, \tag{A7}$$

and using the quadratic formula to solve for ω'

$$\omega' = \frac{\pm\sqrt{\sigma^2(\tau_1, \tau_2)(\kappa_1 - \kappa_2)^2 + 4\kappa_1 \kappa_2 \sigma^2(\tau_1)\sigma^2(\tau_2)} - \sigma(\tau_1, \tau_2)(\kappa_1 - \kappa_2)}{2\kappa_1 \sigma^2(\tau_2)}. \tag{A8}$$

Given the constraint that both ω_1 and ω_2 be positive, ω' must also be positive. It follows that in Equation A8, the square root of the discriminant must be positive. Thus, the relevant solution is simply

$$\omega' = \frac{\sqrt{\sigma^2(\tau_1, \tau_2)(\kappa_1 - \kappa_2)^2 + 4\kappa_1 \kappa_2 \sigma^2(\tau_1)\sigma^2(\tau_2)} - \sigma(\tau_1, \tau_2)(\kappa_1 - \kappa_2)}{2\kappa_1 \sigma^2(\tau_2)}. \tag{A9}$$

Once ω' is known, $\omega' \omega_1$ may be substituted for ω_2 in Equation A1:

$$\kappa_1 = \omega_1^2 \sigma^2(\tau_1) + \omega_1^2 \omega' \sigma(\tau_1, \tau_2), \tag{A10}$$

which yields the formula for ω_1:

$$\omega_1 = \sqrt{\frac{\kappa_1}{\sigma^2(\tau_1) + w' \sigma(\tau_1, \tau_2)}}. \tag{A11}$$

Note that here too only the principal (i.e., positive) root of Equation A10 satisfies the constraint $\omega_1 \geq 0$.

If preferred, the result from Equation A9 may be used to rewrite Equation A11 in terms of κ_1, κ_2, $\sigma(\tau_1)$, $\sigma(\tau_2)$, and $\sigma(\tau_1, \tau_2)$ alone:

$$\omega_1 = \sqrt{\frac{2\kappa_1^2 \sigma^2(\tau_2)}{2\kappa_1 \sigma^2(\tau_1)\sigma^2(\tau_2) + \sigma(\tau_1, \tau_2)\sqrt{\sigma^2(\tau_1, \tau_2)(\kappa_1 - \kappa_2)^2 + 4\kappa_1 \kappa_2 \sigma^2(\tau_1)\sigma^2(\tau_2)} - \sigma^2(\tau_1, \tau_2)(\kappa_1 - \kappa_2)}}. \tag{A12}$$

Finally, ω_2 is simply the product of ω_1 and ω'; although, like ω_1, ω_2 may also be written in terms of κ_1, κ_2, $\sigma(\tau_1)$, $\sigma(\tau_2)$, and $\sigma(\tau_1, \tau_2)$ if desired:

$$\omega_2 = \frac{\sqrt{2\sigma^2(\tau_1,\tau_2)(\kappa_1-\kappa_2)^2+8\kappa_1\kappa_2\sigma^2(\tau_1)\sigma^2(\tau_2)}-\sqrt{2}\sigma(\tau_1,\tau_2)(\kappa_1-\kappa_2)}{2\sigma(\tau_2)\sqrt{\sigma(\tau_1,\tau_2)\sqrt{\sigma^2(\tau_1,\tau_2)(\kappa_1-\kappa_2)^2+4\kappa_1\kappa_2\sigma^2(\tau_1)\sigma^2(\tau_2)}+2\kappa_1\sigma^2(\tau_1)\sigma^2(\tau_2)-\sigma^2(\tau_1,\tau_2)(\kappa_1-\kappa_2)}}. \tag{A13}$$

Acknowledgments

I would like to express my gratitude to Su Baldwin of the National Board of Medical Examiners for alerting me to the problem addressed by this article. I am also delighted to thank the National Board of Medical Examiners for supporting this work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Although, strictly speaking, they are both based on expert opinion, some authors (e.g., Bobko, Roth, & Buster, 2007) have found it useful to distinguish between unit weights and all other weights based on expert judgments. In practice, the decision to use unit weights follows from a belief that differential weights are unnecessary. For this reason, it is unlikely that users of unit weights would apply the strategy for estimating nominal weights proposed in this article.
2. The notion that relative importance varies across components implies that the composite construct is multidimensional. Indeed, without multidimensionality, the choice of component weights does not affect the nature of the composite construct, it only affects its reliability. In such cases, it may be preferable to simply choose weights that maximize reliability or (if possible) to model both components together using a unidimensional model.
3. Here, the definition of effective weights espoused by Wang and Stanley (1970) is used; however, some later authors (e.g., Brennan, 2001; Clauser, Harik, & Margolis, 2006) prefer to define effective weights as the *proportion* of variance of the composite explained by each component—that is, $\omega_s^* = \kappa_s \sigma^{-2}(\tau)$, where ω_s^* is such a proportion for component s. As this more recent formulation may be viewed as a special case of Wang and Stanley's earlier conceptualization (specifically, the case wherein $\boldsymbol{\kappa}$ sums to unity), it is not treated separately here.
4. Under the constraint that effective weights and observed weights each sum to unity, an analytical solution for the two component condition can also be found in Clauser et al. (2006).
5. Note that the concern here is the relative weight applied to each component—not the absolute weight. For example, $\mathbf{w} = \begin{bmatrix} .25 \\ .50 \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} .40 \\ .80 \end{bmatrix}$ should be considered equivalent because in both cases, Component B has twice the weight as Component A. It follows that prior to comparing $\hat{\mathbf{w}}$ and $\mathbf{w}$, each weight was rescaled such that it represented the proportion of combined weights—that is, each set of weights was rescaled such that they summed to 1.
6. Note that because reliability was a study parameter (i.e., it was not estimated as part of the study design), these bootstrap results only account for sampling error due to simulees in the calculation of the observed score variances and covariances. In practice, one should also account for estimation error in the reliabilities.

References

American Institute of Certified Public Accountants. (2011). *How is the CPA exam scored?* Jersey City, NJ: Author.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, *10*, 689-709.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Burt, C. (1950). The influence of differential weighting. *British Journal of Statistical Psychology*, *3*, 105-125.

Clauser, B. E., Harik, P., & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. *Journal of Educational Measurement*, *43*, 173-191.

Corey, S. M. (1930). The effect of weighting exercises in a new-type examination. *Journal of Educational Psychology*, *21*, 383-385.

Dunnette, M. D., & Hoggatt, A. C. (1957). Deriving a composite score from several measures of the same attribute. *Educational and Psychological Measurement*, *17*, 423-434.

Edgerton, H., & Kolbe, E. L. (1936). The method of minimal variation for the combination of criteria. *Psychometrika*, *1*, 183-187.

Elster, R. S. (1972). *Controlling the influences of component variables*. Monterey, CA: Naval Postgraduate School.

Eva, K. W., & Reiter, H. I. (2004). Where judgement fails: Pitfalls in the selection process for medical personnel. *Advances in Health Sciences Education*, *9*, 161-174.

Govindarajulu, Z. (1988). Alternative methods for combining several test scores. *Educational and Psychological Measurement*, *48*, 53-60. doi:10.1177/001316448804800107

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, *1*, 53-60.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, *26*, 139-142.

Kaiser, H. F. (1967). Uncorrelated linear composites maximally related to a complex of correlated observations. *Educational and Psychological Measurement*, *27*, 3-6.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1-14.

Kreiter, C. D., Yin, P., Solow, C., & Brennan, R. L. (2004). Investigating the reliability of the medical school admissions interview. *Advances in Health Sciences Education*, *9*, 147-159.

Peel, E. A. (1947). A short method for calculating maximum battery reliability. *Nature*, *159*, 816-817.

Peel, E. A. (1948). Prediction of a complex criterion and battery reliability. *British Journal of Statistical Psychology*, *1*, 84-94.

Perloff, J. M., & Persons, J. B. (1988). Biases resulting from the use of indexes: An application to attributional style and depression. *Psychological Bulletin*, *103*, 95-104.

Ryans, D. G. (1954). An analysis and comparison of certain techniques for weighting criterion data. *Educational and Psychological Measurement*, *14*, 449-458.

Thomson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology*, *30*, 357-366.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213-217.

Wang, M., & Stanley J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*, 663-705.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*, 23-40.

Woodbury, M. A., & Lord, F. M. (1956). The most reliable composite with a specified true score. *British Journal of Statistical Psychology*, *9*, 21-28.