The Reliability of a Linear Composite of Nonequivalent Subtests

William W. Rozeboom
University of Alberta

Traditional formulas for estimating the reliability of a composite test from its internal item statistics are inappropriate to judge the reliability of multiple regressions and other weighted composites of subtests that are appreciably nonequivalent. Formulas are provided here for the reliability of such a composite given the reliabilities of its component subtests, followed by a comparison of the composite's reliability to that of its components. Compositing can easily incur a substantial loss of reliability, though gains are entirely possible as well. *Index terms: combining nonequivalent subtests, composite reliability, item weighting, nonequivalent subtests, nonhomogeneous item composites.*

Suppose that examinees' scores are interpreted on an m-tuple $X = (x_1, \ldots, x_m)$ of observational measures whose reliability coefficients in examinee population P are known (in practice, estimated) to be respectively $r_{x_1}, \ldots, r_{x_m}$. Some of these x_i may be predictors—such as age, gender, and socioeconomic status—that are treated as errorless, in which case $r_{x_i} = 1$.

Suppose also that these measures have been combined into a linear composite $x^* = w_1x_1 + \ldots + w_mx_m$ that is judged suitable for predicting some target variable y in P. (In practice, composite x^* may also include an additive constant suppressed here in light of its irrelevance to x^*'s reliability.) For example, weights $\{w_i\}$ may be coefficients in the estimated linear regression of y on X in P. Or they may have been chosen for practical convenience as when each w_i, if not simply unity, is an integer roughly proportional to $\sigma_{x_i}^{-1}$. Or it may be desired to study the sensitivity of x^*'s reliability to arbitrary assignment of its subtest weights. Whatever its motivation, the question is: How can the reliability of composite predictor x^* be determined simply from the subtest information already available?

Because subtests $x_1, \ldots, x_m$ are *not* presumed equivalent (i.e., they are not all thought to have nearly the same true-score component), the reliability of x^* cannot be appropriately estimated by any of the traditional internal-consistency formulas. Even so, under the classic presumption that measurement errors are uncorrelated between subtests, the reliability of x^* (in P) is straightforwardly determined by the subtests' reliabilities (in P), their empirical variances/covariances (in P), and their compositing weights.

Specifically, when subtests $\{x_i\}$ are theoretically decomposed into their true-score components $\{t_i\}$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 3, September 1989, pp. 277–283
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/030277-07$1.60

and measurement-error residuals $\{e_i\}$ uncorrelated with the former, that is,

$$x_i = t_i + e_i \qquad (i = 1, \ldots, m) \quad ; \tag{1}$$

$$\mathrm{Cov}(t_i, e_j) = 0 \qquad (i,j = 1, \ldots, m) \quad , \tag{2}$$

the individual subtests have reliability coefficients

$$r_{x_i} = \frac{\sigma_{t_i}^2}{\sigma_{x_i}^2} \qquad (i = 1, \ldots, m) \tag{3}$$

while the true-score/error composition of $x^* = t^* + e^*$ and its resultant reliability are

$$t^* = \sum_{i=1}^{m} w_i t_i \quad , \quad e^* = \sum_{i=1}^{m} w_i e_i \tag{4}$$

and

$$r_{x^*} = \frac{\sigma_{t^*}^2}{\sigma_{x^*}^2} \qquad \left(x^* = \sum_{i=1}^{m} w_i x_i \right) \tag{5}$$

respectively (see Rozeboom, 1966, p. 385ff.). The assumption of uncorrelated errors between subtests entails $\mathrm{Cov}(e_i, e_j) = 0$ for all $i \neq j$, so that

$$\sigma_{e^*}^2 = \mathrm{Var}\left(\sum_{i=1}^{m} w_i e_i \right) = \sum_{i=1}^{m} w_i^2 \sigma_{e_i}^2 \quad . \tag{6}$$

But also

$$\sigma_e^2 = \sigma_x^2 - \sigma_t^2 = \sigma_x^2 (1 - r_x) \tag{7}$$

for all $x = x^*, x_1, \ldots, x_m$. Therefore

$$\sigma_{x^*}^2 (1 - r_{x^*}) = \sum_{i=1}^{m} w_i^2 \sigma_{x_i}^2 (1 - r_{x_i}) \quad , \tag{8}$$

or

$$r_{x^*} = 1 - \frac{\sum_{i=1}^{m} w_i^2 \sigma_{x_i}^2 (1 - r_{x_i})}{\sigma_{x^*}^2} = 1 - \sum_{i=1}^{m} \left(\frac{w_i \sigma_{x_i}}{\sigma_{x^*}} \right)^2 (1 - r_{x_i}) \tag{9}$$ [1]

where

$$\sigma_{x^*}^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j \mathrm{Cov}(x_i, x_j) \quad . \tag{10}$$

An important special case of Equation 9 is where the weights in x^* are regression coefficients of predictor m-tuple X for an empirical criterion y. Assume that y and the x_i have all been standardized to unit variance, so that the compositing weights are standardized beta coefficients $\beta_1, \ldots, \beta_m$. Then the variance of the (standardized) regression $\hat{y} = \sum_{i=1}^{m} \beta_i x_i$ is the squared multiple correlation R_{yX}^2 of y with X. Thus, from Equation 9, the regression's reliability coefficient is

$$r_{\hat{y}} = 1 - \frac{\sum_{i=1}^{m} \beta_i^2 (1 - r_{x_i})}{R_{yX}^2} \quad . \tag{11}$$

[1] Similar formulas have appeared in the recent literature on generalizability theory. See in particular Jarjoura and Brennan, 1982; Brennan, 1983.

Of course, if y is literally what the researcher hopes to predict from $\dot{y}$, then learning this test's reliability adds nothing to what R_{yx} already reveals about its value. But it is more likely that y is simply an observable surrogate for some less accessible variable that is the real target; in that case, $r_{\dot{y}}^{1/2}$ remains useful as an upper bound on this test's validity for the latter.

An Empirical Example

The author's tardy recognition of need for composite-reliability formulas such as Equations 9 and 11 arose only recently when analyzing a classroom problem of test construction using material from Kingma (1983, 1984) on quantitative thinking in young children. Kingma's data were performance on 34 Piagetian tasks administered over a period of two days to each of 170 children ranging in age from 4 to 8 years. These were binary items categorized on grounds of task similarity as Conservation (13 items), Transitivity (7 items), Seriation (6 items), and Correspondence (8 items). Scores were also obtained for these examinees one year later on Thurstone's Primary Mental Abilities Quantitative (PMA-Q) subtest. The class project was to construct from the 34 Piagetian items, together perhaps with age and gender, a test predicting these children's PMA-Q performance a year later, and to appraise this test's reliability and validity. (Complications of sampling noise were ignored here.)

The structure of these 37 measures was evaluated by a technique (Rozeboom, in press) that allows latent factors to be extracted and obliquely rotated in a joint space with selected observation variables—in this application, age and gender—that are treated as causal sources of the observed output. The analysis revealed a remarkably clean solution with three latent factors in which all of the Conservation items loaded only on one factor, almost all of the Transitivity items only on the second, and almost all of the Seriation and Correspondence items only on the third. Meanwhile, PMA-Q loaded appreciably on the first and third latent factor but not on the second, and also showed a considerable age effect unaccounted for by age differences on the Piagetian factors. (Gender was everywhere irrelevant.)

These results suggested that the Piagetian items could best be compacted into three subtests, one summing the Conservation items (a homogeneous test of latent Factor 1), the second summing the Transitivity items (a homogeneous test of latent Factor 2), and the third summing the Seriation and Correspondence items (a homogeneous test of latent Factor 3). From there, the preferred composite predictor of PMA-Q should be this criterion's regression on the three Piagetian subtests together with age. (Considering PMA-Q's negligible loading on Factor 2, the Transitivity subtest's regression weight was expected to be small compared to the others.)

Although test/retest data for these examinees were lacking, the reliabilities of the Piagetian subtests can be estimated by their alpha coefficients (Cronbach, 1951; Rozeboom, 1966, p. 411ff.), while the reliability of age is presumably 1. The Kingma-sample correlations among these subtests and the criterion, together with the estimated subtest reliabilities, are given in Table 1 (x_1 = Age in months, x_2 = Conservation sum, x_3 = Transitivity sum, x_4 = Seriation and Correspondence sum, y = PMA-Q one year later).

Table 1
Correlations Among the Subtests
and Their Prediction Target,
and Alpha Reliabilities (Diagonal Entries)

	x_1	x_2	x_3	x_4	y
x_1	(1.0)	.610	.481	.753	.782
x_2		(.923)	.494	.686	.736
x_3			(.719)	.531	.498
x_4				(.907)	.779

Writing z_y and $Z = (z_1, \ldots, z_m)$ for the unit-variance scalings of y and $\{x_i\}$, the standardized regression $\dot{z}_y$ of the criterion on all four predictors is

$$\dot{z}_y = .382z_1 + .309z_2 + .019z_3 + .269z_4 \tag{12}$$

and the corresponding multiple correlation is

$$R_{yZ} = .863 \quad . \tag{13}$$

Therefore, from Equation 11 and the estimated subtest reliabilities, the reliability of this regression is estimated to be

$$r_{\dot{y}} = 1 - \frac{(.382)^2(1-1) + (.309)^2(1-.923) + (.019)^2(1-.719) + (.269)^2(1-.907)}{(.863)^2} = .988 \quad . \tag{14}$$

Given the statistics in Table 1, Equation 9 also allows easy estimation of reliability for any other composite of these subtests. For example, if for purposes of predicting PMA-Q the useless x_3 is ignored and unit weights are assigned to the remaining variance-normalized subtests, the variance of $z^* = z_1 + z_2 + z_3$ is $\sigma^2_{z^*} = 3.0 + 2(.610 + .753 + .686) = 7.098$. From Equation 9, the estimated reliability of z^* is then $r_{z^*} = 1 - [(1 - 1) + (1 - .923) + (1 - .907)]/7.098 = .976$. [The validity of z^* for predicting y is $r_{yz^*} = (.782 + .736 + .779)/(7.098)^{1/2} = .862$, showing that the more elaborate four-predictor regression gains nothing.]

When Does Compositing Enhance Reliability?

Equation 9 shows how to compute a composite test's reliability from its compositing weights $\{w_i\}$, its subtest covariances $\{\mathrm{Cov}(x_i,x_j)\}$, and the individual subtest reliabilities $\{r_{x_i}\}$. But for conceptual clarity rather than computational ease, all variables in $x^* = \sum_{i=1}^{m} w_i x_i$ are best rescaled to have unit variances (in P). This permits the composite to be rewritten as

$$z_{x^*} \equiv \frac{x^*}{\sigma_{x^*}} = \sum_{i=1}^{m} \left(\frac{w_i \sigma_{x_i}}{\sigma_{x^*}}\right)\left(\frac{x_i}{\sigma_{x_i}}\right) \quad , \tag{15}$$

or

$$z_{x^*} = \sum_{i=1}^{m} a_i z_i \qquad \left(a_i \equiv \frac{w_i \sigma_{x_i}}{\sigma_{x^*}} \quad , \quad z_i \equiv \frac{x_i}{\sigma_{x_i}}\right) \quad , \tag{16}$$

where $a_1, \ldots, a_m$ are the respective weights of variance-normalized subtests $z_1, \ldots, z_m$ in their variance-normalized composite z_{x^*}. Equation 9 then becomes

$$r_{x^*} = 1 - \sum_{i=1}^{m} a_i^2 (1 - r_{x_i}) \tag{17}$$

or equivalently, to make standardized scaling notationally explicit,

$$r_{z_{x^*}} = 1 - \sum_{i=1}^{m} a_i^2 (1 - r_{z_i}) \quad . \tag{18}$$

Let γ_{wX} be the sum of the squared standard-scale compositing weights, and let p_i be the proportionate contribution of each a_i^2 to γ_{wX}, that is,

$$\gamma_{wX} \equiv \sum_{i=1}^{m} a_i^2 = \sum_{i=1}^{m} \frac{w_i^2 \sigma^2_{x_i}}{\sigma^2_{x^*}} \quad , \tag{19}$$

$$p_i^2 \equiv \frac{a_i^2}{\gamma_{wX}} \quad . \tag{20}$$

This yields

$$r_{x^*} = 1-\gamma_{wX}\sum_{i=1}^{m} p_i(1-r_{x_i}) = 1-\gamma_{wX}\left(1-\sum_{i=1}^{m}p_i r_{x_i}\right) \quad , \qquad (21)$$

or more perspicuously

$$1-r_{x^*} = \gamma_{wX}(1-\tilde{r}_x) \quad , \qquad (22)$$

where

$$\tilde{r}_x \equiv \sum_{i=1}^{m} p_i r_{x_i} \quad . \qquad (23)$$

Because $\sum_{i=1}^{m} p_i = 1$ with each p_i non-negative, $\tilde{r}_x$ is simply a weighted average of the individual subtest reliabilities, with weights interpretable as the differential saliences assigned to the individual subtests in their composite. The term $1 - r_{x^*}$ can evidently be viewed as the *un*reliability of test x^*, with similar interpretations for $1 - r_{x_i}$ and $1 - \tilde{r}_x$. Thus Equation 22 shows the unreliability of composite x^* to be a salience-weighted average of the unreliabilities of its component subtests, modulated by a coefficient γ_{wX}. In the library of data-space test statistics especially germane to reliability, coefficient gamma warrants status comparable to coefficient alpha. [Indeed, $(m/[m - 1])(1 - \gamma_{wX})$ equals the alpha coefficient of weighted subtests $\{w_i x_i\}$.] But the nature of γ_{wX} must still be clarified.

To the extent that γ_{wX} is less than 1, r_{x^*} conforms to the classic expectation that the reliability of a composite test exceeds that of its components. But if the subtest weights in x^* drive γ_{wX} above 1, then Equation 22 reveals that the composite's reliability is inferior to that of its components. To appreciate how this can occur, write $g_i \equiv w_i x_i$ for each weighted subtest in $x^* = \sum_{i=1}^{m} w_i x_i = \sum_{i=1}^{m} g_i$. Then

$$a_i^2 = \frac{\sigma_{g_i}^2}{\sigma_{x^*}^2} \qquad (24)$$

where

$$\sigma_{x^*}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}(g_i,g_j) \quad . \qquad (25)$$

Thus the reciprocal of coefficient gamma can be written as

$$\gamma_{wX}^{-1} = \frac{\sum_i \sigma_{g_i}^2 + \sum_{\substack{i=1\\ i\neq j}}^{m}\sum_{j=1}^{m}\mathrm{Cov}(g_i,g_j)}{\sum_i \sigma_{g_i}^2} = 1+(m-1)\frac{c_{gg}}{v_g} \quad , \qquad (26)$$

where v_g is the mean variance of weighted subtests $\{g_i\}$, c_{gg} is their mean proper covariance (i.e., excluding self-covariances), and c_{gg}/v_g roughly equals the average correlation among the weighted subtests. Clearly $\gamma_{wX} < 1$ if $c_{gg} > 0$, in which case the compositing does indeed enhance reliability. But just as clearly, when the mean covariance between the weighted subtests is negative, γ_{wX} exceeds 1 and hence degrades the composite reliability.

But is not a negative value of c_{gg} a pathological condition that should seldom if ever arise? Not at all—at least not in principle. Suitably chosen subtest weights can align x^* with any axis in the space spanned by its subtests, and the c_{gg} expected from an arbitrary choice of x^*'s direction in this space is generally negative. Specifically (for γ_{wX} rather than directly for c_{gg}), let $\mathbf{Z} = (z_1, \ldots, z_m)$ be the column vector of variance-normalized but otherwise unweighted subtests in Equation 16, while $\mathbf{F} = (f_1, \ldots, f_m)$

comprises these subtests' variance-normalized principal factors identified by the eigenstructure of their covariance matrix $\mathbf{C}_{ZZ}$. Specifically, analyzing $\mathbf{C}_{ZZ}$ (equivalently, the unstandardized subtests' correlation matrix) as

$$\mathbf{C}_{ZZ} = \mathbf{T}\mathbf{D}_\lambda\mathbf{T}' \tag{27}$$

with $\mathbf{T}$ orthonormal (i.e., $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$) and $\mathbf{D}_\lambda$ a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_m$, let F be an m-tuple of data-space axes such that

$$\mathbf{Z} = (\mathbf{T}\mathbf{D}_\lambda^{1/2})F, \quad \mathbf{C}_{FF} = \mathbf{I} \quad . \tag{28}$$

(It is well known that such an F always exists for Z, albeit not altogether uniquely.) Setting $\mathbf{a} = (a_1, \ldots, a_m)$ for the row vector of subtest weights in the variance-normalized composite, $z_{x*} = \sum_{i=1}^m a_i z_i$ can be written more compactly as $\mathbf{a}Z$. It can easily be seen that this is equivalent to a weighted composite $z_{x*} = \sum_{i=1}^m b_i f_i = \mathbf{b}F$ of the subtests' principal factors. Specifically,

$$z_{x*} = \mathbf{a}Z = (\mathbf{a}\mathbf{T}\mathbf{D}_\lambda^{1/2})F = \mathbf{b}F \quad , \tag{29}$$

where

$$\mathbf{b} \equiv \mathbf{a}\mathbf{T}\mathbf{D}_\lambda^{1/2} \quad . \tag{30}$$

Presuming that all eigenvalues of $\mathbf{C}_{ZZ}$ are positive,[2] this yields $\mathbf{a} = \mathbf{b}\mathbf{D}_\lambda^{-1/2}\mathbf{T}'$ and hence

$$\gamma_{wX} = \mathbf{a}\mathbf{a}' = \mathbf{b}\mathbf{D}_\lambda^{-1/2}\mathbf{T}'\mathbf{T}\mathbf{D}_\lambda^{-1/2}\mathbf{b}' = \mathbf{b}\mathbf{D}_\lambda^{-1}\mathbf{b}' = \sum_{i=1}^{m} b_i^2\lambda_i^{-1} \quad . \tag{31}$$

It is easily seen that each b_i is the covariance of z_{x*} with the ith factor in F. Moreover, $\sum_{i=1}^m b_i^2 = 1$, because by scaling stipulation

$$1 = \sigma^2_{z_{x*}} = \mathbf{b}\mathbf{C}_{FF}\mathbf{b}' = \mathbf{b}\mathbf{I}\mathbf{b}' = \mathbf{b}\mathbf{b}' \quad . \tag{32}$$

Equation 31 thus shows γ_{wX} to be a weighted average of the eigenvalue reciprocals $\{\lambda_i^{-1}\}$, where the weight b_i^2 of λ_i^{-1} is the squared correlation of composite x^* with the ith principal factor of Z. Inasmuch as the unweighted average of eigenvalues $\{\lambda_i\}$ is

$$m^{-1}\mathrm{Tr}[\mathbf{C}_{ZZ}] = m^{-1}\sum_{i=1}^{m}\sigma^2_{f_i} = 1 \quad , \tag{33}$$

the unweighted average of their reciprocals cannot be less than 1 and may greatly exceed that if some of the λ_i are quite small.[3] Thus, although subtest weights that put x^* mainly in the subspace of its subtests' leading principal factors will achieve $\gamma_{wX} < 1$ as desired, a random choice of x^*'s orientation in its subtest-space when the λ_i vary appreciably is expected to yield $\gamma_{wX} > 1$ with r_{x*} correspondingly less than $\bar{r}_x$.

In short, the more highly subtests $x_1, \ldots, x_m$ are intercorrelated, the more likely it is that selection of compositing weights for a purpose other than trying to predict what is most common to them will incur a composite reliability less than the average reliability of the constituent subtests. It is entirely possible, though far from necessary even for subtests with a strong first factor, that this reliability loss is severe.

As a rule of thumb, the time to worry about degrading reliability by compositing subtests is when the latter are a positive manifold while broadly half of the compositing weights are negative. But such generalized apprehensions are pointless in particular applications. Instead, the researcher can simply fill

[2] This assumption is robustly realistic. But at the price of a more advanced argument, the essential conclusion to follow can also be reached without it.

[3] It is easy to show that for any concave monotonic transformation $\phi(c)$ of quantities $c_1, \ldots, c_m$, the mean value of $\{\phi(c_i)\}$ exceeds the mean of $\{c_i\}$ by an amount that becomes increasingly large as the nonlinearity of $\phi(c)$ over the range of $\{c_i\}$ becomes conspicuous.

in the right-hand terms of Equation 9 and judge whether the reliability that results on the left is tolerable for the purpose at hand.

Note. Empirical estimates of reliability are quite generally grounded in idealized presumptions about error independence and/or true-score equivalences, and practical applications of Equation 9 are no exception to this rule. First, use of Equation 9 requires that each subtest reliability r_{x_i} be estimated somehow (preferably by test/retest studies, but by some internal-consistency measure if necessary), and the imperfections of these evidently contaminate the estimate of r_{x*} computed by Equation 9 as well. Second, although Equation 9 assumes nothing about subtest equivalence, it does presume zero correlations among measurement errors between subtests; as is (or should be) well known, this is flagrantly unrealistic for tests taken by the examinee at the same time and place.

How severely internal-consistency estimates of reliability are inflated by correlated errors is still an empirical unknown. But when the subtests $\{x_i\}$ composited in $x*$ are not mere subgroupings of items administered at a single sitting but rather are different kinds of measures obtained in diverse ways and varied settings, Equation 9's presumption of uncorrelated errors between subtests should be as true to fact as this classic ideal ever approaches in practice.

References

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City IA: American College Testing Program.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6*, 161–171.

Kingma, J. (1983). Seriation, correspondence, and transitivity. *Journal of Educational Psychology, 75*, 763–771.

Kingma, J. (1984). Criterion problems in conservation research reconsidered from a psychometric point of view. *Journal of General Psychology, 111*, 109–129.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood IL: Dorsey Press.

Rozeboom, W. W. (in press). HYBALL: A method for subspace-constrained oblique factor rotation. *Multivariate Behavioral Research*.

Acknowledgments

Johannes Kingma graciously made available the data on which the empirical example was based.

Author's Address

Send requests for reprints or further information to William W. Rozeboom, Center for Theoretical Psychology, University of Alberta, Edmonton, Alberta T6G 2E9, Canada.