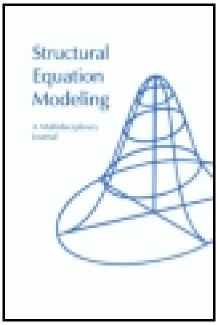

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

Point and Interval Estimation of Reliability for Multiple-Component Measuring Instruments via Linear Constraint Covariance Structure Modeling

Tenko Raykov

Available online: 19 Nov 2009

STRUCTURAL EQUATION MODELING, *11*(3), 342–356

Point and Interval Estimation of Reliability for Multiple-Component Measuring Instruments via Linear Constraint Covariance Structure Modeling

Tenko Raykov
Fordham University

A widely and readily applicable covariance structure modeling approach is outlined that allows point and interval estimation of scale reliability with fixed components. The procedure employs only linear constraints introduced in a congeneric model, which after reparameterization permit expression of composite reliability as a function of appropriate parameters. Unlike the popular Cronbach's coefficient alpha that already at the population level is generally a misestimator of scale reliability, this method is based on the formal definition of the reliability coefficient. The proposed approach is illustrated by means of a numerical example.

Multiple-component measuring instruments are highly popular in social, behavioral, and educational research primarily due to their being constructed to provide converging pieces of information about underlying latent dimensions that can only be measured with fallible indicators. The current, widely followed practice of their reliability estimation is based mainly on Cronbach's coefficient alpha (Cronbach, 1951). As has been known for more than 35 years, however, unless practically rather restrictive conditions are fulfilled (viz., essential tau equivalence of the components), alpha is not equal to the overall composite's reliability even at the population level (e.g., McDonald, 1999; Novick & Lewis, 1967; Raykov, 1997b; Zimmerman, 1972). The goal of this article is to contribute to the methodology of scale reliability evaluation by providing a widely applicable approach developed within the popular covariance structure modeling (CSM) methodology, which permits point and interval estimation of reliability for composites with prespecified components and is easily employed by the general social or behavioral researcher with any CSM software allowing introduction only of linear constraints.

Requests for reprints should be sent to Tenko Raykov, Department of Psychology, Fordham University, Bronx, NY 10458. E-mail: raykov@fordham.edu

MOTIVATION

As is well documented in the methodological literature, a long-standing tension has existed between coefficient alpha and the fact that, in general, it does not represent a dependable estimator of a scale's reliability. A number of researchers have shown over the past 35 years or so, beginning with the instructive article by Novick and Lewis (1967), that already at population level alpha has certain deficiencies making reliance on it, in general, a suboptimal strategy; and, in fact, one that can be misleading. Specifically, as demonstrated by Novick and Lewis and Zimmerman (1972), as well as further elaborated more recently (e.g., Green & Hershberger, 2000; Komaroff, 1997; McDonald, 1981; Raykov, 1997b, 1998a, 2001a, 2001b),

1. With uncorrelated errors of measurement, alpha is a lower bound of composite reliability regardless of the factorial structure of the instrument.
2. With such error terms, alpha is only then equal to scale reliability when the components are (essentially) tau equivalent; however, this strong requirement is hard to reconcile with the well-known fact that social and behavioral assessment is typically carried out in arbitrary units of measurement (whereas the requirement necessitates them to be, in addition, equal to one another).
3. With correlated errors, alpha can be an underestimate or, conversely, can be an overestimate of composite reliability.

This article is based on the logical premise that a researcher interested in multicomponent instrument reliability is, in fact, concerned with the composite reliability coefficient itself, not with any other quantity—like alpha—that only under generally rather restrictive conditions equals reliability (Raykov, 2003a). Hence, unlike much of past and present empirical research in the social, behavioral, and educational disciplines, this article is in full agreement with what could be considered a fundamental principle of science: Questions phrased in terms of concept A ought to be answered in terms of the same concept A, not in terms of another concept B that only in special cases equal A. Therefore, the article develops independently of coefficient alpha and is focused on the scale reliability coefficient that is the quantity of actual relevance when questions about reliability of multiple-component measuring instruments are to be answered.

POINT ESTIMATION OF SCALE RELIABILITY

The following developments utilize the classical definition of reliability as the ratio of true variance to observed variance (e.g., Lord & Novick, 1968). Therefore, the reliability coefficient can be viewed as an overall (unconditional) index of "precision" of measurement, which represents the proportion of individual ability differences in the observed score variance.

Notation and Definitions

For the specific purposes of this article, we assume that a set of components Y_1, Y_2, …, Y_k is given ($k > 1$) and that one is sampling from a participant population in which interest is in the reliability of the overall scale score $Y = Y_1 + Y_2 + \ldots + Y_k$ (Type 1 sampling, Lord, 1955; the case of weighted scales is included in the following). These components can be parts of an overall test or test battery, testlets, subscales, portions of exam papers, questions, self-report sections, or items. We are concerned with point and interval estimation of the scale's reliability coefficient, ρ_Y, defined as

$$\rho_Y = Var(T)/Var(Y)$$

where T_1, T_2, … , T_k are the true scores of the components Y_1, Y_2, … , Y_k, respectively; $T = T_1 + T_2 + \ldots + T_k$ is the true score pertaining to the observed score Y; and $Var(.)$ denotes variance in the studied population (Lord & Novick, 1968; Zimmerman, 1975).

As is frequently the case in empirical social and behavioral research, we assume that the components Y_1, Y_2, …, Y_k are congeneric (Jöreskog, 1971; cf. Raykov, 2003a). These measures represent the most general case of assessment of a common underlying dimension with possibly different units of measurement and precision or error variances. Then, by definition

$$Y_j = a_j + b_j\eta + E_j \tag{1}$$

holds true, where a_j and b_j are appropriate constants, η is the common true or latent variable score (e.g., $\eta = T_1$ can be taken), and $E_j = Y_j - T_j$ are the corresponding error scores ($j = 1, 2, \ldots, k$; Lord & Novick, 1968). For identifiability reasons, we assume $Var(\eta) = 1$. All models referred to in this article are also presumed to be identified (e.g., in the case $k = 2$, additional identifying restrictions must be imposed), and the assumption of uncorrelated errors is not needed.

In the congeneric case under consideration, the reliability coefficient of the scale score Y is readily shown to equal

$$\rho_y = \frac{(\sum_{i=1}^{k} b_i)^2}{(\sum_{i=1}^{k} b_i)^2 + \sum_{i=1}^{k} \theta_{ii}} \tag{2}$$

where $\theta_{jj} = Var(E_j)$ denotes the error variances ($j = 1, 2, \ldots, k$; e.g., Bollen, 1989; numerators and denominators of reliability coefficients are assumed throughout distinct from zero, a typically fulfilled assumption in empirical research). With correlated errors (e.g., Williams & Zimmerman, 1996),

$$\rho_Y = \frac{(\sum_{i=1}^{k} b_i)^2}{(\sum_{i=1}^{k} b_i)^2 + \sum_{i=1}^{k} \theta_{ii} + 2 \sum_{1 \le i < j \le k} \theta_{ij}} \tag{3}$$

holds true, where $\theta_{ij}(1 \le i < j \le k)$ are the nonzero error covariances and i and j vary across all pairs of correlated errors (given model identification). For a weighted test score, $Y = w_1Y_1 + w_2Y_2 + \ldots + w_kY_k$, where the weights $w_1, w_2, \ldots, w_k$ are known beforehand or to be estimated simultaneously as, say, equal to the indicator loadings (i.e., $w_j = b_j; j = 1, \ldots, k$) or to other quantities with certain properties of interest (e.g., Raykov, 2004),

$$\rho_Y = \frac{(\sum_{i=1}^{k} w_i b_i)^2}{(\sum_{i=1}^{k} w_i b_i)^2 + \sum_{i=1}^{k} w_i^2 \theta_{ii}} \tag{4}$$

holds in the uncorrelated error case, and in that of correlated errors

$$\rho_Y = \frac{\sum_{i=1}^{k} (w_i b_i)^2}{\sum_{i=1}^{k} (w_i b_i)^2 + \sum_{i=1}^{k} w_i^2 \theta_{ii} + 2 \sum_{1 \le i < j \le j} w_i w_j \theta_{ij}} \tag{5}$$

with the same notation for the error covariances as in Equation 3. Therefore, for the purposes of this article, the weighted scale case in Equations 4 and 5 is directly reducible to the initially described unweighted corresponding case via appropriate substitutions (e.g., Raykov, 2001b); therefore, we can focus on the remainder in the unweighted case covered by Equations 2 and 3 (i.e., when $w_j = 1, j = 1, \ldots, k$; Raykov, 2003a).

Reliability Re-Expression

In this subsection, we describe the basis of a widely applicable method of estimation of scale reliability, which can be readily used by the general social or behavioral researcher and is outlined later in detail. A main feature of this method, which sets it apart from earlier ones (e.g., Feldt, Woodruff, & Salih, 1987; Raykov, 1998b, 2002a), is that it allows both point and interval estimation of composite reliability—rather than coefficient alpha—within a single modeling session using a readily applicable CSM approach.[1]

To describe the method, we first observe that with the assumptions made so far—which represent practically no restriction of generality and can be presumed fulfilled in empirical research—Equation 2 is equivalent to the equality of the reciprocals of its right- and left-hand sides (for simplicity of notation, the summation index i running from 1 to k is dropped in this and in the next subsections):

$$\rho_Y^{-1} = 1 + (\Sigma\theta_{ii})/(\Sigma b_i)^2 \tag{6}$$

Denoting $C = (\Sigma b_i)^2$, Equation 6 is rewritten as

$$\rho_Y^{-1} = 1 + \Sigma\theta^*_{ii} \tag{7}$$

where we set

$$\theta^*_{jj} = \theta_{jj}/C \ (j = 1, \ldots, k) \tag{8}$$

Equations 7 and 8 suggest a useful reparameterization of the model defined in Equation 1 that we conceptualize as a covariance structure model (which obviously represents no loss of generality; cf. Raykov, 2002b). The reparameterization is accomplished by a rescaling of the metric of the error terms, E_j, which consists of a division of their units of measurement by $\Sigma b_i = C^{1/2}$, and is of high practical utility in its own right. After this rescaling, the error term variances become θ_{jj}/C; that is, the error variances of the reparameterized model are θ^*_{jj} defined in Equation 8 ($j = 1, \ldots, k$). Indeed, the model definition in Equation 1 can be rewritten as

$$Y_j = a_j + b_j\eta + (b_1 + b_2 + \ldots + b_k)E^*_j \tag{9}$$

where $E^*_j = E_j/(b_1 + b_2 + \ldots + b_k)$; hence, $Var(E^*_j) = Var(E_j)/(b_1 + b_2 + \ldots + b_k)^2 = \theta_{jj}/C = \theta^*_{jj}$ (see Equation 8; $j = 1, \ldots, k$). Also, note then that all initial error covariances, if any, become rescaled in the same manner (i.e., divided by C). Furthermore, this reparameterization only changes the scale of the error variances (and covariances, if any), but not that of the construct loadings or the latent variance assumed fixed at one. Moreover, reparameterization (Equation 9) does not change the model identification status—an identified congeneric model remains identified after this rescaling (demonstration available from the author on request).

[1]The method described in this and the next subsection is not the only alternative to coefficient alpha, which allows reliability estimation using its coefficient formula (Equation 3). Raykov (1997a, 2001a) described other approaches based on Equation 3 that were, however, either based on nonlinear constraints or highly computer intensive when employed for purposes of interval estimation. Unlike those methods, the one in this article utilizes only linear constraints and yields, within a single modeling session, also an interval estimate in addition to a point estimate of scale reliability. Another approach based on Equation 3 was indicated by a referee, which estimates reliability as the complement to one of the ratio of reproduced error variance for the scale score Y to the latter's observed variance. In difference to this estimation procedure, however, that approach does not yield in the same modeling session all quantities needed to obtain a composite reliability confidence interval (see the following in main text).

Reparameterization (Equation 9) is practically accomplished, in terms of the covariance structure model (Equation 1) of concern, by restricting the relations between error terms and corresponding observed variables. Specifically, setting the path from each error, E_j, into its corresponding manifest variable, Y_j, to be equal to the sum of construct loadings (i.e., to $C^{1/2} = \Sigma b_i$) achieves this reparameterization ($j = 1, \ldots, k$). The resulting model is depicted in Figure 1 following widely used conventions for graphical display of structural models (e.g., Jöreskog & Sörbom, 1996).

We note that in the congeneric model of Figure 1, only linear constraints are imposed that can be introduced readily with widely available CSM programs (e.g., LISREL, Jöreskog & Sörbom, 1996, see Appendix; EQS, Bentler, 2004; M*plus*, Muthén & Muthén, 2004; and similarly, Mx, Neale, 1997).

Point Estimation of Scale Reliability

Once the reparameterization in Equation 9 is carried out, fitting the resulting model (i.e., Equation 1 with Equation 9) to the analyzed data using LIREL, EQS, M*plus*, or Mx, one obtains estimates of the rescaled error variances $\theta^*_{jj}(j = 1, \ldots, k)$. Then, according to Equation 7, a point estimate of the overall scale score's reliability results by adding 1 to the sum of these k estimates and taking the inverse. That is, the scale reliability estimator of this article is

$$\hat{\rho}_Y = (1 + \sum_{i=1}^{k} \hat{\theta}_{ii}{}^*)^{-1} \tag{10}$$

where a caret signals estimator of pertinent parameter.

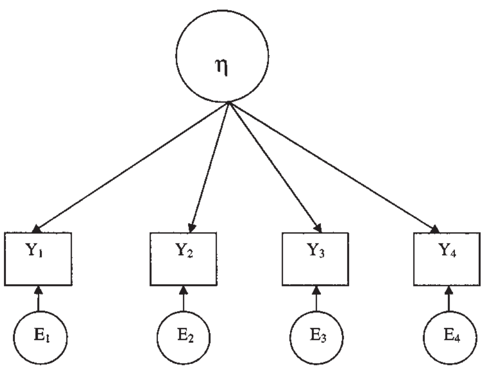

FIGURE 1 Reparameterized congeneric component model (see Equations 1 and 9). The path from each error term E_j into its corresponding observed variable Y_j is set equal to the sum of the loadings of all observed variables on the latent variable η; that is, to $C^{1/2}(j = 1, \ldots, k$; see Equations 6 and 7 and the Appendix for details of implementation).

In empirical research, all quantities appearing on the right-hand side of Equation 10 are found in pertinent sections of the software output associated with the model (see illustration section). Note that if the maximum likelihood (ML) method has been used for model fitting, due to the invariance property of ML estimators (e.g., Rao, 1973) $\hat{\rho}_Y$ in Equation 10 is a ML estimator of scale reliability as well and hence enjoys all properties of such estimators (Raykov, 2003a). In particular, the estimator in Equation 10 is consistent, efficient, and normally distributed (with mean being the true reliability coefficient; Rao, 1973). Therefore, with large samples, the overall test reliability point estimator (Equation 10) has highly desirable optimality properties that ensure its utility in social and behavioral research. Note that the method of this section is distinct from the ones described in Raykov (1997a) and Raykov (2001b) in that it (a) estimates directly the test reliability coefficient and (b) achieves this via simple linear rather than nonlinear constraints and, hence, can be employed also with many popular CSM programs. Moreover, this procedure (c) is substantially easier to apply than either of those two earlier approaches, and, in addition, (d) provides as a byproduct all quantities needed for obtaining a large sample confidence interval of the scale reliability coefficient as outlined next.

INTERVAL ESTIMATION OF SCALE RELIABILITY

The described method for point estimation of scale reliability yields only a numerical guess, in the form of a single number, about the population value of the reliability coefficient (Raykov, 2003a). Although this is an optimal estimate, realizing that in general it does not equal its population counterpart, one is still left wondering how far it may lie from that true reliability coefficient of actual interest. Interval estimation answers this question by furnishing a plausible range of values for this population quantity. Over the past several decades, other methods have also been developed that aimed at interval estimation of scale reliability. Apparently, highest popularity have been procedures for constructing confidence intervals of coefficient alpha (e.g., Feldt, 1965; Feldt et al., 1987). Due to the previously explicated limitations of coefficient alpha, however, in particular its population slippage relative to the scale reliability coefficient of real concern (e.g., Raykov, 1997b), as well as the fact that it has, in general, a different sampling distribution, it is easily observed that the resulting intervals would attain prespecified confidence levels as such intervals for composite reliability only in the frequently unrealistically restrictive case of essential tau equivalence (i.e., all components assessing the same latent construct with the same units of measurement). Therefore, in the general case of congeneric measures of concern in this article, the so-obtained intervals for alpha would be, possibly, with intolerably different from nominal coverage of the true scale reliability of interest; this may suggest misleading conclusions about the used instrument precision of measurement. As an alternative to this traditional methodology, Raykov (1998b) discussed a bootstrap method for approximate evaluation of a standard error and confidence interval of scale reliability. The method rests on a computer-intensive procedure that remains rather time and resource consuming for routine applications in empirical research. More recently, Raykov (2002a) outlined an analytic method for approximate interval estimation of scale reliability that, however, is based on a two-step approach capitalizing on linear approximation of the nonlinear reliability function of construct loadings and error variances (Equation 2).

These limitations of alternative methods for interval estimation of composite reliability are readily resolved by the CSM approach outlined in the preceding section, as shown next.

Confidence Interval of Rescaled Error Variance Sum

Examining again Equation 7, one sees that if (L, U) is a $(1 - \gamma)$ 100% confidence interval for its right-hand side, then $(1/U, 1/L)$ is a $(1 - \gamma)$ 100% confidence interval for the scale reliability coefficient ρ_Y $(0 < \gamma < 1)$. Therefore, all one needs to do to obtain a reliability confidence interval is to invert, and exchange places of, the lower and upper confidence interval limits of the right-hand side of Equation 7. To this end, denote by

$$V = \sum_{i=1}^{k} \theta_{ii}{}^{*} \tag{11}$$

the sum of rescaled error variances in the reparameterized congeneric test model defined by Equations 1 and 9. To obtain a standard error of V (denoted "S.E." in the following), take the square root of the variance of the right-hand side of Equation 11:

$$S.E.(V) = \sqrt{Var(\theta^{*}_{11}) + \ldots + Var(\theta^{*}_{kk}) + 2Cov(\theta^{*}_{11}, \theta^{*}_{22}) + 2Cov(\theta^{*}_{11}, \theta^{*}_{33}) + \ldots + 2Cov(\theta^{*}_{k-1,k-1}, \theta^{*}_{kk})} \tag{12}$$

applying the well-known expression for variance of a sum of random variables (e.g., Hays, 1994), where $Cov(.,.)$ denotes covariance in the population. Using the fact that the covariance structure model parameter estimator is asymptotically normal with mean the estimated true parameter (e.g., Bollen, 1989), and thus V is so as well (see Equation 11); based on the empirical data, a large sample $(1 - \gamma)$ 100% confidence interval for V is obtained as

$$(\hat{V} - z_{\gamma/2} \times S.E.(\hat{V}); \hat{V} + z_{\gamma/2} \times S.E.(\hat{V})) \tag{13}$$

where "×" denotes multiplication, $z_{\gamma/2}$ is the pertinent quantile of the standard normal distribution (e.g., $z_{\gamma/2}$ = 1.96 if a 95% confidence interval is sought, or $z_{\gamma/2}$ = 1.64 for a 90% confidence interval); and $S.E.\ (\hat{V})$ is the value of the right-hand side of Equation 12 when variances and covariances of the error variance estimates appearing there are substituted. These variances and covariances are obtained from the information matrix at the solution point (e.g., Bollen, 1989) and are found, correspondingly, in the covariance matrix of parameter estimate section in the output of used CSM software (see illustration section and Appendix).

Confidence Interval of the Composite Reliability Coefficient

With the lower and upper limits of the confidence interval (Equation 13), as indicated earlier, Equation 6 yields the following large sample (1 – γ) 100% confidence interval for the reliability coefficient of the overall scale score:

$$[(1 + \hat{V} + z_{\gamma/2} \times S.E.(\hat{V}))^{-1}; (1 + \hat{V} - z_{\gamma/2} \times S.E.(\hat{V}))^{-1}] \tag{14}$$

(The initial and final brackets are meant only to be symbolically distinct from the parentheses appearing within them, rather than denote a closed interval.)

The outlined point and interval estimation method for composite reliability is demonstrated next on a numerical example.

ILLUSTRATION ON DATA

Here, simulated data are used to demonstrate an application of the proposed scale reliability point and interval estimation procedure. We are employing such data because they have the advantage that due to being generated beforehand, one knows all underlying model parameters and can therefore determine the true scale reliability coefficient as well as compare it with the one estimated using the described method. To this end, multivariate normal zero-mean data is simulated on $k = 4$ variables on $N = 400$ cases according to the following model:

$$\begin{aligned} Y_1 &= \eta_1 + E_1 \\ Y_2 &= 1.5\ \eta_1 + E_2 \\ Y_3 &= 2\ \eta_1 + E_3 \\ Y_4 &= .1\ \eta_1 + E_4 \end{aligned} \tag{15}$$

where $Var(\eta_1) = 1$ is set, and the error covariance matrix is taken as diagonal with consecutive diagonal elements 1, 2, 3, and 4 (for E_1 through E_4, respectively; for details on the data simulation procedure, see Raykov, Marcoulides, & Boyd, 2003).

The covariance matrix of the simulated data is displayed in Table 1. (The source code needed for the following analyses is provided in the Appendix.)

Fitting to this covariance matrix the congeneric test model in Equation 1 with reparameterization (Equation 9) yields, not surprisingly, acceptable goodness-of-fit indexes: $\chi^2 = .18$, $df = 2$, $p = .91$, root mean square error of approximation = 0 with a 90% confidence interval (0; .03; e.g., Jöreskog & Sörbom, 1996; the same goodness-of-fit indexes are obtained without that reparameterization because it does not have any implications for the covariance structure and thus has no effect on model fit). The error variance estimates are found to be as follows (standard errors in parentheses): 0.0354 (0.0053), 0.0894 (0.0133), 0.1393 (0.0228), and 0.1289 (0.0158) for E_1 through E_4, respectively. Using Equation 10, the estimate of reliability of the overall test score $Y = Y_1 + Y_2 + Y_3 + Y_4$ of interest results with these estimates as

$$\hat{\rho}_Y = (1 + \sum_{i=1}^{k} \hat{\theta}_{ii}{}^*)^{-1} = (1 + 0.0354 + 0.0894 + 0.1393 + 0.1289)^{-1} = (1 + .393)^{-1} = .718. \tag{16}$$

Because we know all model parameters in this case, the true reliability of Y is determined using Equation 2:

$$\rho_Y = \frac{(\sum_{i=1}^{k} b_i)^2}{(\sum_{i=1}^{k} b_i)^2 + \sum_{i=1}^{k} \theta_{ii}} = (1 + 1.5 + 2 + .1)^2/[(1 + 1.5 + 2 + .1)^2 + 1 + 2 + 3 + 3] = 21.16/30.16 = .702 \tag{17}$$

that is quite close to its estimated reliability (see Equation 16). (By comparison, the estimate of coefficient alpha results for this data set as .631 is considerably lower, a finding consistent with the literature on limitations of alpha; e.g., Novick & Lewis, 1967; Raykov, 1997b; Zimmerman et al., 1973; and references therein.)

To obtain a 95% confidence interval for the reliability of Y, we use Equations 12 through 14. First, according to Equation 12, the standard error of the sum V or rescaled error variances in the right-hand side of Equation 7 results as the sum of the corresponding diagonal entries of the inverted information matrix provided in Table 2, plus twice the sum of pertinent off-diagonal elements in that matrix. Note

TABLE 1
Covariance Matrix of Simulated Data ($N = 400$)

Variable	Y_1	Y_2	Y_3	Y_4
Y_1	1.8870			
Y_2	1.5744	4.3173		
Y_3	2.1767	3.1596	7.5345	
Y_4	0.1015	0.2136	0.2690	2.9428

that due to the reparameterization in Equation 9, the rescaled error variances, that is, the consecutive θ^*s, are referred to as PS(2,2) through PS(5,5) in the LISREL source code in the Appendix. The resulting standard error of the error variance sum V in the fitted model is .0424 (see Equation 11).

With this standard error of V, a 95% confidence interval for the right-hand side of Equation 7 is obtained as [1 + .393 – 1.96 × .0424; 1 + .393 +1.96 × .0424] = [1.3099; 1.476]. Finally, using Equation 6, a 95% confidence interval for the scale score's reliability coefficient ρ_Y is furnished by inverting (and exchanging places of) the lower and upper limit of the last interval, leading to

$$[.6775; .7634]. \tag{18}$$

that covers the true reliability of Y, .702 (see Equation 17).

Returning to the results of the initial CSM analysis, the loading of the fourth scale component, b_4, is found to be nonsignificant: Its estimate is 0.1215, with a standard error of .0962 (t = 1.263). This suggests that a revised version of the four-component scale, resulting after deleting the last component Y_4, might be associated with considerably higher reliability. To explore this hypothesis, we fit the model only to the first three observed variables. In the same way as highlighted earlier, the estimated reliability of $Y_{(3)} = Y_1 + Y_2 + Y_3$ is found to be .782. (The true reliability of $Y_{(3)}$ is determined with Equation 2 to be 20.25/26.25 = .7714, which is very close.) The 95% confidence interval for the reliability of $Y_{(3)}$ similarly results as [.7462; .8222] and is markedly above that for the initial scale score Y in (18). This finding possibly suggests considerable increase in re-

TABLE 2
Covariance Matrix of Parameter Estimates

Parameter	λ_{11}	λ_{21}	λ_{31}	λ_{41}	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}
λ_{11}	0.0046								
λ_{21}	0.0015	0.0106							
λ_{31}	0.0017	0.0029	0.0184						
λ_{41}	0.0001	0.0002	0.0003	0.0093					
y_{22}	–0.0002	–0.0002	–0.0003	–0.0001	**0.0000**				
ψ_{33}	–0.0002	–0.0009	–0.0007	–0.0004	**0.0000**	**0.0002**			
ψ_{44}	–0.0003	–0.0007	–0.0023	–0.0006	**0.0000**	**0.0001**	**0.0005**		
ψ_{55}	–0.0004	–0.0008	–0.0013	–0.0006	**0.0000**	**0.0001**	**0.0002**	**0.0003**	
λ_{12}	0.0079	0.0151	0.0233	0.0099	–0.0008	–0.0022	–0.0038	–0.0031	0.0532

Note. LISREL notation is used to symbolize model parameters: λ_{11} through λ_{41} are the indicator loadings associated with Y_1 through Y_4, respectively; ψ_{22} through ψ_{55} the rescaled error variances pertaining to E_1 through E_4, respectively; and λ_{12} is the path connecting E_1 with Y_1 (see Appendix for software code details; matrix found in output section titled "Covariance matrix of parameter estimates"). Numbers in bold are the entries needed for interval estimation of the sum V of rescaled error variances (see Equations 11 and 12).

liability after deletion of Y_4 from the original composite (a correct suggestion given the notable true reliability increase of .064 = .782 – .718). We emphasize that the last interval comparison is not a statistical test of the null hypothesis of equality of scale reliability before and after deletion of Y_4; in fact, comparison of these nonsimultaneous confidence intervals cannot yield conclusive information allowing testing of this hypothesis. A formal test of the hypothesis of no reliability change after scale revision is accomplished with the procedure described in Raykov and Grayson (2003), which should be applied on a sample independent from the one used for this hypothesis's generation.

DISCUSSION AND CONCLUSION

This article outlined a widely applicable CSM method of point and interval reliability estimation for multiple-component measuring instruments, which is readily utilized by the general social, behavioral, or educational researcher. The approach is based only on linear constraints imposed in a congeneric test model and can be used with any CSM software capable of introducing these restrictions, such as LISREL, EQS, M*plus*, and Mx (see also Footnote 1). By comparing pertinent reliability interval estimates, the method may also be used to generate hypotheses about (a) possible change in reliability of scales undergoing development (for a formal test of the hypotheses of change on an independent sample, regardless whether components are added or deleted and irrespective of their position in the scale, see Raykov & Grayson, 2003), (b) group differences in composite reliability (for a testing procedure of a so-generated hypothesis on independent samples, applicable both to related and unrelated groups, see Raykov, 2002b), and (c) discrepancies in reliability of different versions of a measuring instrument (e.g., paper-and-pencil vs. computer-based forms administered to related or unrelated groups; formal testing of such a hypothesis can be conducted with the method in Raykov, 2002b, on an independent sample from the one used to generate it).

A limitation of the proposed method stems from the fact that it is based on the CSM methodology that is a large-sample modeling approach. Therefore, the outlined estimation procedure is best used with large samples (in general, at least with several hundred participants), and similarly with approximately continuous components. Such components may be obtained by initial exploratory factor analysis of the polychoric (tetrachoric) correlation matrix of possibly highly discrete items (on an independent or halved sample; Jöreskog & Sörbom, 1996; Raykov, 2003a) and subsequent construction of pertinent parcel scores, a procedure frequently used in practice; alternatively, in case of unidimensional initial scale, one could build two or more components through random or appropriate split. It is worthwhile emphasizing that this requirement for large samples pertains only to the number of studied participants and not to the number of components, which can be as low as three in an original scale (cf. item response theory-based methods for scale construction and revision that are best used with a large number of items; e.g., Lord, 1980).

In conclusion, a natural question that can be raised at this point is when one could use coefficient alpha and have confidence in its results (Raykov, 2003a). In the context of this article and, in particular, based on research over the past several decades that bears on this query, one may suggest that alpha be considered in empirical settings with discrete data (e.g., dichotomous or trichotomous items) and uncorrelated errors and when, for substantive or other reasons, one may not be willing to parcel the initial set of scale items as indicated earlier; in which case, one can make use of alpha's lower bound feature. A potential problem with applying alpha then is that no widely and easily applicable test of uncorrelated measurement errors seems to be available with such data; yet, as mentioned earlier and has been well-documented in the literature, alpha can also overestimate test reliability with correlated error terms. Alternatively, with approximately continuous data (whether initial data or such arrived at after parceling carried out as outlined earlier or differently), it could be suggested that there is no pressing need to use alpha unless one has (a) a relatively large number of components (more than ½ doz., say); that (b) load highly uniformly on a common latent construct (at least around .6 on a 0–1 metric); and (c) uncorrelated measurement errors (Raykov, 1997b). At least one of conditions (a) through (c) appears likely to be violated especially in early work on scale construction and development in social and behavioral research, however. Such violations should make one very cautious about applications of alpha and consider, instead, use of the scale reliability point and interval estimation method described in this article (see also Footnote 1).

ACKNOWLEDGMENTS

This research was supported by grants from The Law School Admission Council and The College Entrance and Examination Board.

I am indebted to C. Lewis for valuable and stimulating discussions on reliability evaluation. I thank the editor and two anonymous referees for helpful criticism and suggestions on an earlier draft.

REFERENCES

Bentler, P. M. (2004). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

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

Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11,* 93–103.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7,* 251–270.

Hays, W. L. (1994). *Statistics.* Orlando, FL: Harcourt Brace.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and correlated errors on coefficient alpha. *Applied Psychological Measurement, 21,* 337–348.

Lord, F. M. (1955). Sampling fluctuations resulting from sampling of items. *Psychometrika, 20,* 1–22.

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

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Readings, MA: Addison-Wesley.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Muthén, B. O., & Muthén, L. K. (2004). *MPLUS user's guide.* Los Angeles: Author.

Neale, M. (1997). *Mx user's guide.* Richmond, VA: Virginia Commonwealth University, Department of Psychiatry.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika, 32,* 1–13.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 22,* 173–184.

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329–353.

Raykov, T. (1998a). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375–385.

Raykov, T. (1998b). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22,* 369–374.

Raykov, T. (2001a). Bias of Cronbach's coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25,* 69–76.

Raykov, T. (2001b). Estimation of congeneric scale reliability using covariance structure models with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54,* 213–221.

Raykov, T. (2002a). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37,* 89–103.

Raykov, T. (2002b). Examining group differences in reliability of multiple-component measuring instruments. *British Journal of Mathematical and Statistical Psychology, 55,* 145–158.

Raykov, T. (2003). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35,* 299–331.

Raykov, T. (2004). Estimation of maximal reliability: A covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology, 57,* 21–27.

Raykov, T., & Grayson, D. A. (2003). A test for change of composite reliability in scale development. *Multivariate Behavioral Research, 38,* 143–159.

Raykov, T., Marcoulides, G. A., & Boyd, J. (2003). Using SEM programs to perform matrix manipulation and data simulation. *Structural Equation Modeling, 10,* 312–322.

Williams, R. H., & Zimmerman, D. T. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 21,* 59–69.
Zimmerman, D. W. (1972). Test reliability and the Kuder–Richarson formulas: Derivation from probability theory. *Educational and Psychological Measurement, 32,* 939–954.
Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395–412.
Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate or reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33–49.

APPENDIX
LISREL Input File for Point and Interval Estimation of Test Reliability

POINT AND INTERVAL ESTIMATION OF TEST RELIABILITY

DA NI = 4 NO = 400 ! Data file contains $k = 4$ variables for $N = 400$ participants (see main text)

RA = <data file name>! This is the name of the file where the raw data resides or filename of covariance matrix

LA ! The following line contains labels for the four observed variables

Y_1 Y_2 Y_3 Y_4 ! These are the actual labels for Y_1 through Y_4, respectively

MO NY = 4 NE = 5 TE = ZE ! There are four *Y*s and five ηs; last four ηs are dummy variables

LE ! Following are the labels for the latent variables

ETA THETA*11 THETA*22 THETA*33 THETA*44 ! Last four are the rescaled errors

FI PS 1 1 ! Fix underlying latent variable variance ($Var(\eta_1) = 1$)

VA 1 PS 1 1 ! to achieve model identification (as usually done)

FR LY 1 1 LY 2 1 LY 3 1 LY 4 1 ! Free indicator loadings (the *b*s in Equation 1), as usual

FR LY 1 2 LY 2 3 LY 3 4 LY 4 5 ! These are the paths from error terms into their *Y*s

CO LY(1,2) = LY(1,1)+LY(2,1)+LY(3,1)+LY(4,1) ! This is the rescaling constraint (Equation 9)

EQ LY 1 2 LY 2 3 LY 3 4 LY 4 5 ! All paths from *E*s to *Y*s are set equal to $C^{1/2}$ (see Equation 6)

ST .5 ALL ! Need start values in model with more ηs than *Y*s (may need others for other data)

OU ALL ND = 4 AD = OFF NS ! Need all output to get the estimates' covariance matrix

Note. The last four ηs (denoted earlier THETA*11 THETA*22 THETA*33 THETA*44) are dummy latent variables whose variances are set equal to the rescaled error variances for model Equation 1 (see Equations 8 and 9); that is, to θ_{11}^* through θ_{44}^*, respectively.