Appeared in *British Journal of Mathematical and Statistical Psychology*, 2008, *61*, 275-285

"Alpha if Item Deleted: A Note on Criterion Validity Loss in Scale Revision if Maximising Coefficient Alpha

Tenko Raykov

Michigan State University

Author Note:

I am grateful to B. Muthen for instructive comments on constraint evaluation, as well as to the Editor and two anonymous Referees for valuable criticism on an earlier draft of the paper, which contributed considerably to its improvement. Correspondence on this paper may be addressed to Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, East Lansing, MI 48824, USA; email: raykov@msu.edu .

RUNNING HEAD: "ALPHA IF ITEM DELETED" AND VALIDITY LOSS

Abstract

This note is concerned with a validity-related limitation of the widely available and used index "alpha if item deleted" in the process of construction and development of multiple-component measuring instruments. Attention is drawn to the fact that this statistic can suggest dispensing with such scale components, whose removal leads to loss in criterion validity while maximising the popular coefficient alpha. As an alternative, a latent variable modelling approach is discussed that can be used for point and interval estimation of composite criterion validity (as well as reliability) after deletion of single components. The method can also be utilised to test conventional or minimum level hypotheses about associated population change in measurement quality indices.

Keywords: *coefficient alpha*, *criterion validity, interval estimation, latent variable modelling, multiple-component measuring instrument, reliability*

"Alpha if Item Deleted": A Note on Loss of Criterion Validity in Scale Development If Maximising Coefficient Alpha

Multiple-component measuring instruments are highly popular in psychology and the behavioural sciences. Before being widely used they typically need to undergo a process of development through possibly repeated revisions that aim to ensure high psychometric quality of finally recommended scales, in particular high reliability and validity. A rather frequently used statistic for these purposes in empirical research is Cronbach's coefficient alpha (α; e.g., Cronbach, 1951), and especially the index "alpha if item deleted" that represents the increment or drop in the sample value of α if dispensing with a scale component. Recently, however, Raykov (2007a) showed that in certain circumstances that do not appear rare in behavioural research, this index can suggest the deletion of such instrument components whose removal leads to maximal increment in α but entails considerable loss in composite reliability. This results from the fact that α in general incorrectly evaluates scale reliability already at the population level (e.g., Novick & Lewis, 1967; Zimmerman, 1972), and points out the possibility that while seeking components to remove in order to maximise coefficient alpha, a psychologist can in fact seriously compromise reliability of an instrument being developed.

The present note deals with an additional aspect of this potentially serious limitation of the popular alpha coefficient, and in particular of the widely utilised statistic "alpha if item deleted". The remainder indicates that criterion validity can similarly decrease as a result of removing a component from a tentative scale while maximising coefficient alpha, even if data were available from an entire studied population. As an alternative to this statistic, therefore, an extension of the latent variable modelling procedure in Raykov (2007a) is recommended that yields point and interval estimation of criterion validity, in addition to that of reliability, after dispensing with single components. The method provides ranges of plausible population values for these measurement quality indices following any component's deletion, and can be used for testing conventional as well as minimum level hypotheses about them.

Loss in Criterion Validity When Deleting Components to Maximise Coefficient Alpha

This discussion is based on the assumption that a set of congeneric measures is given, denoted $X_1, X_2, \ldots, X_p$ ($p > 2$; Jöreskog, 1971), that is,

$$X_i = T_i + \varepsilon_i = \gamma_i + \beta_i \xi + \varepsilon_i \qquad (1)$$

holds ($i = 1, \ldots, p$), where $T_1, T_2, ..., T_p$ and $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ are respectively their true and error scores, and ξ designates the common latent dimension evaluated by the measures (e.g., $\xi = T_1$ can be taken; Lord & Novick, 1968). [1] For identifiability reasons, $Var(\xi) = 1$ is also set, where $Var(.)$ denotes variance in a studied population, and with respect to ε_i it is only required that their covariance matrix Ψ be positive definite (e.g., Zimmerman, 1975; $i = 1, \ldots, p$). Assume also that a criterion variable, C, is pre-specified. In the rest of this article, the reliability of the composite $Y = X_1 + X_2 + \ldots + X_p$ will be of interest as well as that of closely related versions of it, along with their criterion validity as reflected in the correlation coefficient $Corr(Y, C)$ (e.g., Crocker & Algina, 1986). [2]

In the process of instrument construction and development, behavioural scientists commonly follow the wide-spread practice of repeatedly examining the change in coefficient alpha after single component removal, typically referred to as "alpha if item deleted". This procedure is based on the sample value of the gain or loss in α occurring if say the jth component is dropped from a tentative scale:

$$\Delta\alpha_{Y,-j} = \alpha_Y - \alpha_{Y,-j} \ , \qquad (2)$$

where

$$\alpha_Y = \frac{p}{p-1}\left[\frac{\sum_{i \neq j} Cov(X_i, X_j)}{Var(Y)}\right] \qquad (3)$$

is coefficient alpha for the scale Y, and $\alpha_{Y,-j}$ denotes this coefficient for the composite $Y_{-j} = Y - X_j$, i.e., represents alpha if this item X_j is deleted ($j = 1, \ldots, p$). The estimates of $\alpha_{Y,-j}$ are furnished by widely circulated software, e.g., SPSS, SAS, STATISTICA, and are at present nearly routinely utilised in the behavioural and social sciences for purposes of instrument revision ($j = 1, \ldots, p$). Use of this procedure is based on the tacit, but in general incorrect (see above), assumption that $\alpha_{Y,-j}$ represents the change in reliability following deletion of the jth component ($j = 1, \ldots, p$). On this presumption, a widely adhered to practice in empirical research is to inspect the index "alpha if item deleted" for each component in a tentative scale, in an effort to identify a way of maximally enhancing reliability via single item deletion; then scholars commonly proceed with the scale version which results from dropping the component associated with the highest $\alpha_{Y,-j}$ such that $\alpha_{Y,-j} > \alpha_Y$ ($j = 1, \ldots, p$). As shown recently in Raykov (2007a), however, this procedure cannot be generally trusted because of two important reasons. On the one hand, it depends critically on the sample estimate of the gain or drop in coefficient alpha due to component removal, and in addition α in general incorrectly evaluates scale reliability already at the population level, as mentioned earlier. Consequently, a researcher relying on the index "alpha if item deleted" could decide to proceed with such a revision of a tentative composite, which is associated with maximal increase in α but in actual fact leads to lower reliability, even if data were available from an entire population.

This limitation of the popular statistic "alpha if item deleted" turns out to have further consequences to that just indicated. Specifically, dispensing with a component for which $\alpha_{Y,-j} > \alpha_Y$ can lead also to loss in criterion validity, a major aspect of what may well be considered the bottom line in behavioural measurement. To see this, from

Equation (1) follows

$$Y = (\sum_{i=1}^{k} \beta_i)\xi + \sum_{i=1}^{k} \varepsilon_i \ , \qquad (4)$$

and hence (e.g., Lord & Novick, 1968)

$$Corr(Y,C) = \frac{Cov[C,(\sum_{i=1}^{k} \beta_i)\xi + \sum_{i=1}^{k} \varepsilon_i]}{\sqrt{Var(C)Var[(\sum_{i=1}^{k} \beta_i)\xi + \sum_{i=1}^{k} \varepsilon_i]}} = \frac{\sigma_{\xi C}}{\sigma_C} \frac{\sum_{i=1}^{k} \beta_i}{\sqrt{(\sum_{i=1}^{k} \beta_i)^2 + \sum_{i=1}^{k} \theta_i}} = \frac{\sigma_{\xi C}}{\sigma_C} \sqrt{\rho_Y} = \omega_Y \ , \qquad (5)$$

say, where $\sigma_{\xi C}$, σ_C, and ρ_Y denote the latent-criterion covariance, criterion standard deviation and scale reliability, respectively, while $\theta_j = Var(\varepsilon_j)$ ($j = 1, \ldots, p$).

Now denote by $\omega_{Y,-j}$ and $\rho_{Y,-j}$ correspondingly the criterion validity and reliability of a tentative scale from which the *j*th component is dropped ($j = 1, \ldots, p$). As shown in Raykov (2007a), there exist theoretically and empirically relevant settings that do not appear rare in psychological research, where removal of a component associated with the maximal increase in coefficient alpha entails in fact a loss in reliability. [3] Let in such a setting the *k*th component possess accordingly the properties that (i) $\alpha_{Y,-k} > \alpha_Y$, (ii) $\alpha_{Y,-k}$ is highest for all *k* ($1 \le k \le p$), and (iii) $\rho_{Y,-k} < \rho_Y$. From the inequality in (iii) and Equation (5) (with its corresponding modification for the so-revised composite), it obviously follows

$$\omega_{Y,-j} = Corr(Y - X_k, C) = \frac{\sigma_{\xi C}}{\sigma_C} \sqrt{\rho_{Y,-k}} < \frac{\sigma_{\xi C}}{\sigma_C} \sqrt{\rho_Y} = Corr(Y,C) = \omega_Y \ , \qquad (6)$$

that is, after dispensing with its *k*th component the revised scale has actually lower criterion validity than its immediately preceding version from which it is obtained. Equation (6) lets one further observe that the amount by which criterion validity will be

compromised in this way, $\omega_Y - \omega_{Y,-j}$, depends on the associated loss in reliability index and the correlation between latent and criterion variables.

Therefore, at least in the circumstances outlined in Raykov (2007a; see Footnote 3), the resulting trimmed scale score, $Y - X_k$, will have lower criterion validity. Thus it is possible that a psychologist involved in instrument development who follows the wide-spread practice of removing a component from a tentative scale, which is associated with the highest increase in coefficient alpha, in actual fact arrives at a revised scale that has considerably inferior criterion validity (in addition to such reliability) relative to its version before dropping that component. When this practice is adhered to across several consecutive revisions, as is commonly the case in empirical research, due to accumulation of this negative effect it is obvious that the end version may have substantially lower criterion validity as well as reliability compared to an initial scale.

A Latent Variable Modelling Approach to Evaluation of Measurement Quality Following Single Component Deletion

In order to resolve these potentially serious deficiencies of the popular statistic "alpha if item deleted", the latent variable modelling approach in Raykov (2007a) can be extended to accomplish point and interval estimation of criterion validity after deletion of any component from a tentative scale. This procedure is not concerned with the statistic "alpha if item deleted" but is instead entirely based on the coefficient of criterion validity, as well as that of reliability, for the composite resulting from deleting the jth component of a given scale ($j = 1, .., p$). To this end, Equation (1) is first considered as defining a latent variable model (e.g., Muthén, 2002), and then $2p$ external parameters (new parameters, or 'auxiliary' parameters) are introduced (cf. Raykov, 2007b). The first p of them, denoted $\pi_1, \pi_2, \ldots, \pi_p$, are defined as the criterion validity coefficients of the version resulting after deleting the jth component (see Equation (5)):

$$\pi_j = \frac{\sigma_{\xi C}}{\sigma_C}\left[\frac{\sum_{\substack{i=1,\\ i\neq j}}^{k}\beta_i}{\sqrt{(\sum_{\substack{i=1\\ i\neq j}}^{k}\beta_i)^2 + \sum_{\substack{i=1\\ i\neq j}}^{k}\theta_i}}\right] = \omega_{Y,-j}\ , \qquad (7)$$

while the second set of p external parameters, $\pi_{p+1}, \pi_{p+2}, \ldots, \pi_{2p}$, are defined as the reliability coefficients of the corresponding scale versions

$$\pi_{p+j} = \frac{(\sum_{\substack{i=1\\ i\neq j}}^{p}\beta_i)^2}{(\sum_{\substack{i=1\\ i\neq j}}^{k}\beta_i)^2 + \sum_{\substack{i=1\\ i\neq j}}^{k}\theta_i} = \rho_{Y,-j}\,, \qquad (8)$$

($j = 1, \ldots, p$). (With error covariances, the denominators in the right-hand side of Equations (7) and (8) are extended by the sum of non-zero error covariance estimates; e.g., McDonald, 1999.) It is emphasised that (7) and (8) are not model parameters but are functions of the latter, and hence can be estimated once those are so. When the maximum likelihood (ML) method is used for model fitting purposes, due to the invariance property of ML, the right-hand sides of (7) and (8) written in terms of the participating parameter estimates represent correspondingly the ML estimates of criterion validity and scale reliability after removing the jth component ($j = 1, \ldots, p$). Therefore, the latter estimates share all desirable large-sample properties of ML estimates—consistency, unbiasedness, normality and efficiency (e.g., Rao, 1973).

The estimates (7) and (8) do not address the question of how close they are to the respective population criterion validity and reliability coefficients after dispensing with a given component from a tentative scale, which are the actual coefficients of interest. To this end, a standard error and confidence interval for these quantities is needed. Using the delta method (e.g., Rao, 1973), an approximate standard error and confidence interval was furnished in Raykov (2007a) for scale reliability following deletion of any

component, and the same method can be used here to also render an approximate standard error and confidence interval for the criterion validity of each scale version obtained in this way. Denote first the $(2p - 2) \times 1$ vector of parameters in model (1) after deleting the jth scale component by $\psi_{-j} = (\psi_1,\ldots,\psi_{j-1}, \psi_{j+1}, \ldots, \psi_p, \psi_{p+1}, \ldots, \psi_{p+j-1}, \psi_{p+j+1}, \ldots, \psi_{2p})'$, where priming stands for transposition and the notation $\psi_1 = \beta_1, \ldots, \psi_{p-1} = \beta_p$, $\psi_p = \theta_1, \ldots, \psi_{2p} = \theta_p$ is used for ease of reference (see Equation (1); $j = 1, \ldots, p$). The first-order Taylor expansion of the criterion validity coefficient (7) around the population parameter

$$\psi_{0,-j} = (\beta_{0,1},\ldots,\beta_{0,j-1},\beta_{0,j+1},\ldots,\beta_{0,p},\theta_{0,1},\ldots,\theta_{0,j-1},\theta_{0,j+1},\ldots,\theta_{0,p})'$$
$$= (\psi_{0,1},\ldots,\psi_{0,j-1},\psi_{0,j+1},\ldots,\psi_{0,p-1},\psi_{0,p},\ldots,\psi_{0,p+j-1},\psi_{0,p+j+1},\ldots,\psi_{0,2p})'$$

is

$$\hat{\omega}_{p,-j} \approx \hat{\omega}_{p,-j}(\psi_{0,-j}) + \sum_{\substack{i=1,\\ i\neq j,\\ i\neq p+j}}^{2p} \hat{D}_i(\hat{\psi}_{-j,i} - \psi_{0,-j,i}) \ , \qquad (9)$$

where '≈' denotes 'approximately equal' and $\hat{D}_i = \dfrac{\partial \hat{\omega}_{p,-j}}{\partial \psi_i}$ is the partial derivative of $\hat{\omega}_{p,-j}$ with respect to its ith argument, taken at the parameter estimate point ($i = 1, \ldots, j-1, j+1,\ldots, p+j-1, p+j+1, \ldots, 2p$; $j = 1, \ldots, p$). (The explicit expressions for these derivatives can be rendered following well-known rules for differentiation, but are actually not needed for the purposes of this note, as indicated below.) Hence, an approximate standard error for criterion validity following removal of the jth component is obtained from Equation (9) as:

$$\hat{S.E.}(\omega_{p,-j}) = \sqrt{\frac{\partial \hat{\omega}_{p,-j}}{\partial \psi_{-j}}' Cov(\hat{\psi}_{-j}) \frac{\partial \hat{\omega}_{p,-j}}{\partial \psi_{-j}}} \ , \qquad (10)$$

where $\frac{\partial \hat{\omega}_{p,-j}}{\partial \psi_{-j}'} = \left(\hat{D}_1, \hat{D}_2, ..., \hat{D}_{j-1}, \hat{D}_{j+1}, ..., \hat{D}_p, ..., \hat{D}_{p+j-1}, \hat{D}_{p+j+1}, ..., \hat{D}_{2p}\right)$ is the row vector of above mentioned derivatives and $Cov(\hat{\psi}_{-j})$ is the covariance matrix of pertinent parameter estimators, evaluated at the model solution (j = 1, …, p; cf. Raykov, 2007a). With this standard error, an approximate 100(1-δ)%-confidence interval ($0 < \delta < 1$) for criterion validity after dropping the jth component results as follows by capitalising on the asymptotic normality of the latent variable model parameter estimator (e.g., Muthén, 2002):

$$(\max(0, \hat{\omega}_{p,-j} - z_{1-\delta/2}\, S.\hat{E}.(\omega_{p,-j})),\ \min(1, \hat{\omega}_{p,-j} + z_{1-\delta/2}\, S.\hat{E}.(\omega_{p,-j}))), \qquad (11)$$

where $z_{1-\delta/2}$ denotes the $\delta/2$th quantile of the standard normal distribution while max(.,.) and min(.,.) stand for the larger and smaller numbers following in parentheses, respectively (j = 1, …, p).

The confidence interval (11) provides a range of plausible values, at a confidence level δ, for the population criterion validity of the composite of all components but the jth (j = 1, …, p). From the duality between hypothesis testing and confidence interval (e.g., Hays, 1994) it follows that (11) could also be used to test, as well known, conventional hypotheses at a significance level 1- δ about the criterion validity (or, for the same matter, reliability) coefficient after the jth component is dropped (j = 1, …, p). Moreover, (11) can be used to test minimum level hypotheses about this coefficient. Such a hypothesis states that after dispensing with a component from a given composite, the criterion validity (or reliability) is equal to at least w_0 say ($0 < w_0 < 1$), where w_0 is a substantively desirable threshold for criterion validity (or reliability) that a psychologist requires the composite to attain before being recommended for wider use. Accordingly, the pertinent null hypothesis is H_0: $\omega_j \geq w_0$ (or H_0: $\rho_j \geq w_0$), with corresponding alternative hypothesis H_1: $\omega_j < w_0$ (or H_1: $\rho_j < w_0$) (j = 1, …, p). This hypothesis is tested by examining whether the left-endpoint of the corresponding confidence interval is

entirely above the threshold, in which case the null hypothesis is considered retainable. Otherwise the alternative hypothesis is accepted and further instrument revision may be called for in order to accomplish the desired minimal level of validity (or reliability).

Empirical implementation

The described approach can be implemented in behavioural research with the increasingly popular latent variable modelling program M*plus* (Muthén & Muthén, 2006). This software incorporates recent advances in numerical optimization, which allow one to utilise readily the delta method application outlined in the preceding section for obtaining approximate standard errors and confidence intervals. Specifically, fitting model (1) with the added 2*p* external parameters in Equations (7) and (8), upon a request for confidence interval evaluation, yields point as well as interval estimates of these parameters, i.e., for the criterion validity (and reliability) coefficients after removing any component from a tentative scale. (The code accomplishing this goal, with annotations, is provided in Appendix 1 where it is applied with data used in the illustration section.) It is emphasised that this approach yields interval estimates of criterion validity (or reliability) following single component removal, as well as of an initially considered scale, whereas there is no counterpart interval estimate available when one adheres to the widely followed practice of using the index "alpha if item deleted" for scale revision purposes.

The discussed procedure is also directly applicable in settings with missing data that are frequently encountered in behavioural research dealing with scale construction and development. The method is then straightforwardly employed via use of full information maximum likelihood or multiple imputation if their assumptions are plausible—viz. data missing at random and normality (e.g., Muthén & Muthén, 2006; Little & Rubin, 2002). Last but not least, the proposed approach can be repeatedly used on scale versions resulting from preceding measure removal, in the search of yet further improvement in their criterion validity (and reliability) following deletion of any of their own components. Final recommendations regarding composite revision should be based,

however, on results from a replication study on an independent sample from the same population, due to the possibility of capitalization on chance.

Illustration on Data

In order to demonstrate the possibility that the widely used statistic "alpha if item deleted" can suggest misleading avenues of scale 'improvement' that are in fact associated with pronounced loss in criterion validity, simulated multinormal data are employed in this section. These data will also allow illustration of the utility of the discussed approach to evaluation of criterion validity (and reliability) after single component deletion. To this end, multivariate, zero-mean normal data were generated for $N = 500$ cases and $k = 5$ components Y_1 through Y_5 according to the model

$$\begin{aligned} X_1 &= \xi + \varepsilon_1 \\ X_2 &= \xi + \varepsilon_2 \\ X_3 &= \xi + \varepsilon_3 \\ X_4 &= \xi + \varepsilon_4 \\ X_5 &= 6\ \xi + \varepsilon_5\,, \end{aligned} \qquad (12)$$

where ξ was standard normal and the error terms ε_1 through ε_5 were independent zero-mean normal variables with variance 1.3 each; the criterion variable was generated as having correlation of .80 with ξ. The resulting covariance matrix is presented in Table 1.

Insert Table 1 about here

Conventional scale analysis on the initial composite containing all 5 components Y_1 through Y_5 (i.e., of the scale score $Y = Y_1 + \ldots + Y_5$) reveals an estimated alpha coefficient of .702. The widely used statistic "alpha if item deleted" indicates then that alpha will be maximised if the last component, Y_5, is removed from this composite. Specifically, according to that statistic, this removal would yield a four-component scale

with an alpha of .749 that is more than .1 higher than the alpha resulting from dropping instead any of the other four components (i.e., Y_1 to Y_4) from the initial composite. It is stressed that "alpha if item deleted" suggests here dropping the most reliable component of all five, in order to maximise alpha. Indeed, Equations (12) and immediately following discussion imply that reliability of each of the first four components is under .50, while that of Y_5 is in excess of .95.

To see the effect of deleting the last component on criterion validity and reliability, the latent variable modeling approach of this article is applied. First, fitting the congeneric model (1) (with the $2p$ external parameters, which do not affect model fit as they do not have any implications on the covariance structure), one obtains acceptable goodness of fit indices: chi-square = 13.173, degrees of freedom (df) = 9, p-value (p) = .155, root mean square error of approximation (RMSEA) = .030 with a 90%-confidence interval (0, .063). The criterion validity and reliability of the five versions of the initial scale, which result after each of its components is dropped in turn, as well as of that starting scale are presented in Table 2 along with corresponding standard errors and confidence intervals.

Insert Table 2 about here

As seen from Table 2, removal of Y_5—as suggested by the statistic "alpha if item deleted"—in fact leads to a substantial decrement in criterion validity from .773 (initial scale) to .687 (composite of first four components only), that is a drop by more than 10%. (Using the data generation parameters, this criterion validity loss is found to be equal to .096 in the population.) Similarly, reliability drops from .938 to .742, i.e., by more than 20%. (In the same way, this reliability decrement is found to be .217 in the population.) These effects represent pronounced losses in measurement quality, which result if one were to follow the wide-spread practice of deleting the single component whose removal maximises coefficient alpha. Note also that precision of estimation, as judged by the width of the associated confidence intervals, also drops if one were to dispense with the last component following that popular procedure. Further, from Table 2 it is seen that deletion of any of the first four components instead does not have a notable effect on the

point estimate of criterion validity or reliability, while leading to some relatively minimal loss of estimation precision. This demonstration exemplifies the point that adhering to the widely used statistic "alpha if item deleted" for purposes of scale revision can be associated with a marked loss in criterion validity and reliability, two measurement quality indices of special relevance for psychology and the behavioural sciences.

Conclusion

For a number of decades, a wide-spread practice has been followed by behavioural scientists involved in instrument development. Accordingly, the sample values of the popular coefficient alpha before and after single component removal have received critical attention in an effort to find ways of revising tentative scales so as to maximally enhance their reliability. In particular, the index "alpha if item deleted" has been rather frequently inspected for this purpose. The present note highlights a validity-related limitation of this statistic. The article shows that dispensing with a scale component to maximally increase coefficient alpha, can in fact entail considerable loss in criterion validity, a major aspect of behavioural measurement quality. In addition to a recent demonstration in Raykov (2007a) that such a revision path can lead to loss in composite reliability, this note further cautions psychologists engaged in instrument development that use of "alpha if item deleted" can be seriously misleading in more than one important way. As an alternative, the note discusses a latent variable modelling procedure that provides point and interval estimates of both criterion validity and reliability following deletion of each component in a tentative scale. In addition, the outlined approach allows simultaneous examination of the factorial structure of a given set of measures considered as its components. Moreover, the method is straightforwardly applicable in cases with missing data using maximum likelihood or multiple imputation, when their assumptions are plausible (viz. multi-normality and data missing at random), which is quite often the case in empirical contexts where instrument development is conducted. The discussed procedure, being based on latent variable modelling that is grounded in an asymptotic theory (e.g., Muthén, 2002), yields most trustworthy results with large samples, and similarly with (approximately) continuous components. Further,

being concerned with criterion validity, the proposed method may yield limited information about other relevant types of validity of measurement (e.g., Crocker & Algina,1986). In addition, its results depend on the choice of a criterion variable, which should be made in empirical research based on detailed knowledge of a substantive domain of concern. Finally, as presented in this note, the procedure utilises the assumption of congeneric measures, but it is stressed that it is readily extended to the case of more than a single underlying source of latent variability (see Footnote 1).

References

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

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

Hays, W. L. (1994). *Statistics.* Fort Worth, TX: Harcourt Brace Jovanovich.

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

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

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

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

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81-117.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

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. (2007a). Reliability if deleted, not "alpha if deleted": Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology, 60,* 201-216.

Raykov, T. (2007b). Estimation of revision effect on criterion validity of multiple-component measuring instruments. *Multivariate Behavioral Research, 42,* 415-434.

Zimmerman, D. W. (1972). Test reliability and the Kuder-Richardson formulas: Derivation from probability theory. *Educational and Psychological Measurement, 32,* 939-954.

Zimmerman, D. W. (1975). Probability measures, Hilbert spaces, and the axioms of classical test theory. *Psychometrika*, 30, 221-232.

Table 1

Covariance matrix of five congeneric measures (*N* = 500)

Variable	Y_1	Y_2	Y_3	Y_4	Y_5	C
Y_1	2.267					
Y_2	1.029	2.232				
Y_3	0.912	1.060	2.249			
Y_4	0.977	0.936	0.964	2.291		
Y_5	5.731	5.842	5.681	5.709	36.020	
C	0.793	0.784	0.818	0.718	4.857	1.056

Note. N = sample size, C = criterion variable.

Table 2

Point and interval estimates of criterion validity and reliability of composite resulting after indicated component is dropped from initial scale with all five components, and for that initial scale

DM	CV	SE	CI(CV)	R	SE	CI(R)
Y_1	.773	.017	(.739, .807)	.940	.008	(.924, .960)
Y_2	.773	.017	(.739, .807)	.939	.008	(.923, .959)
Y_3	.773	.017	(.739, .807)	.940	.008	(.924, .960)
Y_4	.774	.017	(.740, .808)	.940	.008	(.925, .961)
Y_5	.687	.020	(.648, .727)	.742	.018	(.707, .777)
None	.773	.017	(.739, .807)	.938	.007	(.925, .952)

Note. DM = dropped measure; CV = criterion validity, R = reliability; SE = standard error, CI(CV) and CI(R) = 95%-confidence interval of criterion validity and of reliability, respectively. Entries in row "None" pertain to estimates and standard errors for criterion validity and reliability of the initial scale with all five components (i.e., when none of the latter is removed).

Footnotes

[1] If $p = 2$, additional identifying restrictions will be needed, such as indicator loading equality (true score-equivalent measures) and/or error variance equality (e.g., parallel measures; Lord & Novick, 1968). Since the location parameters $\gamma_1, \gamma_2, ..., \gamma_k$ are not consequential for reliability in the setting underlying this paper, for convenience they are all assumed equal to zero (e.g., Bollen, 1989). The developments in this note can be directly generalised to the case where more than a single latent dimension is evaluated by a considered set of measures, following the corresponding approach in McDonald (1999; "omega" coefficient).

[2] The procedure discussed below is readily extended to the case when C is a latent variable with at least two indicators. As is common in latent variable modelling, C is also assumed unrelated to the error terms in the observed measures $X_1, \ldots, X_p$ (e.g., Bollen, 1989).

[3] As outlined in Raykov (2007a), at least the following general setup belongs to these empirical settings, with a single latent variable ξ and $p = q + 1$ indicators ($p > 2$; see Footnote 1):

$$X_i = \beta \xi + \varepsilon_i \ (i = 1, \ldots, q), X_{q+1} = \gamma \xi + \varepsilon_{q+1},$$

where $\gamma > \beta$ is sufficiently large, $Var(\varepsilon_{i+1}) \leq Var(\varepsilon_i) = \theta \ (i = 1, \ldots, q$; obviously, without limitation of generality one can also presume that $\beta > 0$; in the last $q+1$ equations consideration of component intercepts is dispensed with as they are inconsequential for reliability; see Footnote 1). As shown in the last cited source, deletion of the last component in this setup, which leads to the highest increment in coefficient alpha, entails substantial loss of reliability. (Note that this setup describes a case of $q+1$ congeneric measures, of which the first q are parallel while the last one is the most reliable of all; see also Appendix 2 in that source.)

Appendix 1

M*plus* Code for Evaluation of Criterion Validity and Reliability After Single Component Deletion

```
TITLE:      EVALUATION OF CRITERION VALIDITY/RELIABILITY AFTER COMPONENT DELETION
DATA:       FILE = <file name>      ! PROVIDES NAME OF RAW DATA FILE.
VARIABLE:   NAMES = Y1-Y6;          ! ATTACHES LABELS TO OBSERVED VARIABLES
MODEL:      KSI BY Y1* (P1)         ! THIS AND NEXT 4 LINES DEFINE THE COMPONENTS
            Y2* (P2)                ! AND ATTACH TO THEM PARAMETER SYMBOLS TO BE
            Y3* (P3)                ! USED BELOW (SEE MODEL CONSTRAINT SECTION).
            Y4* (P4)
            Y5* (P5);
            Y1* (P6);        ! THIS AND NEXT 4 LINES DEFINE THE ERROR VARIANCES
            Y2* (P7);        ! AND ATTACH TO THEM PARAMETER SYMBOLS TO BE USED
            Y3* (P8);        ! BELOW (SEE MODEL CONSTRAINT SECTION).
            Y4* (P9);
            Y5* (P10);
            C BY Y6*1; Y6@0; C WITH KSI* (P11); ! C IS THE CRITERION VARIABLE
            KSI@1; C@1; ! FIXES LATENT VARIANCE AT 1, FOR MODEL IDENTIFICATION
    MODEL CONSTRAINT:
            NEW(PI_1 PI_2 PI_3 PI_4 PI_5 PI_6 PI_7 PI_8 PI_9 PI_10 PI_11 PI_12);
            ! INTRODUCES THE AUXILIARY PARAMETERS π1, π2, …, π12 (SEE EQ. (7), (8))
            PI_6=(P2+P3+P4+P5)**2/
            ((P2+P3+P4+P5)**2+P7+P8+P9+P10); ! = RELIABILITY W/OUT 1ST COMPONENT
            PI_7=(P1+P3+P4+P5)**2/
            ((P1+P3+P4+P5)**2+P6+P8+P9+P10); ! = RELIABILITY W/OUT 2ND COMPONENT
            PI_8=(P1+P2+P4+P5)**2/
            ((P1+P2+P4+P5)**2+P6+P7+P9+P10); ! = RELIABILITY W/OUT 3RD COMPONENT
            PI_9=(P1+P2+P3+P5)**2/
            ((P1+P2+P3+P5)**2+P6+P7+P8+P10); ! = RELIABILITY W/OUT 4TH COMPONENT
            PI_10=(P1+P2+P3+P4)**2/
            ((P1+P2+P3+P4)**2+P6+P7+P8+P9);  ! = RELIABILITY W/OUT 5TH COMPONENT
                   PI_12=(P1+P2+P3+P4+P5)**2/
            ((P1+P2+P3+P4+P5)**2+P6+P7+P8+P9+P10); ! = RELIABILITY WITH ALL COMP.
            PI_1=P11*SQRT(PI_6); ! = CRITERION VALIDITY W/OUT 1ST COMPONENT.
            PI_2=P11*SQRT(PI_7); ! = CRITERION VALIDITY W/OUT 2ND COMPONENT.
            PI_3=P11*SQRT(PI_8); ! = CRITERION VALIDITY W/OUT 3RD COMPONENT.
```

```
            PI_4=P11*SQRT(PI_9); ! = CRITERION VALIDITY W/OUT 4TH COMPONENT.
            PI_5=P11*SQRT(PI_10); ! = CRITERION VALIDITY W/OUT 5TH COMPONENT.
            PI_11=P11*SQRT(PI_12); ! = CRITERION VALIDITY OF SCALE WITH ALL COMP.
OUTPUT:     CINTERVAL SAMPSTAT; ! REQUESTS INTERVAL ESTIMATES FOR ALL 12 COEFFICIENTS
```