Evaluation of Scale Reliability with Binary Measures Using Latent Variable Modeling

Tenko Raykov
Michigan State University

Dimiter M. Dimitrov
George Mason University

Tihomir Asparouhov
Muthén & Muthén

Author Note:

The contributions of the authors were equivalent, and their listing is reverse alphabetical. We are indebted to L. K. Muthén and B. O. Muthén for instructive comments on constraint evaluation, as well as to S. Penev for a valuable discussion on reliability estimation. We are grateful to the Editor and three anonymous Referees for a number of critical and suggesting comments on an earlier version that have contributed substantially to its improvement. Address correspondence on this manuscript to Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443A Erickson Hall, East Lansing, MI, 48824; e-mail: raykov@msu.edu .

Running head: SCALE RELIABILITY WITH BINARY MEASURES

Abstract

A method for interval estimation of scale reliability with discrete data is outlined. The approach is applicable with multi-item instruments consisting of binary measures, and is developed within the latent variable modeling methodology. The procedure is useful for evaluation of consistency of single measures and of sum scores from item sets following the two-parameter logistic model or the one-parameter logistic model. An extension of the method is described for constructing confidence intervals of change in reliability due to instrument revision. The proposed procedure is illustrated with an example.

Keywords: binary measure, coefficient alpha, factor analysis, latent variable modeling, multi-item measuring instrument, two-parameter logistic model, scale reliability.

Evaluation of Scale Reliability with Binary Measures Using Latent Variable Modeling

During the past century, precision of measurement has been one of the most researched topics in the social, behavioral, and educational sciences (e.g., Bollen, 1989; Crocker & Algina, 1986; Li, Rosenthal, & Rubin, 1996, and references therein). The majority of the work on it has been concerned with reliability of multiple-item measuring instruments consisting of continuous components. Empirical data in these and related disciplines, however, are frequently collected in discrete form. For instance, often used scales, tests, subscales, inventories, etc. consist of binary items or of components that are typically evaluated in a dichotomous format, e.g., true/false answer, present/absent symptom, endorsed/non-endorsed attitude. Treating in these cases the resulting data as (approximately) continuous, in particular when the goal is scale reliability evaluation, may yield misleading statistical and substantive conclusions.

As a popular index informing about reliability, coefficient alpha (α) has been widely used in behavioral and social research for more than 50 years (Cronbach, 1951). In spite of this popularity, a drawback of α was demonstrated in the 1960s (Novick & Lewis, 1967), which has important theoretical and empirical implications. Accordingly, unless the scale components are (essentially) tau-equivalent, α underestimates the composite reliability coefficient already at the population level (with unrelated errors). The amount of this slippage has been quantified subsequently and shown to be substantial under certain circumstances (Raykov, 1997). Moreover, with correlated errors α can overestimate scale reliability in the population, or conversely underestimate it, depending on parameter constellation (e.g., Zimmerman, 1972).

This and related research implies that coefficient alpha cannot be considered in general a dependable estimator of scale reliability with continuous or discrete components. Whereas alternatives to α have long been available for continuous measures (e.g., Bollen, 1989), the discrete case when approximate continuity cannot be reasonably assumed (and possibly handled with robust estimation methods) has received substantially less attention. Bartholomew & Schuessler (1991) and Bartholomew, Bassin, & Schuessler (1993) proposed an estimator of reliability for weighted scales of homogeneous sets of binary items with uncorrelated errors (see also Bartholomew & Knott, 1999). However, this estimator did not handle the commonly used unweighted sum score in the social and behavioral sciences, and in addition could not be considered readily applicable by the general researcher as it utilizes repeatedly numerical integration via subroutines currently not widely circulated in these disciplines. Recently, Dimitrov (2003) developed an estimator of dichotomous item and sum score reliability, capitalizing on a classical test theory approach. Yet his procedure did not provide interval estimates of item or composite reliability, in particular for the overall scale score, and used the assumption that the instrument under consideration included already calibrated items. The cited work by Bartholomew and colleagues and by Dimitrov did also not include a method for interval estimation of change in scale reliability following revision. As is well known, point estimates contain limited information about population parameters they purport to evaluate, and in particular cannot be used on their own to make statements as to how far they could be from those parameters that are of actual interest (e.g., Wilkinson & The Task Force on Statistical Inference, 1999). This serious theoretical and empirical limitation is counteracted by the provision of confidence intervals that represent ranges of plausible values for the population parameters and have begun to be increasingly used in empirical research over the last decade or so (e.g., Schmidt, 1996).

To respond to these limitations of past research, the present paper discusses an approach to interval estimation of reliability for homogeneous binary items and of their sum score. As illustrated in a later section, this procedure is preferable to an application of coefficient alpha in the setting of concern (cf. Novick & Lewis, 1967). Further, the article outlines a method for interval estimation of the loss or gain in composite reliability that results from deleting or adding one or more dichotomous measures to a unidimensional instrument (referred to as 'revision' in the sequel). The proposed procedure permits also simultaneous estimation of item parameters and does not assume the involved dichotomous items to have been previously calibrated.

Background, Notation, and Assumptions

In this paper, we assume that a given multi-component measuring instrument consists of p binary items, denoted $Y_1, Y_2, \ldots, Y_p$ ($p > 2$). Examples of such instruments are frequently used tests, scales, self-reports, inventories, subscales, testlets, or questionnaires (all referred to as "scale" or "instrument" in the remainder). These scale components are assumed to follow the two-parameter logistic (2PL) model that is widely utilized in the behavioral and social sciences (e.g., Lord & Novick, 1968). In addition, since the one-parameter logistic model is a special case of the 2PL model (e.g., see below), the following developments also cover the case when the items follow the former model. The measures $Y_1, Y_2, \ldots, Y_p$ may have been originally devised as dichotomous items, or alternatively resulted following specific scoring rules for polytomous items leading eventually to recording of true/false, present/absent, endorse/not-endorsed, or similar binary answers. (We will refer to the "true", "present", or "endorsed" response as "correct" answer in the remainder of this discussion.)

The rest of this article will be concerned with evaluation of (i) the reliability of each dichotomous item $Y_1, Y_2, \ldots, Y_p$; (ii) the reliability ρ_Y of their sum score

(1) $Y = Y_1 + Y_2 + \ldots + Y_p$;

and (iii) the change in the scale's reliability, $\Delta\rho_Y$, which results after removing or adding one or more items to the instrument. (When adding items, these are assumed to follow the 2PL model, or 1PL model if under consideration, along with the ones already in the scale.) To this end, we will capitalize on the approach in Dimitrov (2003), and will outline a method for obtaining interval estimates of (a) item reliability, (b) scale reliability, and (c) revision effect upon reliability.

The 2PL model assumption is tantamount to the following expression for the probability of correct response on the ith item from a considered set of dichotomous measures (e.g., Lord & Novick, 1968):

(2) $P(Y_i = 1|\theta) = exp(Da_i(\theta - b_i))/[1 + exp(Da_i(\theta - b_i))]$,

where P(.|.) denotes conditional probability, θ is the underlying trait being measured (e.g., attitude, ability, or in general a latent dimension), $exp(.)$ symbolizes exponent, a_i is the item's discrimination parameter, b_i its difficulty parameter, and $D = 1.702$ is a scaling constant to achieve close comparability of the item parameters to those of the two-parameter normal ogive (2PNO) model ($i = 1, \ldots, p$; in the latter, the relationship between trait and probability of correct response is described via the standard normal cumulative distribution function).

Model equivalence

For the aims of this paper, of basic importance will be the equivalence of the 2PNO model and the congeneric model for latent normal variables assumed to underlie the binary items $Y_1, \ldots, Y_p$ (Takane & de Leeuw, 1987; Jöreskog, 1971). Denoting by $Y_1^*, \ldots, Y_p^*$ these corresponding variables, the latter model assumes

(3) $Y_i^* = \lambda_i \eta + \zeta_i$

where η is a common factor with variance equal to 1, ζ_i are latent disturbances, and the probability of correct response on Y_i equals the area under the standard normal curve to the right of a pertinent threshold κ_i (i = 1, …, p). With this equivalence, estimates of the item discrimination and difficulty parameters for the 2PL model can be obtained by fitting the model in Equation (3) to data and rescaling counterpart estimates as follows (cf. Kamata & Bauer, 2008):

(4) $$\hat{a}_i = \hat{\lambda}_i / D \text{ , and } \hat{b}_i = \hat{\kappa}_i / \hat{\lambda}_i$$

(i = 1, …, p; see Muthén & Muthén, 2008). We note in passing that the one-parameter model, being a special case of the 2PL model and resulting when all item discrimination indices are the same, is obtained from the model in Equation (3) by imposing the equality restriction on all factor loadings, viz. $\lambda_1 = \lambda_2 = \ldots = \lambda_p$ (i.e., $a_1 = a_2 = \ldots = a_p$).

Point estimation of item and scale reliability

According to the classical test theory (CTT; e.g., Zimmerman, 1975), for the ith item's observed score the decomposition $Y_i = \tau_i + e_i$ holds, where τ_i is its true score and e_i its error score (i = 1, …, p). As is well known, the expected true score π_i of the item in the studied subject population can be obtained as

(5) $$\pi_i = \int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)d\theta \text{ ,}$$

with $\varphi(\theta)$ being the latent trait distribution and $P_i(\theta)$ symbolizing the probability of correct response (i = 1, …, p; e.g., Lord & Novick, 1968). Using this framework, with the assumptions of uncorrelated errors and normal trait distribution that are also adopted in this paper, Dimitrov (2003) provided the following analytic expressions for several population parameters associated with individual items, which will be capitalized on in the rest of this article. Specifically, for the item's mean true score,

(6) $$\pi_i = \frac{1 - erf(X_i)}{2}$$

was shown (i = 1, …, p), where $X_i = a_i b_i / \sqrt{2(1+a_i^2)}$ and $erf(.)$ is a known mathematical function called the "error function" that is numerically obtained (with an absolute error smaller than 0.0005) as

(7) $$erf(X_i) = 1-(1+m_1 X_i + m_2 X_i^2 + m_3 X_i^3 + m_4 X_i^4)^{-4} ,$$

with m_1 = .278393, m_2 = .230389, m_3 = .000972 and m_4 = .078108, assuming $X > 0$ (when $X < 0$, the property $erf(-X) = -erf(X)$ is used). Similarly, for the item error variance, denoted $\sigma^2(e_i)$, he showed

(8) $$\sigma^2(e_i) = m_i \exp[-.5(b_i / d_i)^2] ,$$

with

(9) $$m_i = 0.2646 - 0.118a_i + 0.0187a_i^2 \text{ and } d_i = 0.7427 + 0.7081/a_i + 0.0074/a_i^2 .$$

Further, for the item true variance, symbolized by $\sigma^2(\tau_i)$,

(10) $$\sigma^2(\tau_i) = \pi_i(1-\pi_i) - \sigma^2(e_i)$$

was deduced. This entailed for the item reliability, ρ_i,

(11) $$\rho_i = \frac{\sigma^2(\tau_i)}{\sigma^2(\tau_i)+\sigma^2(e_i)}$$

(i = 1, …, p). Finally, the scale reliability coefficient was rendered as

(12) $$\rho_Y = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)}}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)} + \sum_{i=1}^{p}\sigma^2(e_i)} .$$

The preceding discussion in this section evolved at the population level and no sampling or estimation was involved. From Equations (11) and (12), estimators of item reliability and of scale reliability can be furnished by substituting estimators of item difficulty and discrimination parameters within the expressions appearing in the right-hand sides of these equations. In this way, one obtains

$$(13) \qquad \hat{\rho}_i = \frac{\hat{\sigma}^2(\tau_i)}{\hat{\sigma}^2(\tau_i) + \hat{\sigma}^2(e_i)}$$

and

$$(14) \qquad \hat{\rho}_Y = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)}}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)} + \sum_{i=1}^{p}\hat{\sigma}^2(e_i)} \, ,$$

where a caret denotes estimator of the quantity underneath that results thereby ($i = 1, \ldots, p$). Hence, when the maximum likelihood (ML) method is used for parameter estimation and model testing purposes, due to its invariance property Equations (13) and (14) yield ML estimators of item reliability and scale reliability, respectively. The estimators furnished by Equations (13) and (14) possess therefore all desirable large-sample properties of ML estimators, viz. consistency, unbiasedness, normality, and efficiency (e.g., Roussas, 1997).

Point estimation of loss or gain in reliability following scale revision

When developing a multiple-component measuring instrument, behavioral, social and educational scholars oftentimes undertake revisions consisting of deleting items from a tentative version of it, or alternatively adding items that are congeneric with the ones already in a scale under consideration. Without loss of generality, assuming that say the last k items are considered for deletion ($1 \le k < p$), the change in reliability that would be incurred then, $\Delta\rho_Y$, can now be evaluated using Equation (12) as

(15) $$\Delta\rho_Y = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)}}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)}+\sum_{i=1}^{p}\sigma^2(e_i)} - \frac{\sum_{i=1}^{p-k}\sum_{j=1}^{p-k}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)}}{\sum_{i=1}^{p-k}\sum_{j=1}^{p-k}\sqrt{\sigma^2(\tau_i)\sigma^2(\tau_j)}+\sum_{i=1}^{p-k}\sigma^2(e_i)}.$$

In the right-hand side of Equation (15), the second term is the reliability of the revised scale, which is obtained with the same method as that of the initial scale. We would like to note in passing that the sign of reliability change, $\Delta\rho_Y$, can be negative or positive (or alternatively $\Delta\rho_Y = 0$ could hold), depending on the deleted k items and their psychometric properties. In particular, it is possible to enhance scale reliability when deleting one (or more) inappropriate items, which for instance could have disproportionally large error variances relative to the strength of their relationships with the underlying construct being measured. [1]

In empirical research, this reliability change due to instrument revision is estimated by substituting item parameter estimators into the item error variance and true variance expressions (see Equations (8) and (10)), and the resulting into Equation (15), leading to the following estimator of the revision effect:

(16) $$\hat{\Delta}\rho_Y = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)}}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)}+\sum_{i=1}^{p}\hat{\sigma}^2(e_i)} - \frac{\sum_{i=1}^{p-k}\sum_{j=1}^{p-k}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)}}{\sum_{i=1}^{p-k}\sum_{j=1}^{p-k}\sqrt{\hat{\sigma}^2(\tau_i)\hat{\sigma}^2(\tau_j)}+\sum_{i=1}^{p-k}\hat{\sigma}^2(e_i)}.$$

When ML is employed for estimation purposes, Equation (16) represents an ML estimator of the revision effect upon scale reliability, which thus possesses all earlier mentioned asymptotic properties, such as consistency, unbiasedness, normality and efficiency.

Relationship to coefficient alpha

Equation (12) expresses the population scale reliability coefficient with binary items. An index of reliability, which is widely used in empirical settings in such cases, is coefficient alpha. Despite its high popularity, as indicated earlier α underestimates scale reliability even if an entire population of interest were observed, unless the items are (essentially) tau-equivalent, i.e., evaluate the same underlying construct in the same units of measurement (with uncorrelated errors, Novick & Lewis, 1967; see previous discussion for correlated errors). The tau-equivalence condition is, however, a rather restrictive constraint that in general cannot be assumed fulfilled in social and behavioral research. Hence, α cannot be considered generally a dependable estimator of scale reliability also in case of binary items. In a later section, we demonstrate this underestimation property of α on dichotomous measure data.

Instead of using α for estimation of scale reliability with binary components, this paper advocates utilization of the estimator (14) for the composite reliability coefficient itself. This coefficient is logically the quantity of actual interest when questions about instrument reliability are being raised. Next we discuss a widely applicable procedure that furnishes confidence intervals for the reliability coefficients of interest in this paper (in addition to yielding point estimates of them). The resulting ranges of plausible population values for these coefficients are valid with large samples when the 2PL model is tenable.

Interval Estimation of Item and Scale Reliability and Change in Reliability Due to Revision

The above mentioned property of asymptotic normality of the estimators in Equations (13), (14) and (16) can be used to obtain interval estimates of item and scale reliability as well as of the gain or drop in the latter following revision. These estimates provide further important information about the psychometric qualities of a multi-component instrument under consideration, which is not contained in point estimates or results from hypothesis testing. To obtain the interval estimates, one can employ the well-known delta method (e.g., Raykov & Marcoulides, 2004). In order to outline this procedure next, for simplicity denote generically by ρ the item reliability ρ_i, scale reliability ρ_Y or the change $\Delta\rho_Y$ in the latter following revision, whichever is of concern in a particular analytic session (i = 1, …, p). The first-order Taylor approximation of this coefficient about the vector $\underline{\gamma}_0$ of population parameters is

(17) $$\rho(\underline{\gamma}) \approx \rho(\underline{\gamma}_0) + \left[\frac{\partial \rho(\underline{\gamma})}{\partial \underline{\gamma}'} \Big|_{\underline{\gamma}=\underline{\gamma}_0} \right] (\underline{\gamma} - \underline{\gamma}_0) \ ,$$

where the symbol '≈' denotes 'approximately equal', $\underline{\gamma}$ is the parameter vector of the 2PL model, and bracketed is the vector of partial derivatives of ρ with respect to the parameters it depends on (see Equations (13), (14) or (16), respectively). More specifically, when interval estimating a given item's reliability, the latter vector consists of 2 components—the discrimination and difficulty parameters of that item. When interval estimating the scale reliability coefficient, this vector $\underline{\gamma}$ consists of $2p$ parameters—these are all items' discrimination and difficulty parameters. And when interval estimating the change in reliability following revision, the vector $\underline{\gamma}$ consists of $4p$-$2k$ parameters—the parameters of the items in the initial and revised scales (see Equation (15)). From (17), a squared large-sample standard error of item reliability, scale reliability or change in reliability due to revision follows straightforwardly as

$$(18) \qquad Var(\hat{\rho}) \approx \left[\frac{\partial \hat{\rho}(\hat{\underline{\gamma}})}{\partial \hat{\underline{\gamma}}'} \Big|_{\underline{\gamma}=\hat{\underline{\gamma}}} \right] Cov(\hat{\underline{\gamma}}) \left[\frac{\partial \hat{\rho}(\hat{\underline{\gamma}})}{\partial \hat{\underline{\gamma}}} \Big|_{\underline{\gamma}=\hat{\underline{\gamma}}} \right],$$

where $Cov(\hat{\underline{\gamma}})$ is the observed inverted information matrix (or part of it pertaining to $\underline{\gamma}$). Based on (18), a large-sample 100(1 - δ)%-confidence interval (0 < δ < 1) for the item reliability, scale reliability or revision effect upon reliability, is readily furnished as

$$(18) \qquad \hat{\rho} \pm z_{\delta/2}\sqrt{Var(\hat{\rho})} \;,$$

where $z_{\delta/2}$ is the (1-δ/2)th quantile of the standard normal distribution.

Procedure application in empirical research

Evaluation of the reliability-related point and interval estimates in Equations (13), (14), and (16) through (18) in behavioral and social research can be carried out using the increasingly popular latent variable modeling M*plus* (for software related details, see Appendixes 1 and 2; Muthén & Muthén, 2008). In particular, interval estimation of these three parameters does not involve then explicit estimation by the researcher of partial derivative values and respective information matrix parts. The reason is that this evaluation can be carried out as a byproduct of fitting the 2PL model. To this end, in a special model constraint section one introduces as "new parameters" the item, scale, and change in reliability coefficients; these functions of item parameters are defined as identical to the right-hand sides of the respective Equations (13), (14) and (16). Approximate standard errors and confidence intervals, both at 95%- and 99%-confidence levels, for each of these "new parameters" using the delta method are then provided by the software (see (17) and (18)). The inclusion of the 'new parameters' into the 2PL model does not affect its fit to data, since this extension does not have any consequences for the items Y_1, Y_2, …, Y_p or their distribution. The outlined procedure is demonstrated next.

Illustration on Data

To illustrate the discussed reliability evaluation method, we use simulated data on n = 1000 subjects for p = 5 binary items complying with the 2PL model and possessing item discrimination and difficulty parameters correspondingly as follows: $a_1 = 1.8$, $b_1 = .2$; $a_2 = .5$, $b_2 = .75$; $a_3 = 1.25$, $b_3 = -1$; $a_4 = 1$, $b_4 = 1.5$; $a_5 = .2$, $b_5 = -1.5$. (To this end, first the probabilities for correct response were worked out using Equation (2), for each of these 5 item parameter combinations and n corresponding standard normal draws for θ. Then pertinent binary data were generated on the five items using the resulting as respective probabilities for 'success'.) With these parameters, the true item reliability coefficients are obtained via Equation (11) as $\rho_1 = .547$, $\rho_2 = .148$, $\rho_3 = .356$, $\rho_4 = .213$, and $\rho_5 = .034$ (see also Equations (6) through (10)). Similarly, from Equation (12) the true reliability of their sum score, $Y = Y_1 + Y_2 + \ldots + Y_5$, results as $\rho_Y = .597$.

Fitting to this data set the 2PL model yields a Pearson chi-square value (χ^2) of 13.374, for degrees of freedom (df) = 21 and associated p-value (p) of .90, as well as a likelihood ratio $\chi^2 = 15.029$, $df = 21$, $p = .82$. (See Appendix 1 for software source code with annotating comments.) These goodness-of-fit indices suggest a tenable model. The obtained estimates, standard errors and approximate 95%-confidence intervals for the item parameters, reliability, scale reliability, and related parameters are presented in Table 1.

Insert Table 1 about here

As seen from Table 1, the composite reliability estimate of $\hat{\rho}_Y = .621$, with a standard error (SE) of .017, is fairly close to its population (true) value $\rho_Y = .597$. Moreover, this population value ρ_Y is covered by the resulting 95%-confidence interval (.569, .673) for scale reliability (see final row of Table 1). In contrast, coefficient alpha is estimated as $\hat{\alpha}_Y = .525$, and thus is (i) markedly lower than the true scale reliability coefficient of .597, as well as (ii) positioned to the left of the above reliability confidence interval. Even more importantly, via the definitional formula for coefficient alpha (e.g., Crocker & Algina, 1986), its population value is readily obtained as α = .566, and hence is notably lower than the population reliability ρ_Y = .597. Furthermore, the population α = .566 is located to the left of the scale reliability's 95%-confidence interval, (.569, .673). That is, the population alpha coefficient is not even a plausible value (at the 95% confidence level) for the population scale reliability coefficient.

These results about alpha's performance are not unexpected and are consistent with the earlier mentioned population underestimation feature of α with respect to composite reliability (Novick & Lewis, 1967). In fact, these findings represent a specific illustration of this drawback of coefficient alpha with binary items. (Alpha's underestimation feature with scales consisting of continuous components has been well documented in earlier research; e.g., Li et al., 1996, and references therein.)

Table 1 also shows that while the first four item reliability estimates are associated with confidence intervals substantially above the zero point, that of the fifth item is not so. For the sake of illustrating the outlined interval estimation method for reliability change due to revision, we next evaluate the drop or gain in reliability resulting from say deleting the last item of the initial scale. In particular, for the purposes of this section we are also interested in examining if this change in measurement consistency is significant in the population. (In social and behavioral research, such a decision needs to be based in addition on substantive and validity related considerations.)

To accomplish this goal, we utilize the confidence interval resulting from an application of the delta method on the pertinent difference in scale reliability coefficients, as stated in Equation (15). To this end, all we need to do is include in the model fitting process Equation (15). (See Appendix 2 for source code with annotating comments.) This inclusion does not affect as mentioned the model fit or any estimate, standard error or confidence interval reported in Table 1, while yielding a notably higher estimate of the reliability ρ' for the revised scale consisting of the first four items: $\hat{\rho}' = .681$, SE = .022, and 95%-confidence interval: (.637, .725). In addition, we obtain the estimate of revision effect upon reliability as $\hat{\Delta}\rho = -.06$, SE = .017, with a 95%-confidence interval (-.093, -.027). Since the zero point is not covered by the latter interval and is to the right of it, we conclude that dropping the last from the original set of 5 binary items is associated with a significant gain in reliability (at the .05 level). Any other point hypothesis about reliability change due to revision—as well as some one-tailed hypotheses—could be tested in the same manner (at the significance level of .05), viz. by examining whether the hypothetical value is covered by the 95%-confidence interval (e.g., Hays, 1994; a correspondingly modified confidence level needs to be used if another significance level is pre-selected.) This example also provides an illustration of the fact that adding binary items to an existing scale of dichotomous measures can lower the composite reliability (consider 'reversely' the initial 5-item scale as resulting from the later considered 4-item scale version, when extending the latter by adding the item Y_5).

Conclusion

This article was concerned with item and scale reliability for multiple-component instruments consisting of binary measures. Dichotomous items and composites based on them are quite frequently utilized for evaluation of indirectly observable latent dimensions (traits, abilities, attitudes, aptitude) in the behavioral, social, and educational sciences. The paper outlined a method for interval estimation of item and scale reliability, which permits researchers to obtain ranges of plausible values in studied populations for the degree of consistency associated with binary components and their sum score. The approach allowed one also to evaluate the gain or loss in scale reliability following a decision to add or delete certain items from a tentative composite. Interval estimates of item and scale reliability as well as change in it following revision, provide important information not contained in their point estimates. This information can be especially useful in instrument construction and development frequently carried out by social and behavioral educational researchers (e.g., Wilkinson, 1999; Schmidt, 1996). With its focus on interval estimation, the proposed procedure further permits testing simple and especially composite hypotheses if of interest with regard to any of the three reliability coefficients or change quantity of concern. In particular, minimum effect hypotheses (e.g., Rindskopf, 1997) about item reliability, scale reliability, or revision effect upon reliability can be readily tested by examining whether the pertinent confidence interval is entirely within the null hypothesis or the alternative hypothesis tail (e.g., Roussas, 1997).

The described method is best utilized with large samples of subjects when a considered scale complies with the popular two-parameter logistic model (or, as a special case, the one-parameter logistic model) and is associated with uncorrelated errors. The approach includes also a routine test of overall fit of this model, which is conducted by examining the fit of the counterpart congeneric model (e.g., Tanaka & de Leeuw, 1987; see also Kamata & Bauer, 2008). With this feature, the procedure permits one to routinely assess the latent structure underlying a given scale with binary components, and based on examining their factor loadings and standard errors possibly consider revisions aimed at enhancing its psychometric qualities. The method is also straightforwardly employed in cases with missing values that are rather frequent in empirical research, under the assumption of data missing at random (e.g., Little & Rubin, 2002). While the concern of the present paper was primarily with interval estimation of scale reliability for binary items and of the change in it following revision, future research needs to examine the effects of deleting items with various psychometric features, the relationships among the remaining items as well as the number of items deleted (added). This will permit obtaining a more complete picture of revision effect upon reliability of scales consisting of dichotomous measures in the practice of social and behavioral research.

References

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold.

Bartholomew, D. J., Schuessler, K. F. (1991). Reliability of attitude scores based on a latent trait model. In P. V. Marsden (Ed.), *Sociological Methodology,* vol. 21 (pp. 97-123). San Francisco, CA: Jossey-Bass.

Bartholomew, D. J., Bassin, E. L., & Schuessler, K. F. (1993). Properties of a latent trait reliability coefficient. *Sociological Methods and Research, 22,* 163-192.

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.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27,* 440-458.

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.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136-153.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *5*, 98-107.

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

Muthén, L. K., & Muthén, B. O. (2008). *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.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of tau-equivalence for fixed congeneric components. *Multivariate Behavioral Research, 32,* 329-354.

Raykov, T., Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parametric functions in covariance structure models. *Structural Equation Modeling, 11,* 659-675.

Rindskopf, D. (1997). Testing 'small', not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319-334). Mahwah, NJ: Erlbaum.

Roussas, G. G. (1997). *A course in mathematical statistics*. Upper Saddle River, NJ: Prentice Hall.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

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

Parameter estimates, standard errors, and 95%-confidence intervals for the fitted two-parameter logistic model

Parameter	Est.	SE	t-value	95%-CI	
a_1	1.546	0.269	5.738	(1.016,	2.071)
a_2	0.548	0.066	8.273	(.417,	.676)
a_3	1.962	0.494	3.970	(.992,	2.928)
a_4	1.112	0.161	6.890	(.795,	1.427)
a_5	0.164	0.048	3.389	(.069,	.258)
b_1	0.140	0.048	2.900	(.045,	.234)
b_2	0.725	0.105	6.883	(.518,	.931)
b_3	-0.914	0.072	-12.759	(-1.054,	-.773)
b_4	1.471	0.119	12.231	(1.237,	1.705)
b_5	-1.683	0.534	-3.150	(-2.730,	-.636)
$\sigma^2(\tau_1)$	0.122	0.016	7.530	(.090,	.153)
$\sigma^2(\tau_2)$	0.038	0.006	6.534	(.027,	.050)
$\sigma^2(\tau_3)$	0.090	0.018	4.948	(.054,	.126)
$\sigma^2(\tau_4)$	0.029	0.008	3.582	(.013,	.045)
$\sigma^2(\tau_5)$	0.005	0.005	.976	(-.005,	.014)
$\sigma^2(e_1)$	0.126	0.016	7.815	(.095,	.158)
$\sigma^2(e_2)$	0.193	0.006	30.084	(.181,	.206)
$\sigma^2(e_3)$	0.075	0.017	4.313	(.041,	.109)
$\sigma^2(e_4)$	0.089	0.009	10.418	(.072,	.106)
$\sigma^2(e_5)$	0.234	0.006	40.344	(.223,	.245)
ρ_1	0.491	0.065	7.543	(.363,	.619)
ρ_2	0.165	0.025	6.643	(.116,	.213)
ρ_3	0.546	0.105	5.193	(.340,	.752)
ρ_4	0.248	0.063	3.907	(.123,	.372)
ρ_5	0.019	0.020	.976	(-.020,	.058)
ρ_Y	0.621	0.017	23.397	(.569,	.673)

Note. Est. = estimate, SE = standard error, *t*-value = ratio of estimate to standard error (Muthén & Muthén, 2008), 95%-CI = 95%-confidence interval. (Parameter notation defined in main text.)

Footnote

[1] In general, as discussed in the literature (e.g., Lord & Novick, 1968), it is not true that increase in the number of binary items, p, leads to an increase in scale reliability, ρ_Y. (Such an increase would be the case, though, with parallel items, as seen from the well-known Spearman-Brown formula that will be valid then; e.g., Crocker & Algina, 1986. To come up with parallel items, however, especially in large numbers, is exceedingly difficult in empirical social and behavioral research.) This can be seen from the preceding discussion in the main text, and particularly from a comparison of both terms in the right-hand side of Equation (15). Specifically, adding items with weak relationship to the underlying common true score in the currently considered homogeneous measure case (e.g., Equations (4) and (5)), while associated with sufficiently large error variances, can lead to the second ratio in (15) being larger than the first; this will yield a negative sign of the change in scale reliability, i.e., an extended scale version with larger number of items yet lower reliability. An example is provided in the illustration section.

Appendix 1

M*plus* Source Code for Point and Interval Estimation of Item and Scale Reliability with Binary Items

```
TITLE:      POINT AND INTERVAL ESTIMATION OF ITEM AND SCALE RELIABILTY FOR
            DICHOTOMOUS MEASURES.  (ANOTATING COMMENTS ARE ADDED AFTER
            EXCLAMATION MARK WITHIN PERTINENT ROW.)
DATA:       FILE = <name of raw data file>; ! NEED TO ANALYZE THE RAW DATA
VARIABLE:   NAMES = Y1-Y5;
            CATEGORICAL = Y1-Y5; ! STATES CATEGORICAL NATURE OF SCALE COMPONENTS
ANALYSIS:   ESTIMATOR = ML;
MODEL:      F BY Y1* (P1)  ! THIS AND NEXT LINE ASSIGN PARAMETRIC SYMBOLS (P1 TO
            Y2-Y5 (P2-P5); ! P5) TO SUCCESSIVE FACTOR LOADINGS, TO BE USED BELOW
            F@1;            ! FIXES LATENT VARIANCE, FOR MODEL IDENTIFICATION
            [Y1$1-Y5$1] (P6-P10); ! ASSIGNS SYMBOLS TO SUCCESSIVE THRESHOLDS
 MODEL CONSTRAINT:
 NEW(A1 A2 A3 A4 A5 B1 B2 B3 B4 B5 X1 X2 X3 X4 X5 PI_1 PI_2 PI_3 PI_4 PI_5
 TV1 TV2 TV3 TV4 TV5 EV1 EV2 EV3 EV4 EV5 R1 R2 R3 R4 R5 TRUVAR ERRVAR REL);
 A1 = P1/1.702; ! THIS AND NEXT 4 LINES YIELD THE ITEM DISCRIMINATION INDICES
 A2 = P2/1.702; ! (SEE EQUATION (4))
 A3 = P3/1.702;
 A4 = P4/1.702;
 A5 = P5/1.702;
 B1 = P6/P1; ! THIS AND NEXT 4 LINES FURNISH THE ITEM DIFFICULTY INDICES
 B2 = P7/P2; ! (SEE EQUATION (4))
 B3 = P8/P3;
 B4 = P9/P4;
 B5 = P10/P5;
 X1 = A1*B1/SQRT(2+2*A1**2); ! THIS AND NEXT 4 LINES DEFINE AUXILIARY
 X2 = A2*B2/SQRT(2+2*A2**2); ! QUANTITIES THAT SIMPLIFY CODING NEXT (SEE (10))
 X3 = -A3*B3/SQRT(2+2*A3**2); ! (SEE NOTE 2 BELOW)
 X4 = A4*B4/SQRT(2+2*A4**2);
 X5 = -A5*B5/SQRT(2+2*A5**2); ! (SEE NOTE 2 BELOW)
 PI_1 = .5 -.5*(1-1/(1+.278393*X1 + .230389*X1**2
            + .000972*X1**3 + .078108*X1**4)**4); ! = 1ST ITEM MEAN TRUE SCORE
 PI_2 = .5-.5*(1-1/(1+.278393*X2 + .230389*X2**2
            + .000972*X2**3 + .078108*X2**4)**4); ! = 2ND ITEM MEAN TRUE SCORE
 PI_3 = .5+.5*(1-1/(1+.278393*X3 + .230389*X3**2
            + .000972*X3**3 + .078108*X3**4)**4); ! = 3RD ITEM MEAN TRUE SCORE
 PI_4 = .5-.5*(1-1/(1+.278393*X4 + .230389*X4**2
            + .000972*X4**3 + .078108*X4**4)**4); ! = 4TH ITEM MEAN TRUE SCORE
 PI_5 = .5+.5*(1-1/(1+.278393*X5 + .230389*X5**2
            + .000972*X5**3 + .078108*X5**4)**4); ! = 5TH ITEM MEAN TRUE SCORE
 EV1 = (.2646 -.118*A1 + .0187*A1**2)*EXP(-.5*(B1/
             (.7427 + .7081/A1 + .0074/A1**2))**2); ! = ERROR VARIANCE OF ITEM 1
 EV2 = (.2646 -.118*A2 + .0187*A2**2)*EXP(-.5*(B2/
             (.7427 + .7081/A2 + .0074/A2**2))**2); ! = ERROR VARIANCE OF ITEM 2
 EV3 = (.2646 -.118*A3 + .0187*A3**2)*EXP(-.5*(B3/
             (.7427 + .7081/A3 + .0074/A3**2))**2); ! = ERROR VARIANCE OF ITEM 3
 EV4 = (.2646 -.118*A4 + .0187*A4**2)*EXP(-.5*(B4/
             (.7427 + .7081/A4 + .0074/A4**2))**2);! = ERROR VARIANCE OF ITEM 4
 EV5 = (.2646 -.118*A5 + .0187*A5**2)*EXP(-.5*(B5/
            (.7427 + .7081/A5 + .0074/A5**2))**2); ! = ERROR VARIANCE OF ITEM 5
 TV1 = PI_1*(1-PI_1)-EV1; ! THIS IS THE TRUE VARIANCE OF ITEM 1
 TV2 = PI_2*(1-PI_2)-EV2; ! THIS IS THE TRUE VARIANCE OF ITEM 2
 TV3 = PI_3*(1-PI_3)-EV3; ! THIS IS THE TRUE VARIANCE OF ITEM 3
 TV4 = PI_4*(1-PI_4)-EV4; ! THIS IS THE TRUE VARIANCE OF ITEM 4
 TV5 = PI_5*(1-PI_5)-EV5; ! THIS IS THE TRUE VARIANCE OF ITEM 5
 R1 = TV1/(TV1+EV1); ! THIS IS THE RELIABILITY COEFFICIENT OF ITEM 1
 R2 = TV2/(TV2+EV2); ! THIS IS THE RELIABILITY COEFFICIENT OF ITEM 2
 R3 = TV3/(TV3+EV3); ! THIS IS THE RELIABILITY COEFFICIENT OF ITEM 3
 R4 = TV4/(TV4+EV4); ! THIS IS THE RELIABILITY COEFFICIENT OF ITEM 4
 R5 = TV5/(TV5+EV5); ! THIS IS THE RELIABILITY COEFFICIENT OF ITEM 5
```

```
 ERRVAR = EV1 + EV2 + EV3 + EV4 + EV5; ! = ERROR VARIANCE OF THE ITEM SUM SCORE Y
 TRUVAR = TV1 + TV2 + TV3 + TV4 + TV5
             +2*(SQRT(TV1*TV2)+SQRT(TV1*TV3)+SQRT(TV1*TV4)+ SQRT(TV1*TV5)
             + SQRT(TV2*TV3)+SQRT(TV2*TV4)+SQRT(TV2*TV5)
             + SQRT(TV3*TV4)+SQRT(TV3*TV5)
             + SQRT(TV4*TV5)); ! THIS IS THE TRUE VARIANCE OF THE ITEM SUM SCORE
 REL = TRUVAR/(TRUVAR+ERRVAR); ! THIS IS THE SCALE RELIABILITY COEFFICIENT, ρY
OUTPUT:     CINTERVAL; ! REQUESTS CONFIDENCE INTERVALS FOR ALL PARAMETERS
```

Note 1. After assigning parametric symbols to the factor loadings and thresholds for all 5 items (P1 through P10), the "new parameter" section introduces successively the item discrimination and difficulty parameters (the *a*'s and the *b*'s, respectively), auxiliary quantities to simplify following code (the *X*'s), item mean true scores (the π's), item true and error variances (the $\sigma^2(\tau_i)$'s), item reliabilities (the ρ_i's), the true and error variances for the item sum score *Y*, and finally the reliability coefficient (ρ_Y). The following 2 sections yield consecutively the item discrimination and difficulty parameters (see Equation (4)), the auxiliary quantities (see equation following (10)), the item mean true scores (see Equations (10) and (11)), true variances, error variances and reliabilities (see correspondingly Equations (5), (7) and (13)), and finally the sum score true variance, error variance, and reliability coefficient (see Equation (14)).

Note 2. Since $\hat{\kappa}_3 < 0$ and $\hat{\kappa}_5 < 0$ for the analyzed data set (which can be found with an initial model fitting without the "model constraint" section), the indicated anti-symmetric feature of the error function *erf*(*X*) requires (i) definition of the auxiliary quantities X_3 and X_5 with a negative sign (see Note 1), and then (ii) adding to .5, rather than subtracting from .5, half of the pertinent error function value as implemented in the above command file (see Equations (10), (11), and subsequent discussion).

Appendix 2

M*plus* Source Code for Point and Interval Estimation of Gain or Loss in Scale Reliability Following Revision

```
TITLE:      POINT AND INTERVAL ESTIMATION OF CHANGE IN COMPOSITE RELIABILITY DUE
            TO DELETION (OR ADDITION) OF BINARY ITEMS.  THIS IS THE COMMAND FILE
            FOR EXAMINING DROP OR GAIN IN RELIABILTY AS A RESULT OF REMOVING THE
            LAST ITEM (SEE MAIN TEXT FOR A MORE GENERAL NOTE).
DATA:       FILE = <name of raw data file>;
VARIABLE:   NAMES = Y1-Y5;
            CATEGORICAL = Y1-Y5;
ANALYSIS:   ESTIMATOR = ML;
MODEL:      F BY Y1* (P1)
            Y2-Y5 (P2-P5);
            F@1;
            [Y1$1-Y5$1] (P6-P10);
 MODEL CONSTRAINT:
 NEW(A1 A2 A3 A4 A5 B1 B2 B3 B4 B5 X1 X2 X3 X4 X5 PI_1 PI_2 PI_3 PI_4 PI_5
 TV1 TV2 TV3 TV4 TV5 EV1 EV2 EV3 EV4 EV5 R1 R2 R3 R4 R5 TRUVAR ERRVAR REL1 REL2 DR);
 A1 = P1/1.702;
 A2 = P2/1.702;
 A3 = P3/1.702;
 A4 = P4/1.702;
 A5 = P5/1.702;
 B1 = P6/P1;
 B2 = P7/P2;
 B3 = P8/P3;
 B4 = P9/P4;
 B5 = P10/P5;
 X1 = A1*B1/SQRT(2+2*A1**2);
 X2 = A2*B2/SQRT(2+2*A2**2);
 X3 = -A3*B3/SQRT(2+2*A3**2);
 X4 = A4*B4/SQRT(2+2*A4**2);
 X5 = -A5*B5/SQRT(2+2*A5**2);
 PI_1 = .5 -.5*(1-1/(1+.278393*X1 + .230389*X1**2
            + .000972*X1**3 + .078108*X1**4)**4);
 PI_2 = .5-.5*(1-1/(1+.278393*X2 + .230389*X2**2
            + .000972*X2**3 + .078108*X2**4)**4);
 PI_3 = .5+.5*(1-1/(1+.278393*X3 + .230389*X3**2
            + .000972*X3**3 + .078108*X3**4)**4);
 PI_4 = .5-.5*(1-1/(1+.278393*X4 + .230389*X4**2
            + .000972*X4**3 + .078108*X4**4)**4);
 PI_5 = .5+.5*(1-1/(1+.278393*X5 + .230389*X5**2
            + .000972*X5**3 + .078108*X5**4)**4);
 EV1 = (.2646 -.118*A1 + .0187*A1**2)*EXP(-.5*(B1/
             (.7427 + .7081/A1 + .0074/A1**2))**2);
 EV2 = (.2646 -.118*A2 + .0187*A2**2)*EXP(-.5*(B2/
             (.7427 + .7081/A2 + .0074/A2**2))**2);
 EV3 = (.2646 -.118*A3 + .0187*A3**2)*EXP(-.5*(B3/
             (.7427 + .7081/A3 + .0074/A3**2))**2);
 EV4 = (.2646 -.118*A4 + .0187*A4**2)*EXP(-.5*(B4/
             (.7427 + .7081/A4 + .0074/A4**2))**2);
 EV5 = (.2646 -.118*A5 + .0187*A5**2)*EXP(-.5*(B5/
            (.7427 + .7081/A5 + .0074/A5**2))**2);
 TV1 = PI_1*(1-PI_1)-EV1;
 TV2 = PI_2*(1-PI_2)-EV2;
 TV3 = PI_3*(1-PI_3)-EV3;
 TV4 = PI_4*(1-PI_4)-EV4;
 TV5 = PI_5*(1-PI_5)-EV5;
 R1 = TV1/(TV1+EV1);
 R2 = TV2/(TV2+EV2);
 R3 = TV3/(TV3+EV3);
 R4 = TV4/(TV4+EV4);
```

```
 R5 = TV5/(TV5+EV5);
 ERRVAR = EV1 + EV2 + EV3 + EV4 + EV5; ! = ERROR VARIANCE OF THE LONGER SCALE
 TRUVAR = TV1 + TV2 + TV3 + TV4 + TV5
             +2*(SQRT(TV1*TV2)+SQRT(TV1*TV3)+SQRT(TV1*TV4)+ SQRT(TV1*TV5)
             + SQRT(TV2*TV3)+SQRT(TV2*TV4)+SQRT(TV2*TV5)
             + SQRT(TV3*TV4)+SQRT(TV3*TV5)
             + SQRT(TV4*TV5)); ! = TRUE VARIANCE OF THE LONGER SCALE
 REL1 = TRUVAR/(TRUVAR+ERRVAR); ! = LONGER SCALE'S RELIABILITY COEFFICIENT
 REL2 = (TRUVAR-TV5-2*(SQRT(TV1*TV5)+SQRT(TV2*TV5)+SQRT(TV3*TV5)+SQRT(TV4*TV5)))/
         (TRUVAR-TV5-2*(SQRT(TV1*TV5)+SQRT(TV2*TV5)+SQRT(TV3*TV5)+SQRT(TV4*TV5))
         +ERRVAR-EV5); ! = SHORTER SCALE'S RELIABILITY COEFFICIENT
 DR = REL2-REL1; ! THIS IS THE CHANGE IN RELIABILITY DUE TO REVISION, Δρ;
OUTPUT:     CINTERVAL;
```

Note. This command file only extends that in Appendix 1 in two places: (a) the “new parameter” section includes additionally the shorter scale’s reliability coefficient (named ‘REL2’) and the revision effect upon reliability (named ‘DR’); and (b) these two parameters are formally defined with the last two input lines immediately before the “output” command. (For consistency, the reliability coefficient of the longer scale is named here ‘REL1’ and is identical to the parameter named ‘REL’ in the source code of Appendix 1.)