Scale Reliability Evaluation With Heterogeneous Populations

Educational and Psychological

Measurement

2015, Vol. 75(5) 875–892

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414558587

epm.sagepub.com

Tenko Raykov¹ and George A. Marcoulides²

Abstract

A latent variable modeling approach for scale reliability evaluation in heterogeneous populations is discussed. The method can be used for point and interval estimation of reliability of multicomponent measuring instruments in populations representing mixtures of an unknown number of latent classes or subpopulations. The procedure is helpful also for evaluation of possible between-class reliability differences as well as of within-class reliability coefficients. The estimation approach can similarly be used in empirical settings with known class membership when distinct populations are investigated, their number is known beforehand and membership in them is observed for the studied subjects, or alternatively in settings where only the number of latent classes is known. A modification and extension of the method for evaluation of maximal reliability or coefficient alpha in heterogeneous populations are also outlined. The discussed procedure is illustrated with numerical data.

Keywords

coefficient alpha, latent class, maximal reliability, mixture, scale, reliability, unobserved heterogeneity, weighted linear combination

Multiple component measuring instruments are highly popular in the educational, behavioral, social, and biomedical sciences as well as in marketing and business. The widespread use of such instruments in these and cognate disciplines is at least in part

¹Michigan State University, East Lansing, MI, USA

²University of California at Santa Barbara, CA, USA

Corresponding Author:

Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443a Erickson Hall, East Lansing, MI 48824, USA.

Email: raykov@msu.edu

due to their particularly useful feature of providing multiple converging pieces of information about underlying latent constructs of substantive interest (e.g., Raykov & Marcoulides, 2011). Psychometric quality of these instruments (often referred to as “scales” below) is a topic that has received a great deal of attention over the past century, with a large part of it being devoted to the concept of reliability. Point and interval estimation of scale reliability has been thereby at the center of interest by methodologists and substantive researchers, and an impressive body of literature on the topic has accumulated over the last three decades or so.

Populations studied in the social and behavioral disciplines are frequently characterized by pronounced heterogeneity resulting from a multiplicity of underlying distinct subpopulations or classes, whereby their number and size are typically unknown as is individual subject membership in them. The resulting latent class mixtures present serious challenges to classical and standard statistical analysis approaches, including psychometric procedures usually employed for evaluation of measuring instrument quality. In particular, when such mixtures are not adequately accounted for as sources of unobserved heterogeneity, biased estimates and incorrect standard errors as well as hypothesis test results can well ensue, with potentially misleading subject-matter interpretations by empirical researchers unaware of possibly substantial unobserved heterogeneity underlying a population under investigation.

This article intends to contribute to bridging an apparent gap in the literature on reliability estimation when evaluating latent constructs in populations representing finite mixtures of latent classes (for simplicity, often referred to as “mixtures” in the sequel). The method discussed below is useful when one is interested in estimation of reliability of psychometric scales employed in such populations, in particular when he/she is concerned with point and interval estimation of a scale’s reliability in each of the latent classes as well as of the possible between-class differences in “precision” (consistency) of measurement. With this property, the following procedure can be especially helpful in modeling and analytic efforts aimed at scale construction and development in contemporary social and behavioral research that frequently deals with questions concerning populations with unobserved heterogeneity, which may represent mixtures of distinct latent classes of substantive relevance. As a by-product, the method is also straightforwardly applicable in settings with known/observed class membership, for instance when a prespecified number of distinct populations are studied and one is interested in evaluating possible reliability differences for a given measuring instrument across them. Similarly, the following procedure is directly used when the number of latent classes is known but not individual class membership, as well as for evaluation of maximal reliability or of the popular coefficient alpha in mixtures (in particular, when alpha is close to reliability in them).

The article is structured as follows. The next section lays out the background, notation, and assumptions underlying the rest of the discussion. A following section outlines a latent variable approach to reliability evaluation when studied populations represent mixtures of latent classes. A subsequent section deals with the empirical application of the method, and a following one illustrates it on numerical data. A

final section is concerned with limitations of the discussed procedure and concludes the article.

Background, Notation, and Assumptions

In the rest of this article, we assume that a set of (approximately) continuous measures are given, which are denoted $X_1, \dots, X_p$ ($p > 1$; see also Conclusion section). The measures are the components or subscales of a unidimensional scale, inventory, questionnaire, test, self-report, testlet, or in general a measuring instrument (frequently referred to as “scale” in the sequel). This scale is also assumed to have been administered to a sample of independent subjects from a studied population with unobserved heterogeneity, which is a mixture of an unknown number k of classes (subpopulations) whereby subject membership in them is similarly unknown or unobserved ($k > 1$).

With these assumptions, the measures in question represent what has been traditionally referred to as a set of congeneric tests (Jöreskog, 1971), with their true scores being perfectly linearly related. That is, based on the classical test theory decomposition of each observed score $X_j = T_j + E_j$ into the sum correspondingly of true and error score,

$$T_j = a_j + b_j T \quad (1)$$

holds, where T_j is the true score of X_j , a_j is the respective intercept, T is the common true score evaluated by these measures, and b_j is the loading of X_j on T ($j = 1, \dots, p$; e.g., Zimmerman, 1975). The error scores $E_1, \dots, E_p$, associated correspondingly with the measures $X_1, \dots, X_p$, may or may not be correlated; in case of correlated errors, we will assume that the overall Model (1) is identified, as we will also in case of uncorrelated errors.¹ The congeneric Model (1) is empirically indistinguishable, in the setting of interest in this article, from the popular single-factor model that corresponds to the hypothesis of unidimensionality or scale homogeneity (e.g., McDonald, 1999). For the aims of the article, the assumption of measurement invariance (e.g., Millsap, 2011) will be similarly adopted as a necessary condition for measuring the same construct in all classes or subpopulations of a studied population. (Methods for its examination are also widely available—e.g., Millsap, 2011; see also conclusion section and Raykov, Marcoulides, & Li, 2012; Raykov, Marcoulides, & Millsap, 2013).

When psychometric scales are employed in empirical research, usually their reliability coefficients (referred to as “reliability” in the sequel) are measurement quality indexes of particular interest to evaluate. Frequently, the reliability of the overall sum score is considered then (but alternatives such as optimal linear combinations are also possible to pursue; e.g., Li, 1997, and references therein; see also below). This sum score reliability, often called composite reliability, is defined as the ratio of true variance to observed variance in the unit-weighted overall sum $Y = X_1 + \dots + X_p$, that is, as

$$\rho_Y = \text{Var}(T_Y) / \text{Var}(Y), \quad (2)$$

where $\text{Var}(\cdot)$ denotes variance, $T_Y = T_1 + \dots + T_p$ is the true score associated with the scale score Y , and ρ_Y symbolizes the reliability of Y .

As shown for instance in Bollen (1989), the (population) composite reliability coefficient ρ_Y can be expressed in terms of the parameters of Model (1) as follows, assuming $b_1 = 1$ for model identification and denoting for convenience $\text{Var}(T) = \phi$:

$$\rho_Y = \phi(1 + b_2 + \dots + b_p)^2 / [\phi(1 + b_2 + \dots + b_p)^2 + v_1 + v_2 + \dots + v_p], \quad (3)$$

where $v_j = \text{Var}(E_j)$ ($j = 1, \dots, p$). Equation (3) will play an instrumental role in the remainder of this article, and we note that in case of error covariances the denominator in its right-hand side is extended by twice their sum (Bollen, 1989).

A Latent Variable Modeling Approach to Reliability Evaluation in Heterogeneous Populations

In the typical finite mixture setting, a studied population consists of k latent classes, whereby their number k is unknown as is their size and the class membership for the individual subjects in it ($k > 1$; e.g., Everitt & Hand, 1981). With the above assumption of measurement invariance, in each class the congeneric Model (1) similarly holds with the same loadings and intercepts, viz.

$$T_{c,j} = a_{c,j} + b_{c,j}T_c \quad (4)$$

($c = 1, \dots, k$), that is, the following two series of equalities are valid

$$a_{1,j} = \dots = a_{k,j} = a_j, \text{ say, and} \quad (5)$$

$$b_{1,j} = \dots = b_{k,j} = b_j, \text{ say,} \quad (6)$$

which will be used in the following discussion ($j = 1, \dots, p$; see also Conclusion section).

Class-Specific Reliability Coefficient

From Equations (3), (5), and (6), the class-specific scale reliability, denoted $\rho_{c,Y}$, is

$$\begin{aligned} \rho_{c,Y} &= \phi_c(1 + b_{c,2} + \dots + b_{c,p})^2 / [\phi_c(1 + b_{c,1} + \dots + b_{c,p})^2 + v_{c,1} + \dots + v_{c,p}] \\ &= \phi_c(1 + b_2 + \dots + b_p)^2 / [\phi_c(1 + b_2 + \dots + b_p)^2 + v_{c,1} + \dots + v_{c,p}] \end{aligned} \quad (7)$$

($c = 1, \dots, k$). (The right-hand side of Equation 7 is readily extended, as indicated above, to the case of error covariances by adding twice their sum in its denominator; e.g., Bollen, 1989.)

The reliability coefficient in Equation (7) is a nonlinear function of the parameters of Model (1), and hence it can be point and interval estimated once that model is fitted to data. The latter is possible using the popular latent variable modeling (LVM) methodology (e.g., Muthén, 2002; see below for further detail and the illustration section for an example). Thereby, a method appropriate for the scale component distribution needs to be employed (e.g., Bollen, 1989). In particular, if the maximum likelihood (ML) method can be used then, an ML estimator of class-specific reliability is obtained by substituting into the right-hand side of Equation (7) the ML estimators of the model parameters, due to the invariance property of ML (e.g., Casella & Berger, 2002).

In addition, whenever a point estimate of reliability and an associated standard error are available, using for instance the monotone transformation-based procedure in Raykov and Marcoulides (2011; see Browne, 1982) and their R function “ci.rel,” one can readily obtain an approximate confidence interval (CI) at a given $100(1 - \gamma)\%$ confidence level ($0 < \gamma < 1$) for the scale reliability coefficient in class c ($c = 1, \dots, k$; see illustration section for an example). Alternatively, since this reliability in Equation (7) is evidently a continuously differentiable function of the model parameters (e.g., Stewart, 1991), the bootstrap method can be used to obtain such a CI (Efron & Tibshiriani, 1993; Muthén & Muthén, 2012).

Between-Class Differences in Reliability

From Equation (7), it follows that the difference in scale reliability across any two classes, say u th and w th ($1 \leq u < w \leq k$), is

$$\Delta\rho_{u,w} = \phi_u(1 + b_2 + \dots + b_p)^2 / [\phi_u(1 + b_2 + \dots + b_p)^2 + v_{u,1} + \dots + v_{u,p}] - \phi_w(1 + b_2 + \dots + b_p)^2 / [\phi_w(1 + b_2 + \dots + b_p)^2 + v_{w,1} + \dots + v_{w,p}]. \quad (8)$$

Equation (8) entails that there will be no differences in the reliability of the used scale across classes if and only if

$$\Delta\rho_{u,w} = 0 \quad (9)$$

for all pairs of indexes u and w , that is, if and only if

$$\begin{aligned} & \phi_u(1 + b_2 + \dots + b_p)^2 / [\phi_u(1 + b_2 + \dots + b_p)^2 + v_{u,1} + \dots + v_{u,p}] \\ & = \phi_w(1 + b_2 + \dots + b_p)^2 / [\phi_w(1 + b_2 + \dots + b_p)^2 + v_{w,1} + \dots + v_{w,p}] \end{aligned} \quad (10)$$

($1 \leq u < w \leq k$).

When used across all k classes, Equation (10) represents a set of $k-1$ nonlinear constraints in terms of parameters of the congeneric Model (1). This Set (10) is in fact a necessary and sufficient condition for class-invariant reliability of a multicomponent measuring instrument in a population representing a mixture of (k) latent

classes. The condition may or may not be fulfilled for a studied population. When a sample from that population is available, and for a particular k , the restriction Set (10) is testable as usual by employing LVM with a pair of nested models—the k -class congeneric model with Constraint (10) that is nested in the same model without that constraint (see next section for further detail, and the illustration section for an example).

Implications for Contemporary Educational, Behavioral, and Social Research

An inspection of the reliability coefficients in Equations (7) and the set of Equations (10) leads us to the following consequential observation for a psychometric scale under consideration. Despite the facts that (a) the same scale consisting of the same components $X_1, \dots, X_p$ is used in all classes (studied population) and (b) measurement invariance holds across the classes, this scale need not have the same psychometric quality—as reflected in the reliability coefficient—in all classes of a given population that is a mixture of them. (The between-class reliability differences may obviously be even more pronounced in Case 5 and/or 6 do not hold, if the “same” construct substantively is still being evaluated in all classes with the used instrument; see also conclusion section.)

This general lack of reliability identity is in our view highly recommendable to keep in mind in contemporary social and behavioral research that is increasingly frequently concerned with populations characterized by pronounced unobserved heterogeneity. For such finite mixture populations, the preceding discussion implies that in general there may be no single meaningful (i.e., no “such thing as”) “reliability of a multicomponent measuring instrument,” but rather possibly multiple in general scale reliability coefficients that are class-specific. In particular, the potential cross-class differences in these coefficients should be appropriately taken into account when considering use of the instrument in empirical studies of these or related populations. (The discrepancies between these reliability coefficients may be even more pronounced if (5) and/or (6) do not hold in a population of interest, if the “same” construct substantively is still being evaluated with the scale in all classes; see also conclusion section.) Conversely, only when there are no between-class differences in reliability could one in our view refer to “reliability of the measuring instrument” in the studied population.

With this in mind, we should like to raise at this point what we find to be an important caution. We submit that manuals of widely used measuring instruments in educational, social, behavioral, and biomedical research, as well as in cognate disciplines, are likely to be referring to reliability estimates obtained in populations used for the scales’ initial study that were treated—possibly incorrectly—as homogeneous at the stage of instrument construction and development, whereas those populations (a) could have been in fact mixtures of substantively discernible latent classes (subpopulations); or (b) could no longer be considered substantively meaningfully homogeneous due to intervening events, historical trends, or development; or (c) their

contemporary counterparts where the instruments are considered for use could be suspected to possess substantial unobserved heterogeneity, based on subject-matter considerations. Such published manual reliability estimates are in general unlikely to be appropriate in present day empirical research, since they may in fact represent potentially misleading “average” reliability estimates that are not valid in any of the classes of a similar population of interest, in addition to being biased estimates of individual class reliabilities of actual relevance. Instead of referring to those manuals and reliability estimates found there, application of the method discussed in this article could well be recommended when considering use of a particular scale in a study of a population possibly representing a mixture of latent classes.

Factor Mixture Modeling as Statistical Framework for Mixture Reliability Evaluation

The preceding discussion has laid the foundation of a generally applicable procedure for scale reliability evaluation in populations characterized by pronounced, unobserved heterogeneity. Given that such mixtures typically consist of an unknown number of latent classes, this procedure is inextricably intertwined with the increasingly popular latent class analysis (LCA) in the social and behavioral sciences (e.g., Lubke & Muthén, 2005).

Standard or classic LCA assumes that the observed measures are uncorrelated (independent) within class, following its fundamental assumption of “conditional independence” (e.g., McCutcheon, 1987). This assumption does not hold, however, in the setting of concern in this article. The reason is that the latter is based on the congeneric test Model (1) (single-factor model), the most widely used framework presently for unidimensional scales (see, e.g., McDonald, 1999, and references therein, also for possible extensions within the nonlinear factor analysis framework; e.g., Raykov & Marcoulides, 2011). As can be readily seen from Equation (1) (see also Equation 4), the congeneric model implies that within each of the classes the observed measures are (still) interrelated, a general feature resulting from the presence of latent variability and covariability underlying the observed measures within each class (see below and Note 1).

This “residual interrelation” can be readily accommodated with an appropriate extension of conventional or classic LCA, the so-called factor mixture modeling (FMM; e.g., Lubke & Muthén, 2005). Accordingly, in FMM the within-class covariance of observed measures is assumed to be explained by one or more continuous latent variables, in addition to the categorical latent variable representing class membership for the entire (overall) model in the studied population. This is precisely the setting underlying this article, however. The reason is that the congeneric test model it is based on, effectively assumes within each class the common true score as the source of observed measure covariability or interrelation (see Equation 1 and immediately following discussion, as well as Equation 4). In other words, it is this latent variable—the common true score—that explains in the model under consideration

the within-class violations of conditional independence, and thus plays the role of the continuous latent variable in FMM “producing” within-class measure interrelationships. For this reason, FMM is the statistical modeling framework that underlies the remainder of the article, and in fact represents the first step in an application of its procedure.

Empirical Application of Method in Heterogeneous Populations

To accomplish scale reliability evaluation in populations that are mixtures of latent classes (for simplicity referred to as “mixture reliability evaluation” below), a scholar needs to proceed in two steps. In the first, he/she engages in FMM of the sample data from a studied population (assuming the sample is representative of the population and in particular of all its latent classes). That mixture analysis is readily carried out using the popular LVM methodology, aims to “determine” the number of latent classes, and is conducted by fitting the congeneric Model (1) with 2, 3, 4, and so on, classes to carry out model selection with respect to number of latent classes (e.g., Geiser, 2013; see also the model definition related discussion after Equation 1, as well as the conclusion section).

This model selection process can be based as usual on the BIC index, misclassification rate, hypothesis tests for dropping the last added class (including in particular the bootstrap likelihood ratio test; e.g., Nylund, Asparouhov, & Muthén, 2007), and in empirical research especially on substantive considerations (e.g., Hancock & Samuelsen, 2007; Muthén, 2014). Thereby, a test of interest for the researcher may well be that examining the need of considering 2 classes rather than 1 class (in particular, the bootstrap likelihood ratio test for 1 vs. 2 classes; see next section). This test, routinely provided by LVM software such as Mplus (Muthén & Muthén, 2012), addresses the possibility of the studied population being homogeneous enough (i.e., not heterogeneous to a pronounced degree) to be treated as consisting of a single class. If the pertinent test statistic turns out to be nonsignificant, one can retain its null hypothesis of a single class, on which then standard modeling and in particular conventional reliability estimation methods would be applicable (e.g., Raykov & Marcoulides, 2011; see also Raykov, 2012). Otherwise, the procedure of the present article can be used. The end result of this first analytic step is the selection—as most preferable—of a model with a certain number of classes, denoted as before k ($k > 0$; we assume in the rest of this article $k > 1$, since the case $k = 1$ has already been handled in the literature, e.g., in the last cited two sources and pertinent references therein).

In the second step, given a selected model with k classes say from the first step (FMM), the researcher can test for cross-class identity in reliability of the scale under consideration. This is achieved by testing Equations (10) that represent as a set a necessary and sufficient condition for lack of class-differences in scale reliability, as pointed out earlier. This null hypothesis stipulating equality of all k class reliability coefficients, that is,

$$\rho_{c,Y} = \rho_{c+1,Y} \quad (11)$$

($c = 1, \dots, k-1$), can be readily tested using LVM and a pair of nested models as indicated earlier. These are (a) Model (1) with k classes and Constraint (11) (for $c = 1, \dots, k-1$), which is nested in (b) Model (1) with k classes but without this Constraint. We note that with k classes this hypothesis testing is equivalent to testing $k-1$ parameter restrictions simultaneously within the selected model.

If the null Hypothesis (11) is rejected, or instead of testing it, a scholar can also point and interval estimate any two classes' differences in scale reliability. This is achieved by point and interval estimation of the reliability difference $\Delta\rho_{u,w}$ in Equation (8) for a pair of given classes u and w ($1 \leq u < w \leq k$), and is also readily accomplished using LVM. To this end, the difference $\Delta\rho_{u,w}$ is introduced as an "external" model parameter, defined as $\rho_{u,Y} - \rho_{w,Y}$ for classes u and w , and its interval estimation is requested from the software (see Appendix A for the Mplus source code needed then). The resulting CIs for the class differences in reliability, which are based on the popular delta-method (e.g., Raykov & Marcoulides, 2004), provide each a range of plausible values of the corresponding scale reliability class difference in the overall population under investigation (and could also be used for hypothesis testing, in particular of simple or point hypotheses; e.g., Raykov & Marcoulides, 2008).

Method Application With Known Class (Group) Membership

When the number of latent classes and subjects' class membership are known in an empirical setting, the described procedure is straightforwardly utilizable, viz., as a multiple-group LVM application (e.g., Muthén & Muthén, 2012). Specifically, in that case one can (a) test for group (population) differences in scale reliability, (b) point and interval estimate the scale reliability in each group (population), and (c) point and interval estimate the scale reliability differences across any two groups (populations). To this end, all one needs to do is employ the above modeling approach while using the known group membership information (i.e., after providing this information to the software as a separate, observed variable defining the groups; formally one uses then Equations (7) through (10) with the subindex " c " being substituted with " g " for known group/population membership; $g = 1, \dots, G$, where G is number of studied groups, $G > 1$). Hence, when both the number of classes k and the individuals' class memberships are known, the method of this article can be seen as being effectively a LVM procedure for studying reliability differences across a prespecified set of distinct populations under consideration.

Method Application With Known Number of Classes

In settings where only the number k of classes is known but not the subjects' membership in them (e.g., in some biomedical studies), the procedure of this article is used in the following way. One fits then just the mixture model of this article for that known

number k of classes. That is, one does not become involved in “determining” the number of classes, but apart from that activity uses all remaining ones constituting the procedure outlined in this article. In other words, one employs then the Mplus command files in Appendix A just for that known number k of classes (as well as the R-function in Appendix B on the so-obtained results; in Appendix A, that command file is provided for the case $k = 2$).

Modification of Method for Evaluation of Maximal Reliability and Coefficient Alpha in Heterogeneous Populations

In some empirical settings, a scholar may be interested in working with the optimal linear combination of the components of a given measuring instrument rather than, or in addition to, their overall (unweighted) sum score Y of interest thus far, or alternatively with the popular coefficient alpha (see, e.g., Raykov, 2012, for its limitations, and Raykov & Marcoulides, 2014, for empirically testable conditions under which it is close to scale reliability in a studied population). As discussed in detail in the literature, the optimal linear combination is associated with maximal possible reliability for the given instrument (linear combination of its components), and possesses also maximal criterion validity with respect to any criterion variable that is uncorrelated with the error terms in Model (1) (e.g., Li, 1997, and references therein; Penev & Raykov, 2006).

When of concern is to point and interval estimate maximal reliability associated with a given measuring instrument used in a heterogeneous population, on the assumption of Model (1) underlying this article with uncorrelated errors, the method described above is directly applied with a minor modification. The latter consists in using then the corresponding formula for maximal reliability rather than composite reliability any time the latter is used in the above description of this article’s procedure. To this end, all one needs to do is employ instead of unitary weights in the sum score, $Y = X_1 + \dots + X_p$, the weights $w_j = b_j/v_j$ for the corresponding instrument components, X_j ($j = 1, \dots, p$; e.g., Raykov, 2012).

In cases where it is of interest to point and interval estimate coefficient alpha in a heterogeneous population, the method of this article is readily used in conjunction with the approach in Raykov and Marcoulides (2014; in particular, if alpha is close to scale reliability in all classes; a sufficient and testable condition for this proximity of alpha to reliability in a given population or subpopulation of it, is provided there as well). To accomplish this goal all one needs to do is use the expression for coefficient alpha in terms of latent variances and covariances of the pertinent saturated model for the components of the instrument (e.g., Raykov & Marcoulides, 2014), any time the formula for composite reliability is employed in the above description of the method of the present article. (This alpha evaluation is conducted after first “determining” the number of latent classes, as discussed in the previous section of the article.)

We demonstrate next on numerical data the outlined method of reliability evaluation in heterogeneous (mixture) populations.

Table 1. Model Selection Indexes for Fitted Factor Mixture Models.

<i>k</i>	BIC	<i>p</i> -ALMR	<i>p</i> -BLRT
1	14546.219	n/a	n/a
2	14034.648	.0	.0
3	14059.341	.105	1.0

Note. *k* = number of classes; *p*-ALMR = *p*-value for adjusted Lo–Mendell–Rubin test (e.g., Geiser, 2013); *p*-ALMR = *p*-value for bootstrap likelihood ratio test (e.g., Nylund et al., 2007); n/a = not applicable. (See also Note 2.)

Illustration on Data

For the aims of this section, we use simulated data with *n* = 1,000 cases for a scale consisting of *p* = 5 components using the congeneric Model 1 in each of *k* = 2 classes with class (membership) prevalences of .5 each. Specifically, in the first class multi-normal data were generated according to the following model (cf. Equations 4):

$$\begin{aligned} X_1 &= T + E_1, \\ X_2 &= .75T + E_2, \\ X_3 &= .75T + E_3, \\ X_4 &= .75T + E_4, \\ X_5 &= 2T + E_5, \end{aligned} \tag{12}$$

where *T* was standard normal and the error terms were zero-mean independent normal variates with variances .5 each. The data in the second class were generated using the same Model (12), but with *T* normal with mean 2 and variance 2 and the error terms as independent standard normal variates each. (Further details on the simulation procedure can be obtained from the authors on request.)

Next, in Step 1 of the procedure outlined in this article we fit the congeneric (factor mixture) Model (1) successively with *k* = 1, 2, and 3 classes.² (The Mplus command file needed is presented in Appendix A.) The resulting BIC indexes as well as *p* values of the adjusted Lo–Mendell–Rubin test and the bootstrap likelihood ratio test are presented in Table 1.

Table 1 suggests to select the *k* = 2 class model based on the BIC, since the latter index is the smallest at *k* = 2 classes; in addition, the two-class model is found to be also associated with the lowest misclassification rate (Geiser, 2013), providing additional support for that model (see also Note 2). Based on these indexes, we select the two-class model from the fitted model series. This is a correct decision, and one that is not unexpected given that the data were generated using a two-class model.

In the second step of the procedure of this article, we use the selected two-class model to point and interval estimate the possible class difference in reliability of the scale consisting of the 5 components under consideration (overall sum score; see

second Mplus source code in Appendix A). To this end, we introduce the reliability difference in Equation (8) as an “external” parameter and request its interval estimation, which also yields CIs of the two class-specific reliabilities. (By examining the resulting CI for the class-difference in reliability, one can also test the hypothesis of identity in the class-specific reliabilities, as mentioned earlier; see also below.) We note that this external parameter introduction does not change the fitted model or its fit indexes and degrees of freedom, nor does it alter the model parameter estimates and standard errors, since it does not have implications with respect to the analyzed data in addition to those of the earlier fitted two-class model (before this “external” parameter inclusion).

The resulting scale reliability estimates were as follows in the two classes (standard errors presented within parentheses):

$$\hat{\rho}_{1,Y} = .929(.007) \quad (13)$$

and

$$\hat{\rho}_{2,Y} = .860(.013). \quad (14)$$

Using these estimates and standard errors, the R-function “ci.rel” in Raykov and Marcoulides (2011; see also Appendix B) furnishes the 95% CI for the scale reliability in Class 1 as (.914, .942), and that interval for the scale reliability in Class 2 as (.832, .884). We note that these two CIs do not overlap, suggesting that the scale reliability in Class 2 is considerably lower than that reliability in Class 1 (see also below).

The class difference in reliability is also estimated in the last fitted, two-class model as follows (the 95% CI provided by the software and based on the delta-method, is stated within final parentheses)

$$\hat{\Delta\rho}_{1,2} = .070(.014)(.043, .097). \quad (15)$$

Since the reliability class difference CI in (15) does not cover 0, it is suggested that at the .05 significance level the null hypothesis of class identity could be rejected; thereby, it is suggested that a range of plausible population values, at the 95% confidence level, stretches from .043 through .097.

Since we know here the (true) parameters of the model having generated the data, we can use with them the scale reliability parametric expression in Equation (3) to find out the true scale reliability coefficients in each of the two classes. Proceeding in this way, we obtain (e.g., using simple calculators or alternative software) $\rho_{1,Y} = .928$ and $\rho_{2,Y} = .868$. These class reliability coefficients are very close to their estimates reported in Equations (14) and are covered by their 95% CIs presented above. In addition, the true difference in class reliability coefficients is $\Delta\rho_{1,2} = .060$, which is similarly rather close to the estimated class difference in scale reliability in Equation (15) and is covered by its above CI (see Equation 15).

Conclusion

This article was concerned with reliability estimation in populations characterized by unobserved heterogeneity and representing mixtures of latent classes, which are becoming of increasing interest in the educational, behavioral, social, and biomedical disciplines, as well as in marketing and business research. The discussion aimed to contribute to closing an apparent gap in the reliability related literature that did not seem to contain a readily and widely applicable procedure for point and interval estimation of scale reliability in mixture populations. The outlined method permits a researcher (a) to examine for identity class-specific reliability coefficients, (b) to point and interval estimate their difference as well as these within-class reliabilities, and relatedly (c) to ascertain whether there are between-class differences in scale reliability. In addition, the method can also be used for (d) examining scale reliability differences in given distinct populations when class membership is known, and (e) point and interval estimation of their difference then (as well as testing various hypotheses about this difference, as in the typical mixture case). The procedure is also applicable in empirical settings where only the number of latent classes is known but not subjects' membership in them. A direct modification and extension of the described approach is utilizable when of interest is point and interval estimation of maximal reliability or of coefficient alpha in heterogeneous populations. The method is straight-forwardly applicable using popular software, such as the LVM program Mplus (Muthén & Muthén, 2012).

As a main message of this article, we would like to also see the caution we raised that in general there may be no meaningful single scale reliability coefficient of relevance for a population that is a mixture of two or more substantively distinct latent classes. Whether that is the case, can be addressed with the outlined procedure followed by thorough substantive interpretations in an empirical setting. In this connection, we should like to point out that in our view one could only then speak meaningfully of reliability of a multicomponent measuring instrument for a studied (mixture) population, when there is support in a representative data set from it for lack of between-class differences in the scale's reliability, that is, the null hypothesis of no class differences in it is not rejected.

Similarly, we should like to caution in particular educational and behavioral researchers selecting their measurement instruments based on inspection of pertinent manuals provided by their publishers. The reason is that these manuals may be based (a) on outdated population definitions (i.e., populations may have changed qualitatively since the publication of the manuals), and no less importantly (b) on pilot studies that considered their sampled populations as consisting of a single class. With the increasing diversity of the populations of interest in the social and behavioral sciences, this earlier assumption of them representing a single class is in our opinion becoming increasingly likely to be violated. Hence, the reliability estimates found in instrument manuals are likely to be potentially biased and not applicable for one, several, or any of the latent classes (subpopulations) in a contemporary population of interest in these and cognate disciplines, if not even misleading.

Furthermore, we would like to mention that depending on whether a scholar adopts the measurement invariance assumption or not, may have an impact on the number of latent classes “determined” using the corresponding conventional procedure within FMM (e.g., Geiser, 2013; Lubke & Muthén, 2005), which is used in the first step of an application of the method of this article. We would argue for using as often as possible this assumption (or that of partial measurement invariance), which is testable and enhances substantially the trust one could have in the presumption of measuring the same latent construct in all classes (subpopulations) of a given heterogeneous populations (e.g., Millsap, 2011; see also Raykov et al., 2012, 2013).

The discussed reliability estimation procedure has several limitations. One, as described, it is currently applicable with (approximately) continuous scale components. With normal components of a used instrument, direct application of the ML estimation and testing approach is appropriate, which renders ML estimates of the class-specific reliability coefficients as well as of their differences (Equation 8). With up to mild deviations from normality, which do not result from piling at scale end for an individual component(s), use of the robust ML method is recommendable (MLR; Muthén & Muthén, 2012), perhaps with components having as few as 5 to 7 possible answer options and in particular relatively symmetric distributions. We encourage future research to address possible robustness of this approach in situations where one or more scale components exhibit a limited number of possible scores and/or very asymmetric distributions.

Similarly, as indicated at the outset, we assumed throughout that sampled subjects are independent of each other, that is, they are not clustered or nested within Level-2 or higher-order units, such as teams, schools, clinicians, managers, interviewers, cities, companies, physicians, hospitals, neighborhoods, and so on. We may conjecture that the MLR method may also have some robustness to such violations of the classical independence assumption, particularly when their extent is limited as is the degree of nonnormality of scale components (or, preferably, the instrument components are normal or essentially so to begin with). Further research is needed, however, before one may place trust in such a potential recommendation.

Last but not least, the discussed reliability estimation approach is best used with large samples. This follows from the fact that it rests on ML or robust ML estimation, the key methods in mixture analysis at present, which are however grounded themselves in asymptotic statistical theory (e.g., Muthén, 2002). We encourage future research aimed at developing possible guidelines to help determine sample size when one could rely on the underlying large-sample estimation theory as having obtained practical relevance in a given empirical study.

In conclusion, this article offers to educational, behavioral, social, biomedical, marketing, and business scientists a widely applicable means for reliability and related indexes evaluation in populations with unobserved heterogeneity that are mixtures of latent classes. The outlined method permits one to make more informed conclusions than about psychometric scale measurement quality as reflected in a reliability coefficient, and can be readily used in studies of populations with

pronounced unobserved heterogeneity that are becoming of increasing interest in contemporary research in these and cognate disciplines.

Appendix A

Mplus Source Code for Reliability Evaluation in Heterogeneous Populations

```

TITLE:           MIXTURE RELIABILITY EVALUATION. STEP 1.
                  TWO-CLASS MODEL. (SEE NOTE BELOW.)

DATA:           FILE = <NAME OF RAW DATA FILE> ;

VARIABLE:       NAMES ARE y1-y5;
                  CLASSES = c(2) ;

ANALYSIS:       TYPE = MIXTURE;
                  STARTS = 500 50;

MODEL:          %OVERALL%
                  f BY y1-y5;
                  [f@0];
                  %c#2%
                  y1-y5;
                  f;
                  [f];

OUTPUT:         TECH11 TECH14;

```

Note. When fitting the model with fewer/more classes, delete/add the pertinent section for the dropped/added classes (starting with “%class#% and finishing with “[f*1]”; the two-class model is preferred in the illustration example).

```

TITLE:           MIXTURE RELIABILITY EVALUATION. STEP 2.
                  (TWO-CLASS MODEL.)

DATA:           FILE = <NAME OF RAW DATA> ;

VARIABLE:       NAMES ARE Y1-Y5;
                  CLASSES = C(2) ;

ANALYSIS:       TYPE = MIXTURE;
                  STARTS = 500 50;

MODEL:          %OVERALL%
                  F BY Y1@1

```

```

Y2-Y5 (B2-B5) ;
Y1-Y5 (V11-V15) ;
F*1 (LVC1) ;
[F@0] ;
%C#2%
Y1-Y5 (V21-V25) ;
F (LVC2) ;
[F] ;

```

MODEL CONSTRAINT:

```

NEW(RHO1, RHO2, DELTA) ;
RHO1 = LVC1*(1 + B2 + B3 + B4 + B5)**2
/ (LVC1*(1 + B2 + B3 + B4 + B5)**2 + V11 + V12
+ V13 + V14 + V15) ;
RHO2 = LVC2*(1 + B2 + B3 + B4 + B5)**2
/ (LVC2*(1 + B2 + B3 + B4 + B5)**2 + V21 + V22
+ V23 + V24 + V25) ;
DELTA = RHO1-RHO2 ;

```

OUTPUT: CINTERVAL;

Appendix B

R-function for Interval Estimation of Class-Specific Scale Reliability

```

ci_cs.rel = function(r, se) {
  l = log(r/(1-r))
  sel = se/(r*(1-r))
  ci_l_lo = l-1.96*sel
  ci_l_up = l + 1.96*sel
  ci_lo = 1/(1 + exp(-ci_l_lo))
  ci_up = 1/(1 + exp(-ci_l_up))
  ci = c(ci_lo, ci_up)
  ci
}

```


Note 1. This function is a trivial adaptation of the R-function ‘ci.rel’ in Raykov and Marcoulides (2011) for the purposes of this discussion, and is presented here for completeness of the present article.

Note 2. At the R prompt, type/submit ‘ci_cs.rel(rho, se)’, where for ‘rho’ the class-specific reliability estimate is entered and for ‘se’ its associated standard error.

Acknowledgments

We thank B. Muthén and L. Muthén for instructive and helpful comments on latent class analysis and its applications, as well as to G. Day for valuable and insightful discussions on mixture modeling. We are also grateful to two anonymous referees for critical comments on an earlier version of the article, which contributed significantly to its improvement.

Declaration of Conflicting Interests

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

Funding

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

Notes

1. If $p = 2$, we assume that additional constraints are imposed, for example, loading equality or error variance equality, in order for the congeneric Model (1) to be identified. We also presume that (a) for none of the loadings $b_j = 0$ is true, (b) the error scores E_j have each positive variance, and (c) the factor variance is positive, that is, $Var(T) > 0$ ($j = 1, \dots, p$). The assumptions (a) through (c) may be considered hardly restrictive if at all in empirical research, however, as they could be typically expected to be fulfilled then.
2. The 4-class model was associated with numerical problems and formally with a BIC marginally higher than that of the 3-class model. We view these problems as consistent with an attempt to fit a model with a notably higher number of classes than the model used in the data generation process, and for this reason the 4-class model results are not presented in Table 1.

References

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72-141). Cambridge, England: Cambridge University Press.
- Casella, G., & Berger, J. (2002). *Statistical inference*. Monterey, CA: Wadsworth.
- Efron, B. J., & Tibshiriani, R. (1993). *An introduction to the bootstrap*. London, England: Chapman & Hall.
- Everitt, B., & Hand, D. (1981). *Finite mixture distributions*. London, England: Chapman & Hall.

- Geiser, C. (2013). *Data analysis with Mplus*. New York, NY: Guilford.
- Hancock, G. R., & Samuelson, K. M. (2007). *Advances in latent variable mixture models*. New Haven, CT: Information Age.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109-133.
- Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika*, 62, 245-249.
- Lubke, G., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21-39.
- McCutcheon, A. (1987). *Latent class analysis*. Thousand Oaks, CA: Sage.
- McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ: Erlbaum.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor & Francis.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81-117.
- Muthén, B. O. (2014, July 22). *Advances in mixture modeling and more*. Keynote address at the 79th International Psychometric Society Meeting, Madison, WI.
- Muthén, L. K., & Muthén, B. (2012). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- Penev, S., & Raykov, T. (2006). On the relationship between maximal reliability and maximal validity for linear composites. *Multivariate Behavioral Research*, 41, 105-126.
- Raykov, T. (2012). Scale development using structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472-492). New York, NY: Guilford Press.
- 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.
- Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. New York, NY: Taylor & Francis.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- Raykov, T., & Marcoulides, G. A. (2014). A direct latent variable modeling based method for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement*. Advance online publication. doi: 10.1177/0013164414526039
- Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, 72, 954-974.
- Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. *Educational and Psychological Measurement*, 73, 713-727.
- Stewart, J. (1991). *Calculus*. Monterey, CA: Wadsworth.
- Zimmerman, D. W. (1975). Probability measures, Hilbert spaces, and the axioms of classical test theory. *Psychometrika*, 40, 395-412.