Evaluation of Measurement Instrument Criterion Validity in Finite Mixture Settings

Educational and Psychological

Measurement

2016, Vol. 76(6) 1026–1044

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415613542

epm.sagepub.com

Tenko Raykov¹, George A. Marcoulides², and Tenglong Li¹

Abstract

A method for evaluating the validity of multicomponent measurement instruments in heterogeneous populations is discussed. The procedure can be used for point and interval estimation of criterion validity of linear composites in populations representing mixtures of an unknown number of latent classes. The approach permits also the evaluation of between-class validity differences as well as within-class validity coefficients. The method can similarly be used with known class membership when distinct populations are investigated, their number is known beforehand and membership in them is observed for the studied subjects, as well as in settings where only the number of latent classes is known. The discussed procedure is illustrated with numerical data.

Keywords

criterion validity, latent class, mixture, scale, unobserved heterogeneity, validity

Measuring Instrument Validity Evaluation in Finite Mixture Settings

Instrument validity is of paramount importance in educational and psychological research as it is fundamental to any latent construct measurement attempt in these

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, East Lansing, MI 48824, USA.

Email: raykov@msu.edu

and related sciences. Multicomponent instruments are highly popular in these disciplines, and their widespread use is due in part to their particularly useful feature of providing multiple converging pieces of information about underlying constructs of substantive interest (e.g., Raykov & Marcoulides, 2011). The quality of these psychometric instruments (often referred to as “scales” below) is a topic that has received a great deal of attention by methodologists and quantitative scientists over the past century. Although reliability estimation seems to have attracted the larger share of this interest, formal approaches to evaluation of validity have also drawn considerable attention.

While an impressive body of literature has emerged over the last several decades on validity coefficient estimation, the overwhelming part of it has been concerned with what may be referred to as standard single-class methods. A main limitation of these methods is that they do not account for the fact that many studied populations are characterized by substantial unobserved heterogeneity due to multiple underlying distinct subpopulations or classes. To complicate matters, their number and size are typically unknown as is also subject membership in them (e.g., Everitt & Hand, 1981). These latent class mixtures therefore present in general serious challenges to classical and standard statistical analysis approaches, including procedures usually employed for the evaluation of instrument quality. Specifically, when mixtures are not adequately accounted for and thus single-class methods are used, biased estimates and incorrect standard errors as well as hypothesis test results can well ensue, in particular for validity coefficients. As a consequence, when population heterogeneity is not properly handled, misleading subject-matter interpretations are likely to follow.

Recently, Raykov and Marcoulides (2015) outlined a readily applicable latent variable modeling procedure for point and interval estimation of scale reliability in mixture settings, which could also be used to evaluate between-class differences in instrument reliability as well as within-class reliability coefficients. The aim of the present article is to extend that approach to scale criterion validity evaluation, including between-class evaluation and testing of validity differences and within-class validity coefficients. The following procedure is especially helpful in modeling and analytic efforts aimed at instrument construction and development in contemporary educational and psychological research, which frequently deals with populations characterized by unobserved heterogeneity that may be mixtures of distinct latent classes of theoretical and empirical relevance. The method is also straightforwardly applicable with known or observed class membership, for instance, when a prespecified number of populations are studied and one is interested in evaluating possible criterion validity differences for a given multicomponent instrument across them. Similarly, the approach can be directly used when the number of latent classes is known but not individual class membership. In addition, the outlined procedure can be utilized for the evaluation of criterion validity of optimal linear combinations of a considered instrument's components that are associated with maximal reliability and validity, and permits studying the scale's latent structure as well as addressing its measurement invariance properties.

Background, Notation, and Assumptions

In this article, we assume that a set of (approximately) continuous homogeneous measures are given, denoted $X_1, \dots, X_p$, which are frequently referred to as “components” in the sequel ($p > 1$; see also conclusion section).¹ In empirical research, oftentimes these components are elements or subscale scores of a unidimensional scale, inventory, questionnaire, test, self-report, testlet, test battery, or in general a measurement instrument. We also presume that the instrument has been administered to a sample of independent subjects from a studied population with unobserved heterogeneity that is a mixture of an unknown number k of classes (subpopulations; $k > 1$), with subject membership in them being similarly unknown or unobserved.

With these assumptions, the measures $X_1, \dots, X_p$ represent what is widely known as a set of congeneric tests (Jöreskog, 1971). That is, using 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 in this set of manifest variables, where T_j is the true score of X_j , a_j is the associated intercept, T is the common true score evaluated by the p measures, and b_j is the loading of X_j on T (for instance, T could be taken as T_1 ; $j = 1, \dots, p$; e.g., Zimmerman, 1975). The error scores $E_1, \dots, E_p$ may or may not be correlated, with either of these two circumstances being an assumption; in case of correlated errors, it is presumed that the overall Model (1) is identified and plausible for a given data set, and similarly with correlated errors. For identification purposes, we set $b_1 = 1$, with the variance of the common true score being a free parameter denoted $\phi = \text{Var}(T)$, where $\text{Var}(\cdot)$ symbolizes variance. We note in passing that the congeneric Model (1) is empirically indistinguishable, in the setting of concern in the remainder, from the popular single-factor model pertaining to the hypothesis of unidimensionality or scale homogeneity (e.g., McDonald, 1999). Last but not least, for the goals of this article the assumption of measurement invariance (MI; e.g., Millsap, 2011) is adopted, as a necessary condition for measuring the same construct in all classes or subpopulations of a studied population. (Methods for examining MI are widely available—e.g., Millsap, 2011; see also conclusion section and Raykov, Marcoulides, & Li, 2012; Raykov, Marcoulides, & Millsap, 2013.)

When employing psychometric scales in empirical research, usually their overall sum scores are of concern, but also their weighted linear combinations may be more generally of interest. As indicated earlier, of major relevance in either case is the validity of the scale (overall sum or weighted sum score). A particular type of validity, referred to as criterion validity, can be used thereby to address essential aspects of validity (e.g., Allen & Yen, 2001). As is well known, criterion validity can be quantified in the correlation coefficient between the sum score—unweighted or weighted—with a preselected criterion variable, denoted Z in the rest of this discussion, which measure is usually chosen on substantive grounds (e.g., Crocker &

Algina, 2006). That is, this validity can be formally defined as the correlation of the criterion with the sum score:

$$v_Y = \text{Corr}(Y, Z), \quad (2)$$

where $\text{Corr}(\cdot)$ denotes correlation and $Y = w_1X_1 + \dots + w_pX_p$ is the general form of that score, with w_j being the component weights that as a special case can all be equal to 1 ($j = 1, \dots, p$).

Criterion Validity Evaluation in Mixture Settings Using Latent Variable Modeling

To outline the validity point and interval estimation method of concern in this article, we begin by noting that from Equation (2) and the preceding discussion, assuming uncorrelated criterion with error scores, straightforward algebra yields in the uncorrelated error case the criterion validity coefficient as

$$v_Y = \frac{\sum_{j=1}^p w_j b_j \sigma_{T,Z}}{\sigma_Z \sqrt{\phi \sum_{j=1}^p w_j^2 b_j^2 + 2\phi \sum_{i < j}^p w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_j}}, \quad (3)$$

where $\sigma_{T,Z} = \text{Cov}(T, Z)$ is the covariance of the common true score T and the criterion Z (with $\text{Cov}(\cdot, \cdot)$ denoting covariance), $\theta_j = \text{Var}(E_j)$ are the error variances, and as mentioned $b_1 = 1$ is set ($j = 1, \dots, p$). In case of nonzero error covariances, the square-rooted expression in the denominator of Equation (3) is extended by twice their correspondingly weighted sum (Bollen, 1989).

We next observe that in a typical finite mixture setting a studied population consists of k latent classes, whereby k is unknown as is their size and the class membership for individual subjects ($k > 1$; e.g., Everitt & Hand, 1981). Based on the earlier made assumption of ML, in each class the congeneric Model (1) holds with the same loadings and intercepts, viz.

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

where the subindex “ c ” is used to denote latent class and the following two sets of equalities are valid:

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

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

($c = 1, \dots, k$; $j = 1, \dots, p$; see also conclusion section).

Class-Specific Criterion Validity Coefficients

From Equations 3 through 6, the class-specific scale criterion validity coefficient, denoted $v_{Y,c}$, results as follows:

$$v_{Y,c} = \frac{\sum_{j=1}^p w_j b_j \sigma_{c;T,Z}}{\sigma_Z \sqrt{\phi_c \sum_{j=1}^p w_j^2 b_j^2 + 2\phi_c \sum_{i<j} w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_{c;j}}}, \quad (7)$$

where ϕ_c and $\sigma_{c;T,Z}$ are the latent variance and covariance with the criterion, respectively, and $\theta_{c;j}$ are the error variances in the c th class ($c = 1, \dots, k$; $j = 1, \dots, p$). (The right-hand side of Equation (7) is readily extended, as indicated above, in case of error covariances, by adding twice their correspondingly weighted sum under the radical sign; e.g., Bollen, 1989.)

The validity coefficient in Equation (7) is obviously a nonlinear function of the parameters of Model (1). Therefore, it can be point and interval estimated once that model is fitted to data (assuming that it is found plausible, as indicated earlier). This is possible using the popular latent variable modeling (LVM) methodology (e.g., Muthén, 2002; further details are provided below; see the illustration section for an example and the appendix for the needed source code). An estimation method appropriate for the observed component distribution needs to be employed thereby (e.g., Bollen, 1989); in particular, if the maximum likelihood (ML) method can be used, an ML estimator of class-specific validity 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).

Moreover, once a point estimate of criterion validity and an associated standard error become available, using for instance the monotone transformation-based procedure in Raykov and Marcoulides (2011; see also Browne, 1982), one can readily obtain an approximate confidence interval at a given $100(1 - \gamma)\%$ confidence level ($0 < \gamma < 1$) for the scale criterion validity coefficient in class c ($c = 1, \dots, k$; see illustration section for an example and the appendix for an R-function accomplishing this interval estimation). Alternatively, due to the validity coefficient in Equation (7) being a continuously differentiable function of the model parameters (e.g., Apostol, 2007), the bootstrap method can be employed to obtain that confidence interval (Efron & Tibshiriani, 1993; Muthén & Muthén, 2014).

Class Differences in Criterion Validity

From Equation (7), the difference in the criterion validity coefficients across any two classes, say s th and u th ($1 \leq s < u \leq k$), is as follows:

$$\Delta v_{s,u} = \frac{\frac{\sum_{j=1}^p w_j b_j \sigma_{s;T,Z}}{\sigma_Z \sqrt{\phi_s \sum_{j=1}^p w_j^2 b_j^2 + 2\phi_s \sum_{i<j}^p w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_{s;j}}}}{\frac{\sum_{j=1}^p w_j b_j \sigma_{u;T,Z}}{\sigma_Z \sqrt{\phi_u \sum_{j=1}^p w_j^2 b_j^2 + 2\phi_u \sum_{i<j}^p w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_{u;j}}}} \quad (8)$$

Based on Equation (8), there will be no class differences in the criterion validity of the used instrument if and only if

$$\Delta v_{s,u} = 0 \quad (9)$$

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

$$\frac{\frac{\sum_{j=1}^p w_j b_j \sigma_{s;T,Z}}{\sigma_Z \sqrt{\phi_s \sum_{j=1}^p w_j^2 b_j^2 + 2\phi_s \sum_{i<j}^p w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_{s;j}}}}{\frac{\sum_{j=1}^p w_j b_j \sigma_{u;T,Z}}{\sigma_Z \sqrt{\phi_u \sum_{j=1}^p w_j^2 b_j^2 + 2\phi_u \sum_{i<j}^p w_i w_j b_i b_j + \sum_{j=1}^p w_j^2 \theta_{u;j}}}} = 1 \quad (10)$$

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

When considered across all k classes, Equations (10) represent a set of $k-1$ non-linear constraints in terms of parameters of the congeneric test Model (1). Hence, the set of restrictions (10) is in fact a necessary and sufficient condition for class-invariant criterion validity of a multicomponent measuring instrument in a population representing a mixture of (k) latent classes. We should like to emphasize that this condition may or may not be fulfilled in a given population. When a (representative) sample from it is available, and for a particular number k of latent classes, the set (10) is testable by employing LVM with a pair of nested models. These are (a) the k -class congeneric model with constraint (10), which is nested in (b) the same model without that constraint (see next section and the illustration section for an example).

Implications for Present-Day Educational, Behavioral, and Social Research

Extending the discussion in Raykov and Marcoulides (2015) for scale reliability in heterogeneous populations, the preceding developments in this article allow us to

make the following important and consequential observation. Irrespective of the facts that (a) the same measuring instrument with the same components $X_1, \dots, X_p$ is used in all classes (studied population) and (b) MI holds across the classes, the instrument need not have the same (criterion) validity in all classes of a given population that is a mixture of them. (The class differences in validity may obviously be even more pronounced if (5) and/or (6) do not hold, in case the “same” construct substantively is still being evaluated with the scale in all of them; see also conclusion section.)

This general lack of class invariance in validity is in our opinion essential to keep in mind in contemporary behavioral and social research that is increasingly concerned with populations characterized by substantial unobserved heterogeneity, a trend that we submit will likely become only more pronounced with time. For such mixture populations, the preceding discussion implies that in general there may be no single meaningful (i.e., no “such thing as”) “validity of a multi-component measuring instrument.” Rather, the empirical reality may actually be that there may be multiple “validities” of a given scale that are class-specific (i.e., subpopulation-specific; the discrepancies across classes in the validity coefficients may be even more pronounced if (5) and/or (6) do not hold in a population of interest, in case 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 criterion validity could or should one in our view refer to “criterion validity of the measuring instrument” in the studied population (and to “validity of the instrument in the population”). Integrated with the similar warnings in Raykov and Marcoulides (2015) regarding scale reliability, we stress that a given multicomponent measuring instrument may in fact function differently with respect to reliability and/or validity from class to class in a population of interest, depending also on the degree of unobserved heterogeneity in the population.

For these reasons we submit that manuals of widely used measuring instruments in educational, social, behavioral, biomedical research, marketing, and business as well as in cognate disciplines are likely to be referring to (criterion) validity 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 (see Raykov & Marcoulides, 2015, for a similar warning with respect to scale reliability). At the same time, 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. We will even argue further that published manual (criterion) validity estimates are in general not necessarily appropriate in present-day empirical research, since they may in fact represent potentially misleading “average” validity estimates that are not valid and of relevance themselves in any of the classes of a (similar) population of interest, in addition to being biased estimates of individual class validities that should instead be

of actual importance. Hence, rather than referring to those manuals and (criterion) validity estimates found there, application of the method in this article could well be recommended to use for evaluation of criterion validity when considering a particular scale for a study of a population possibly representing a mixture of latent classes (see also Raykov & Marcoulides, 2015, for a similar recommendation with respect to scale reliability).

Mixture Factor Modeling as a Statistical Framework for Validity Evaluation in Heterogeneous Populations

Given the relevance of properly accounting for potentially substantial unobserved heterogeneity in studied populations when evaluating instrument (criterion) validity, standard and single-class analytic methods cannot in general be seen as adequate. Instead, latent class analysis is a viable avenue to follow when concerned with psychometric quality evaluation of scales of interest in mixture settings. As discussed at length elsewhere (e.g., Lubke & Muthén, 2005), standard or classic latent class analysis assumes that the observed measures are uncorrelated (independent) within class, following its fundamental assumption of “conditional independence” (e.g., Collins & Lanza, 2013; McCutcheon, 1987). Accordingly, within each class these manifest variables (latent class indicators) are uncorrelated, that is, independent under normality. This is a rather strong assumption that does not in general hold in the setting of concern to this article as well as in much of present-day empirical research. The reason is that, as elaborated earlier, the currently considered setting is based on the congeneric test Model (1) (single-factor model), which represents effectively 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). Indeed, as seen from Equations (1) and in particular (4), the congeneric test model underlying this article does imply that within each of the classes the observed measures are (still) interrelated (see also Note 1). This is due to the presence of within-class common latent variability and covariability in the latent class indicators X_1 through X_p (see Equations 4). This “residual relationship” is readily accommodated with mixture factor analysis (MFA) (e.g., Lubke & Muthén, 2005; Muthén & Muthén, 2014), which is therefore the modeling framework for accomplishing the aims of this article.

Empirical Application of Validity Evaluation Procedure in Heterogeneous Populations

To point and interval estimate criterion validity of a given multicomponent instrument in a mixture setting, first one needs to conduct MFA on the pertinent data collected from a studied population (assuming the available sample is representative of the population, and in particular, of all its latent classes; see Raykov & Marcoulides, 2015). This finite mixture analysis is readily carried out using LVM, and its goal is

to “determine” the number of latent classes, that is, to conduct model selection with respect to number of classes. To this end, the congeneric Model (1) is fitted with 1, 2, 3, and so on, classes in order to carry out model choice regarding class number (e.g., Geiser, 2013).

Once a latent class model with k classes is selected in this manner, the researcher moves on to testing for cross-class identity in criterion validity of the scale under consideration ($k > 1$; see also Raykov, Marcoulides, & Chang, 2016, for a more general discussion on examining population heterogeneity). This is accomplished by testing Equations (10) that as mentioned earlier represent a necessary and sufficient condition for lack of class-differences in validity. The corresponding null hypothesis stipulates then equality of all k class criterion validity coefficients, that is,

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

($c = 1, \dots, k-1$), and is readily tested using LVM with a pair of nested models as indicated above—Model (1) with k classes and constraint (11), which is nested in Model (1) with k classes without that constraint. We reiterate that with k classes this hypothesis testing is equivalent to testing $k-1$ parameter restrictions simultaneously within the selected latent class model.

In case the null hypothesis (11) is rejected, or instead of testing it, a scholar can also point and interval estimate any two classes’ difference in criterion validity. To this end, he/she can point and interval estimate the validity difference $\Delta v_{s,u}$ in Equation (8) for a pair of given classes s and u ($1 \leq s < u \leq k$), which is also readily accomplished using LVM. This is achieved by introducing the difference $\Delta v_{s,u}$ as an “external” model parameter and requesting its interval estimation from the software (see the appendix for the Mplus source code needed then). The resulting confidence intervals for the class differences in criterion validity, which are based on the popular delta-method (e.g., Raykov & Marcoulides, 2004), provide then ranges of plausible values of the corresponding validity class differences in the overall population under investigation (and could also be used for hypothesis testing, in particular of simple or point hypotheses). Moreover, on rejecting hypothesis (11) one can interval estimate the class-specific criterion validity coefficients using their standard errors provided by the software thereby and the monotone transformation-based approach in Raykov and Marcoulides (2011, chap. 4; see pertinent R-function “ci.cv” in the appendix) or the bootstrap approach mentioned above.

Alternatively, when the null hypothesis (11) is not rejected it may be considered as retainable, suggesting that there are no class differences in the criterion validity of the used scale. It is in this case that one can actually speak of (a single, meaningful) criterion validity of the instrument in question, which is of relevance in the studied population. One can then point and interval estimate in a next step that common criterion validity coefficient using the data from all classes, by implementing the constraints (11) in the corresponding fitting of the k -class congeneric model. Subsequently, employing the resulting standard error one can also obtain an approximate 95% confidence interval, say, of this common criterion validity coefficient in

the population in question using the above mentioned R-function “ci.cv” (see the appendix) or the bootstrap approach.²

The Case of Known Class (Group) Membership

The outlined procedure is also straightforwardly employed as a multigroup LVM application when the number of latent classes (groups) and subjects’ class membership are known (e.g., Muthén & Muthén, 2014). Indeed, in this case one can test for group (subpopulation) differences in criterion validity, point and interval estimate this coefficient in each group (subpopulation), and point and interval estimate the validity differences across any two groups (subpopulations) if warranted (see preceding subsection). To this end, formally one uses Equations (7) through (10) with the subindex “*c*” being substituted say with “*g*” for observed group/subpopulation membership; $g = 1, \dots, G$, where G is the known number of studied groups (populations, $G > 1$). That is, when both the number of classes k and the sample individuals’ class membership are known, the method of this article reduces to a LVM (multigroup structural equation modeling) procedure for studying criterion validity differences across a pre-specified set of distinct (sub)populations under consideration. When the null hypothesis of group validity differences is retained, point and interval estimation of the common criterion validity coefficient can be carried out, as mentioned in more detail in the preceding subsection, by implementing the constraint of invariant group validity coefficients when fitting the corresponding G -group (G -class), single-factor model.

The Case of Known Number of Classes

If 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 similarly used in a straightforward manner. Specifically, one fits then only the mixture model for the known number k of classes and uses all remaining activities associated with the procedure outlined in this article. (In other words, one employs the Mplus command files in the appendix for that known number k of classes as well as the R-function “ci.cv” on the so-obtained results; in the appendix, that command file is provided for the case $k = 2$; see also note to it).

Evaluation of Maximal Validity in Heterogeneous Populations

As discussed in detail in the literature, the optimal linear combination (OLC) associated with maximal reliability for the given instrument (linear combination of its components) 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). Hence, when of concern is to point and interval estimate the maximal criterion validity for a scale 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 instead of unitary weights in the sum score, $Y = X_1 + \dots + X_p$, the weights $w_j = b_j/\theta_j$ for the corresponding instrument components, X_j , which produce their OLC, that is, the linear combination

$$Y^* = (b_1/\theta_1)X_1 + (b_2/\theta_2)X_2 + \dots + (b_p/\theta_p)X_p \quad (12)$$

($j = 1, \dots, p$; e.g., Raykov, 2012; see also Equations 4 through 6).

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

Illustration on Data

For the purposes of this section, we use a simulated data set consisting of $n = 1,000$ cases for a scale with of $p = 5$ components and a criterion measure, Z , using for the scale the congeneric Model (1) in each of $k = 2$ classes with class (membership) prevalence of .6 and .4. In the first class, multi-normal data were generated according to the following model (Equations 4):

$$\begin{aligned} X_1 &= T + E_1, \\ X_2 &= 1.5T + E_2, \\ X_3 &= 1.7T + E_3, \\ X_4 &= 2T + E_4, \\ X_5 &= 2.2T + E_5, \end{aligned} \quad (13)$$

where T was standard normal, the error terms $E_1, \dots, E_5$ were zero-mean independent normal variates with standard deviations 1.5, 1.7, 1.4, 1.4, and 1.4, respectively, and the common true score correlation with the criterion Z was .8, while the criterion variable was standard normal. The data in the second class were generated using the same model (13) but with T normal with mean 3 and variance 1, the error terms being independent zero-mean normal variates with standard deviations 1.7, 1.8, 2, 2.1, and 2.3, respectively, and a common true score correlation of .7 with the criterion. (Further details on the simulation procedure can be obtained from the authors on request.)

In the first step of employing the procedure outlined in this article, we fit the congeneric (factor mixture) Model (1) successively with $k = 1, 2$, and 3 classes.² (The pertinent Mplus command file is presented in the appendix; see also Raykov et al., 2016). The resulting BIC indexes as well as p values of the bootstrap likelihood ratio test (BLRT; Nylund, Asparouhov, & Muthén, 2007) are presented in Table 1.

Table 1 suggests to select the $k = 2$ class model based on the BIC and p value of the BLRT, since the BIC is the smallest at $k = 2$ classes and that p value is for the last

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

k	BIC	p-BLRT
1	24555.208	n/a
2	24331.051	.000
3	24353.697	.227

Note. *k* = number of classes; p-BLRT = *p*-value for the bootstrap likelihood ratio test (e.g., Nylund et al., 2007); n/a = not applicable.

The 3-class model was associated with numerical issues, including, parameter boundary estimates, which is consistent with an attempt to over-extract classes (see Equation 13 and immediately following discussion on the data generation process for the two-class model used thereby).

time significant then (as class number increases). This is a correct decision, given that the data were generated using a two-class model.

In the second step of the applying the procedure of this article, we use the selected two-class model to point and interval estimate the possible class difference in criterion validity of the scale under consideration (overall sum score; see second Mplus source code in the appendix). To this end, we introduce the validity difference in Equation (9) as an “external” parameter and request its interval estimation, which also yields standard errors for the two class-specific validity coefficients. The resulting validity estimates were as follows in the two classes (standard errors presented within parentheses):

$$\hat{v}_{Y,1} = .709 (.025) \quad \text{and} \quad \hat{v}_{Y,2} = .602 (.044) \tag{14}$$

Using these estimates and standard errors, the R-function “ci.cv” in the appendix furnishes the 95% confidence interval (CI) for the scale validity in Class 1 as (.513, .684), indicating moderate criterion validity. Similarly, the 95% CI for the scale validity in Class 2 results then as (.658, .755), suggesting also moderate criterion validity. We note that these two CIs overlap arguably to a minor degree, suggesting that the scale criterion validity in Class 2 is considerably higher than that validity in Class 1 (see also next).

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

$$\hat{\Delta}v_{1,2} = .106 (.054) (.001, .212). \tag{15}$$

Because the confidence interval of the validity class difference does not cover 0 (as its left endpoint is above 0), it is suggested that at the .05 significance level the null hypothesis of class identity in criterion validity could be rejected, with the scale criterion validity in Class 1 being considerably higher than this validity in Class 2.

Thereby, it is suggested that a range of plausible population values for the class difference in scale criterion validity, at the 95% CI, stretches from .001 through .212.

Since we know here the (true) parameters of the model having generated the analyzed data, we can use with them the scale validity parametric expression in Equation (4) to find out the true validity coefficients in each of the two classes. Proceeding in this way, we obtain $v_{Y,1} = .744$ and $v_{Y,2} = .618$, that is, the considered instrument has moderate criterion validities in each of the classes, and its class-specific validity coefficients notably differ from each other. These true class validity coefficients are quite close to their estimates reported in Equations (14) and are covered by their 95% CIs presented earlier. In addition, the true difference in class validity coefficients is $\Delta v_{1,2} = .126$, which is similarly quite close to the estimated class difference in scale validity in (15) and is covered by its confidence interval found above.

Conclusion

This article dealt with validity estimation in populations characterized by unobserved heterogeneity and representing mixtures of latent classes. Such populations are already of particular concern in the educational, behavioral, social, and biomedical disciplines as well as in marketing and business research, and/or are becoming of increased interest in them. The outlined LVM approach permits one (a) to examine for identity class-specific criterion validity coefficients, (b) to point and interval estimate their difference as well as these within-class validities, and relatedly (c) to ascertain whether there are between-class differences in scale criterion validity. Moreover, the method can also be used (d) to examine scale validity differences in given distinct populations in case the subjects' class membership is known, and (e) to point and interval estimate these coefficients' difference then (as well as to test various hypotheses about this difference, as in the typical mixture case). When no class differences in criterion validity are found, the procedure similarly permits more precise estimation of the common criterion validity coefficients (see also Note 2). Furthermore, the method is applicable in settings where only the number of latent classes is known but not subjects' membership in them. An extension of the outlined approach, which was also described, can be used when of interest is point and interval estimation of maximal criterion validity based on the set of components constituting a given measuring instrument. Last but not least, the procedure can provide useful information about the latent structure within each class and can address MI issues associated with an instrument under consideration. The method of this article is readily used with the popular LVM program Mplus (Muthén & Muthén, 2014).

Following Raykov and Marcoulides (2015) for the case of scale reliability in heterogeneous populations, we consider as a main message of this article the warning that in general there may or may not be a meaningful single (criterion) validity coefficient of relevance for a given measurement instrument in a population of interest that is a mixture of two or more substantively distinct latent classes, even when one is concerned solely with criterion validity. Whether this is the case indeed and hence

one could talk meaningfully of “criterion validity for an instrument” under consideration, can be addressed with the outlined procedure followed by thorough substantive interpretation of its results in an empirical setting. Specifically we find that one could only then speak (tentatively) of criterion validity of an instrument in 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 validity, that is, the null hypothesis (11) of no class differences in criterion validity is not rejected (the confidence interval(s) of the class difference(s) in validity contain 0).

Similarly to the last cited source, we find it appropriate to caution educational and behavioral researchers with respect also to validity, who are selecting their measurement instruments based on inspection of pertinent manuals provided by their publishers. The reasons are that these manuals may well 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 for subsequent analytic purposes their sampled populations as consisting of a single class, a critical assumption that is in our opinion becoming increasingly likely with time to be violated. Hence, the validity estimates found in the instrument manuals, and in particular criterion validity estimates that are referred to in them, are not unlikely 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 being misleading.

Furthermore, it is worth pointing out here that depending on whether a scholar adopts the MI assumption may have an impact on the number of latent classes “determined” using the procedure within factor mixture modeling (e.g., Geiser, 2013; Lubke & Muthén, 2005; Raykov & Marcoulides, 2015), which is used in the first step of an application of the method of this article. We wish to argue for using as often as possible this assumption (or that of partial MI), which enhances substantially the trust one could have in the stipulation of measuring the same latent construct in all classes (subpopulations) of populations under investigation (e.g., Millsap, 2011; see also Raykov et al., 2012; Raykov, Marcoulides, & Millsap, 2013).

The discussed validity evaluation procedure has several limitations. As described, it is currently applicable with (approximately) continuous scale components (Raykov & Marcoulides, 2015). With up to mild deviations from normality, which do not result from piling at scale end for an individual component(s), we submit that use of the robust ML method is worth considering (MLR; Muthén & Muthén, 2014), perhaps with components having as few as 5 to 7 possible answer options and in particular relatively symmetric 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 also conjecture that the MLR method may as well have some robustness to violations of this classical independence assumption, especially when their extent is limited as is the degree of non-normality of scale

components. Further research is needed, however, before one may place trust in such a potential recommendation as well as in the above for using MLR with discrete (but not highly discrete) instrument components. Moreover, the discussed approach is best used with large samples, since it rests on ML or robust ML estimation, the key methods in mixture analysis at present (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 theory as having obtained practical relevance in a given empirical study. Last but not least, as stated at the outset, the method in this article is based also on the assumption of tenability of the congeneric Model (1) in all classes of a studied population. The plausibility of this assumption is examinable however by comparing the associated BIC index with that of the model not assuming any structure of the relationship of the used latent class indicators (Raykov et al., 2016).³

In conclusion, this article offers to educational, behavioral, social, biomedical, marketing, and business scientists a widely applicable means for criterion validity evaluation in populations with unobserved heterogeneity that are mixtures of latent classes. Together with the scale reliability procedure in Raykov and Marcoulides (2015), the method outlined here permits one to make more informed conclusions about psychometric scale measurement quality in mixture settings, and can be readily used in studies of populations with pronounced unobserved heterogeneity of increasing interest and relevance in contemporary research in these and cognate disciplines.

Appendix

Mplus Source Codes and R-Function for Criterion Validity Evaluation in Heterogeneous Populations

```
TITLE:      MPLUS COMMAND FILE FOR MIXTURE VALIDITY EVALUATION. STEP 1.
            ILLUSTRATION SECTION EXAMPLE. TWO-CLASS MODEL. (SEE NOTE
            BELOW.)

DATA:      FILE = <name of raw data file>;

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

ANALYSIS:  TYPE = MIXTURE;
            STARTS = 500 50;
            LRTSTARTS = 2 1 50 15;

MODEL:     %OVERALL%
            f BY y1-y5;
            [f@0];
            LZ BY Z@1;
            Z@0;
            [lz@0];
            %c#2%
            f; [f];
            lz; [lz];
            y1-y5;

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 “y1-y5;”; the two-class model is preferred in the illustration example).

```

TITLE:      MPLUS COMMAND FILE FOR MIXTURE VALIDITY EVALUATION. STEP 2.
            (TWO-CLASS MODEL.)
DATA:      FILE = <name of raw data file>;
VARIABLE:  NAMES ARE y1-y5 Z;
            CLASSES = c(2);
ANALYSIS:  TYPE = MIXTURE;
            STARTS = 500 50;
MODEL:     %OVERALL%
            f BY y1@1
            y2-y5 (b2-b5);
            y1-y5 (th11-th15);
            [f@0];
            f (fi1);
            LZ BY Z@1;
            Z@0;
            [lz@0];
            lz (varz1);
            f with lz (fiz1);
            %c#2%
            f (fi2); [f];
            lz (varz2); [lz];
            y1-y5 (th21-th25);
            f with lz (fiz2);
MODEL CONSTRAINT:
            NEW(VY1, VY2, RHO1, RHO2, DELTA_VY);
            RHO1 = FI1*(1+B2+B3+B4+B5)**2
                  / (FI1*(1+B2+B3+B4+B5)**2+TH11+TH12
                    +TH13+TH14+TH15);
            RHO2 = FI2*(1+B2+B3+B4+B5)**2
                  / (FI2*(1+B2+B3+B4+B5)**2+TH21+TH22
                    +TH23+TH24+TH25);
            VY1 = FIZ1*SQRT(RHO1) / (sqrt(fi1)*sqrt(varz1));
            VY2 = FIZ2*SQRT(RHO2) / (sqrt(fi2)*sqrt(varz2));
            DELTA_VY = VY1-VY2;
OUTPUT:    CINTERVAL;

```

Note. The class-specific scale reliability coefficients are point (and interval) estimated in the quantities RHO1 and RHO2 with this command file as well (see Raykov & Marcoulides, 2015), which instrumentally utilizes in its MODEL CONSTRAINT section the popular correction for attenuation formula with respect to true and observed criterion scale-score correlations (e.g., Raykov & Marcoulides, 2011).


```

ci.cv=function(r, se) { # R-function for interval estimation of criterion validity
  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 R-function is a trivial adaptation of the function “ci.pc” in Raykov & Marcoulides (2011, chap. 7) 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.cv(v, se),” where for “v” the class-specific validity estimate is entered and for “se” its associated standard error (see pertinent section of software output).

Acknowledgments

We thank B. Muthén for helpful comments on latent class analysis and its applications as well as to an anonymous referee for valuable criticism on an earlier version of the article.

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 defined in Equations (1) to be identified. We also presume in the sequel 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, $Var(T) > 0$ in the notation of Equation (1) ($j = 1, \dots, p$). The assumptions (a) through (c) may be considered hardly restrictive if at all in empirical behavioral and social research, as they could typically be expected to be fulfilled there.
2. The current paragraph in the main text outlines also an approach to point and interval estimation of the common reliability coefficient for a given multicomponent scale under consideration, in case the null hypothesis of lack of class differences in scale reliability is retained. This approach consists of fitting then the congeneric Model (1) to the data from all classes with the imposed set of constraints ensuring class invariance in scale reliability,

and using subsequently the pertinent R-function “ci_cs.rel” in Raykov and Marcoulides (2015; see their Appendix 2) on the resulting common scale reliability point estimate with its associated standard error provided by the software.

3. To examine for possible error covariances, one may proceed as follows. As one avenue, with a plausible model for a given data set that does not contain error covariances, one could argue that there may be insufficient evidence in the data to warrant an assumption of one or more non-zero error covariances in the studied population. Alternatively, one could examine in that (benchmark) model the modification indices and test subsequently for significance the error covariances found to be associated with sufficiently high such indices (e.g., Muthén & Muthén, 2014). As a third feasible approach, one could successively release in that (benchmark) model all possible error covariances, one at a time, and apply the Benjamini–Hochberg multiple testing procedure on the resulting p values for the associated single degree of freedom likelihood ratio tests applied on the resulting pairs of nested models (with the benchmark model being nested in each of its versions with added error covariance; e.g., Raykov, Marcoulides, Lee, & Chang, 2013).

References

- Allen, W. J., & Yen, M. W. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland.
- Apostol, T. (2007). *Calculus*. New York, NY: Wiley.
- 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.
- Collins, L. M., & Lanza, S. T. (2013). *Latent class analysis*. New York, NY: Wiley.
- Crocker, L., & Algina, J. (2006). *Classical and modern test theory*. Boca Raton, FL: H. B. Jovanovich.
- 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.
- 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, L. K., & Muthén, B. (2014). *Mplus user's guide*. Los Angeles, CA: Author.

- 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. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- Raykov, T., & Marcoulides, G. A. (2015). Scale reliability evaluation in heterogeneous populations. *Educational and Psychological Measurement, 75*, 875-892.
- Raykov, T., Marcoulides, G. A., & Chang, C. (2016). Examining population heterogeneity in finite mixture settings using latent variable modeling. *Structural Equation Modeling, 23*, 726-730.
- Raykov, T., Marcoulides, G. A., Lee, C.-L., & Chang, D. C. (2013). Studying differential item functioning via latent variable modeling: A note on a multiple testing procedure. *Educational and Psychological Measurement, 73*, 898-908.
- 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.
- Zimmerman, D. W. (1975). Probability measures, Hilbert spaces, and the axioms of classical test theory. *Psychometrika, 40*, 395-412.