Published in final edited form as:

Struct Equ Modeling. 2017 ; 24(1): 31–51. doi:10.1080/10705511.2016.1238307.

Examining Trait $\times$ Method Interactions Using Mixture Distribution Multitrait–Multimethod Models

Kaylee Litson,

Utah State University

Christian Geiser,

Utah State University

G. Leonard Burns, and

Washington State University

Mateu Servera

University of the Balearic Islands

Multitrait-multimethod (MTMM) analysis allows social scientists to study multiple traits measured by multiple methods and is commonly used to evaluate convergent and discriminant validity of social science measures (Campbell & Fiske, 1959; Eid, 2000; Widaman, 1985). According to Campbell and Fiske (1959), *convergent validity* requires that independent methods of measurement converge (show agreement) when used to measure the same psychological trait(s). *Discriminant validity* is present when measures of different traits (either by the same or different methods) do not show much overlap (i.e., when measures are not highly correlated).

Currently, confirmatory factor analysis (CFA) is one of the most popular methods to analyze MTMM data (Kenny & Kashy, 1992; Eid, Lischetzke, & Nussbeck, 2006). CFA is often preferred to the original MTMM matrix approach (Campbell & Fiske, 1959) because it allows researchers to calculate correlations among latent factors rather than observed variables, thus accounting for measurement error. This provides for a model-based correction for attenuation and makes it easier to compare convergent and discriminant validity coefficients across measures that may differ in their reliabilities. CFA also has more flexibility in terms of how latent trait and method factors can be defined, interpreted, and related to other variables (e.g., Eid, 2000; Marsh & Hocevar, 1988; Widaman, 1985).

Most current CFA-MTMM approaches assume that trait and method factors are uncorrelated and that traits and methods do not interact (exceptions are discussed later in this article). Therefore, most of these approaches implicitly assume that method effects are homogeneous across the entire range of trait values. Moreover, most current approaches to MTMM analysis assume population homogeneity with respect to both convergent and discriminant

Correspondence concerning this article should be addressed to Kaylee Litson, Department of Psychology, Utah State University, Logan, UT 84322. kaylee.litson@gmail.com.

Kaylee Litson, Department of Psychology, Utah State University; Christian Geiser, Department of Psychology, Utah State University; G. Leonard Burns, Department of Psychology, Washington State University; Mateu Servera, Department of Psychology, University of the Balearic Islands.

validity. That is, it is assumed that the measures are equally valid (or invalid) for all individuals in a given population. This may be an unrealistic assumption in many applications. For example, a researcher might be interested in using a CFA-MTMM design to evaluate the convergent and discriminant validity of anxiety and depression scales among children using multiple informants as methods (e.g., mother reports and father reports). It may be the case that mothers and fathers show strong agreement (high convergent validity) for children who exhibit severe depression and/or anxiety symptoms, but show weaker agreement (low convergent validity) for children who exhibit weak or moderate symptom levels, which may not be as obvious to external observers. Moreover, mothers may have difficulty distinguishing between depression and anxiety symptoms for children who exhibit low symptom levels, leading to low discriminant validity for this method at low trait levels. In contrast, mothers may show no difficulty distinguishing between depression and anxiety symptoms if a child exhibits severe symptoms (high discriminant validity at high trait levels).

In the example, one incorrect conclusion drawn from traditional MTMM analyses might be that there is little convergent validity across methods (e.g., mother reports and father reports) when in fact convergent validity may be quite high in a subgroup showing high symptom levels. Another incorrect conclusion drawn from traditional MTMM analyses could be that anxiety and depression are inseparable (lack discriminant validity), when, in fact, they are distinguishable constructs, albeit only in a subgroup of individuals who show severe symptom levels. All of these hypothetical conclusions suggest that method effects may differ across different trait levels. For a variety of psychological constructs commonly examined through MTMM analyses, method effects—and, consequently, convergent and discriminant validity—may not be constant across the entire range of trait values considered, violating the implicit assumption of population homogeneity with respect to method effects. We refer to potential differences in method effects at different trait levels as *trait × method interactions* (i.e., the strength of method effects might vary across trait levels).

It seems desirable for researchers to test the rather restrictive assumption of equal method effects across the entire range of trait values (i.e., the assumption that *trait × method* interactions are absent) made in traditional MTMM analyses by examining population homogeneity versus heterogeneity. When the assumption of population homogeneity is incorrect, researchers may obtain inaccurate or misleading estimates of convergent and discriminant validity for their measures. In the case of population heterogeneity, the resulting single estimates of convergent and discriminant validity obtained from traditional MTMM approaches would represent weighted averages that may apply to only a small set of observations. For the remaining observations, convergent and discriminant validity would be either over- or underestimated.

In the present paper, we describe a new MTMM mixture approach. This new method allows researchers to examine potential differences in convergent and discriminant validity across a priori unknown subpopulations by combining latent factor mixture models (Lubke & Muthén, 2005; McLachlan & Peel, 2000; Muthén, 2001a) with CFA-MTMM models.

Previous Approaches to Study Trait $\times$ Method Interactions

Prior methods for evaluating trait $\times$ method interactions include using multivariate analysis of variance (MANOVA) as well as multivariate group comparison using multigroup CFA (Cole & Maxwell, 1985; Cole, Truglio, & Peeke, 1997). Both MANOVA and multigroup CFA allow researchers to test whether trait and/or method effects differ across subgroups. In contrast to the approach presented in this paper, both MANOVA and multigroup CFA require that the subgroups to be compared be known to the researcher beforehand (e.g., males versus females). As such, neither method can account for potential latent subgroups, that is, subgroup of individuals that are *not* known to the researcher and for which method effects might differ.

Another method that allows studying trait $\times$ method interactions is the direct-product model (Browne, 1984; Wothke & Browne, 1990), which uses a multiplicative approach applied to traditional MTMM matrices. This model evaluates trait $\times$ method interactions by testing whether statistical trait communalities (i.e., the portion of variance explained by trait variance) remain unchanged across methods and whether statistical method communalities (i.e., the portion of variance explained by method variance) remain unchanged across traits, as presented and tested by Bagozzi & Yi (1992). One downside of the direct-product model is that it does not use latent variables and thus confounds true variability with random error, underestimating the true correlations between variables. Further, the model specifies that observed correlations equal trait correlations, which attenuates (or diminishes) the method-specific effect (Eid, 2006). This model is impractical when one considers that method-specific effects are often present and stronger than trait-specific effects for many psychological constructs (e.g., Campbell & Fiske, 1959; Fiske, 1982; Fiske & Campbell, 1992; Podsakoff, MacKenzie, & Podsakoff, 2012).

Previous CFA-MTMM approaches that include latent variables and that allow modeling trait $\times$ method interactions are latent difference-score based approaches presented by Pohl, Steyer, and Kraus (2008) as well as Pohl and Steyer (2010). In Pohl et al.'s approaches, method factors are defined based on latent difference score variables; difference score method factors can be correlated with trait factors. Such correlations can be interpreted in terms of trait $\times$ method interactions. However, Pohl et al.'s approach does not allow specifically examining differences in convergent and discriminant validity in subpopulations, which is the goal of the MTMM mixture approach that we present below.

The MTMM mixture approach is most directly related to two recently presented approaches. Oberski, Hagenaars, and Saris (2015) combined an MTMM design with latent class factor analysis and a survey response model to examine systematic compared to random error in survey questions. Their justification for such an approach was that different surveying methods may contain more or less error for certain groups of participants (e.g., participants are not a homogeneous sample). For example, phone survey responses from individuals who have excellent communication skills may show less measurement error than responses from individuals with moderate or low communication skills. Oberski et al., however, did not use the MTMM design to examine convergent and discriminant validity, which is a common application of MTMM analysis. The potential to examine differences in convergent and

discriminant validity coefficients across trait levels was our reason for creating the MTMM mixture method.

Another related approach is described by Raykov, Marcoulides, and Li (2015) that allows studying differences in criterion validity across subpopulations using mixture distribution methods. Criterion validity is frequently assessed by correlating a test score with a criterion variable. Traditional methods for evaluating criterion validity assume that the test score and criterion variables show the same correlation at all levels (e.g., at low, medium, and high scores), which is a strong assumption. To account for criterion validity differences across different trait levels, Raykov et al. combined a latent class approach with methods designed to measure criterion validity. In the present paper we make a similar argument with regard to convergent and discriminant validity coefficients obtained through MTMM analyses, and we propose a similar solution. When the assumption of homogeneity is violated, because multiple subpopulations exist within the studied population, convergent and discriminant validity coefficients obtained from traditional MTMM analyses could be misleading. Therefore, estimates of convergent and discriminant validity should be allowed to vary across trait levels.

In the present application, we propose to apply mixture distribution methods (e.g., Collins & Lanza, 2010; Lazarsfeld & Henry, 1968; McCutcheon, 1987; McLachlan & Peel, 2000; Vermunt & Magidson, 2002; Lubke & Muthén, 2005) to CFA-MTMM analysis to examine whether a population is heterogeneous with regard to convergent and discriminant validity. Uncovering potential latent subgroups within a sample expands the applicability of MTMM analysis by allowing researchers to study whether the validity of different measurement methods varies across previously unknown subgroups.

Mixture distribution methods allow researchers to identify population heterogeneity (i.e., latent subgroups that are not known before the analysis) instead of separating a sample into multiple groups based on ad hoc criteria (e.g., quartiles of observed scores). Mixture distribution methods have been successfully applied in many areas of structural equation modeling, for example, to conventional (single-method) CFA models (Lubke & Muthén, 2005), latent state-trait analysis (Courvoisier, Eid, & Nussbeck, 2007), and longitudinal growth models (Muthén & Muthén, 2000). Furthermore, Raykov et al. (2015) successfully applied mixture models to study heterogeneity with regard to criterion validity, and Oberski et al. (2015) applied latent class models to an MTMM framework to study heterogeneity with regard to error in survey responses across methods. In the present paper, we combine mixture distribution methods with a CFA-MTMM framework to study heterogeneity with regard to convergent and discriminant validity.

Combining CFA-MTMM and Factor Mixture Analysis

The starting point for our MTMM mixture model is a single-class, multiple-indicator CFA-MTMM model with a separate common factor (true score variable) for each trait-method unit (TMU) and a congeneric measurement structure (Marsh & Hocevar, 1988; see Figure 1A). In this model, each observed variable (Y_{ijk}) is decomposed into a constant intercept (α_{ijk}), a latent true score variable (τ_{jk}), and a measurement error variable (ϵ_{ijk}):

$$Y_{ijk} = \alpha_{ijk} + \lambda_{ijk}\tau_{jk} + \varepsilon_{ijk}, \quad (1)$$

where i indicates the i th indicator, j indicates the j th trait, k indicates the k th method, and λ_{ijk} is a constant factor loading parameter. The measurement error variables ε_{ijk} have means of zero and are uncorrelated with all true score variables τ_{jk} as well as with all other error variables. The model can be identified by fixing the intercept and loading for one indicator per TMU to 0 and 1, respectively. Figure 1A illustrates the model for two traits, two methods, and three observed variables (indicators) per TMU.

The model in Equation 1 can be seen as a latent variable representation of the original Campbell and Fiske (1959) MTMM matrix. By using multiple indicators, true score variance can be separated from measurement error variance within each TMU. The common true score variables τ_{jk} represent the true score (error free) portion of variability within a given TMU (e.g., true depression score variability measured by self-report), whereas ε_{ijk} represents random measurement error.

Correlations between the τ_{jk} factors across TMUs can be used to study convergent and discriminant validity in line with Campbell and Fiske's (1959) original guidelines. In contrast to the correlations in Campbell and Fiske's (1959) original MTMM matrix, the correlations between τ_{jk} factors are corrected for measurement error (unreliability) and are thus less biased and easier to compare across different TMUs. Following Campbell and Fiske (1959), three general types of correlations can be assessed in this analysis: 1) correlations between different methods measuring the same trait (monotrait-heteromethod correlations; these indicate convergent validity), 2) correlations between the same method measuring different traits (heterotrait-monomethod correlations), and 3) correlations between different methods measuring different traits (heterotrait-heteromethod correlations). Based on comparisons between different types of MTMM correlations, researchers can examine convergent and discriminant validity.¹

One complication with the model in Figure 1A can arise in practical applications when the same (or similar) items or scales are used across methods (e.g., mother and father reports). For example, mothers (Method 1) may be asked to report on the same symptoms of depression as fathers (Method 2). In this case, observed variables that refer to the same items or symptoms may be more highly correlated across methods or informants than observed variables that reflect different items. Such effects are well-known from longitudinal research in which the same indicators tend to be more highly correlated with themselves across time than with other indicators. Similar effects can occur in multi-indicator MTMM data across methods.

The model in Figure 1A does not adequately account for variable- or indicator-specific effects that are shared across methods. As a consequence, the model may not fit well when different methods use similar items or scales, potentially resulting in biased parameter

¹For more detailed guidelines for evaluating convergent and discriminant validity based on MTMM correlations, see Campbell and Fiske (1959).

estimates. This could lead to incorrect estimates of latent correlations (i.e., of convergent and discriminant validity coefficients). This problem can be addressed by adding so-called indicator-specific factors to the model (Geiser, Burns, and Servera, 2014), as shown in Figure 1B.² Indicator-specific factors are residual factors that characterize the deviation of a given indicator from the values that are expected based on a linear regression on the reference indicator. For example, the item “My/this child is depressed” may be chosen as reference indicator for measuring depression through mother and father reports, respectively. A second item, say, “My/this child is withdrawn” may serve as a non-reference item. The indicator-specific residual factor in this case reflects residual variance in the second item (withdrawn) after the variance that *withdrawn* shares with the reference item (depressed) is partialled out. When using this approach, a marker item or indicator (i.e., an indicator that represents the intended construct especially well, either theoretically or empirically) is chosen as reference indicator.³

For the formal description of the extended model and without loss of generality, we select the first indicator ($i = 1$) as reference (marker) indicator. The measurement equations for the CFA-MTMM model with indicator-specific factors are then given by:

$$Y_{1jk} = \tau_{1jk} + \varepsilon_{1jk}, \text{ for } i = 1. \quad (2)$$

$$Y_{ijk} = \alpha_{ijk} + \lambda_{ijk}\tau_{1jk} + \delta_{ijk}\zeta_{ij} + \varepsilon_{ijk}, \text{ for } i \neq 1. \quad (3)$$

Note that the latent variable τ_{1jk} is now specific to the reference Indicator 1 (i.e., represents that variable's true score variable). This can be seen from the fact that the intercept and factor loading for the reference indicator are fixed to 0 and 1, respectively. The parameter δ_{ijk} indicates a constant indicator-specific factor loading. For identification, one δ_{ijk} parameter is fixed to one per indicator-specific factor. The indicator-specific factors (ζ_{ij}) have a mean of zero and are uncorrelated with all true score variables that pertain to the same trait j as well as with all error variables. The indicator-specific factors capture the systematic indicator-specific residual variance in the non-reference indicators that is shared across methods for a given trait. The measurement error variables ε_{ijk} have means of zero and are uncorrelated with all true score variables τ_{1jk} , all indicator-specific factors ζ_{ij} as well as with all other error variables. The indicator-specific model is illustrated in Figure 1B.

A limitation of both previously discussed models is that they do not adequately account for potential population heterogeneity with regard to convergent and discriminant validity. We now present an extension of the single-class CFA-MTMM models described above to factor

²Another way to address this issue is to allow the error variables pertaining to the same indicators across methods to correlate. However, this would lead to an underestimation of indicator reliabilities and generally more highly parameterized models, which is why we did not use this approach in the present paper.

³Reference indicators may be chosen at random if indicators are essentially homogeneous. If indicators are heterogeneous (e.g., indicators represent item subscales or different items with substantively different content), then the reference indicator should be chosen using substantive theory and all results should be interpreted in relation to the reference indicator (similar to dummy coding in regression analysis).

mixture models with multiple latent classes. Unlike the single-class models, the MTMM mixture model specifies multiple latent classes, where c indicates the c th class. In principle, any parameter of the model can differ across classes. For the non-indicator-specific CFA-MTMM mixture model, the measurement equation is given as:

$$Y_{ijkc} = \alpha_{ijkc} + \lambda_{ijkc}\tau_{jkc} + \varepsilon_{ijkc}. \quad (4)$$

For the indicator-specific CFA-MTMM mixture model, the measurement equations are given as:

$$Y_{1jkc} = \tau_{1jkc} + \varepsilon_{1jkc}, \text{ for } i=1 \quad (5)$$

$$Y_{ijkc} = \alpha_{ijkc} + \lambda_{ijkc}\tau_{1jkc} + \delta_{ijkc}\zeta_{ijc} + \varepsilon_{ijkc}, \text{ for } i \neq 1. \quad (6)$$

Of particular interest in the CFA-MTMM mixture models are potential mean differences on the true score variables between classes as well as whether correlations between the true score variables that represent the different TMUs differ across classes. Such correlation differences would indicate that convergent and/or discriminant validity coefficients differ across latent classes, pointing to the presence of trait-method interactions. If there are no such differences between classes (or if a single class model fits better than multi-class models), this would indicate that convergent and discriminant validities apply in the same way to the entire population, which is an implicit assumption made in most traditional MTMM approaches. The CFA-MTMM mixture models thus allow testing this assumption in a formal way.

An important aspect of our proposed mixture CFA-MTMM approach is to determine measurement invariance (MI) within and across latent classes. If certain measurement parameters are invariant (or at least partially invariant) within and/or across classes as described below, one can meaningfully test whether structural (latent variable) parameters (e.g., latent means) differ across methods and classes (Byrne, Shavelson, & Muthén, 1989; Schoot, Lugtig, & Hox, 2012; Widaman & Reise, 1997).

Testing MI across methods can be useful when equivalent items or scales were used by different methods in a study (e.g., when multiple reporters were used as methods and they filled out questionnaires with equivalent items). Researchers can use MI tests to examine the extent to which measures behaved similarly across methods and whether different methods resulted in similar or different latent means, variances, and correlations (Geiser et al., 2014).

In the context of the MTMM mixture model, the MI testing procedure begins by assuming configural MI *across methods*, then testing weak, strong, and strict MI across methods. Next, to understand which structural (latent variable) parameters may be different *across classes*, MI should be tested across classes. Strong MI across classes is typically desirable in order to meaningfully evaluate latent mean differences across classes. To illustrate the mixture CFA-

MTMM model, we now present an application to a sample of children who were rated by mothers and fathers on attention deficit hyperactivity disorder (ADHD) symptoms of inattention (IN) and hyperactivity/impulsivity (HI).

Empirical Illustration

To keep our illustration relatively simple, we chose an example with just two traits (IN and HI) and two methods (mother and father reports). This 2 trait by 2 method design allows us to illustrate all relevant features of the mixture CFA-MTMM approach, including the comparison of both convergent and discriminant validity coefficients across classes. We describe a step-by-step modeling approach that first establishes well-fitting single-class, single-trait baseline models (and tests the degree of MI across methods) before moving on to more complex multi-class, multi-trait models. Beginning the analysis with smaller (sub)models has many practical advantages, as factor mixture analyses can be very complex, especially when multiple traits and multiple methods are included. We found that the step-by-step approach described here facilitated the identification of the final model considerably. We propose to compare multiple models based on chi-square difference tests in the single-class modeling step as well as information criteria such as the Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC) to compare subsequent multi-class models. All analyses were carried out using Mplus version 7.11. Appendix A provides the specific Mplus syntax specification for the final model for which we report detailed outcomes.

Participants, Procedures, & Measures

The data used in the empirical example comes from a larger dataset used to evaluate the progression of ADHD symptoms among $N = 1,045$ first-grade children across 22 schools in the Balearic Islands, Spain, and eight additional schools from Madrid, Spain, totaling 30 elementary schools.⁴ The study protocol was approved by the University of the Balearic Islands Institutional Review Board. Each child in the study was evaluated at four measurement occasions by mothers, fathers, teachers, and aides using the Child and Adolescent Disruptive Behavior Inventory (CADBI, Burns & Lee, 2011). Symptom levels were evaluated in the home by parents using a 6-point scale (i.e., $0 = \text{nearly occurs none of the time (e.g., 2 or fewer times per month)}$, $1 = \text{seldom occurs (e.g., once per week)}$, $2 = \text{sometimes occurs (e.g., a few times per week)}$, $3 = \text{often occurs (e.g., once per day)}$, $4 = \text{very often occurs (e.g., several times per day)}$, and $5 = \text{nearly occurs all the time (e.g., many times per day)}$). Items for both HI and IN are relevant to ADHD symptoms found in the DSM-5.

Items were averaged to three parcels: each containing three items and each ranging from 0 – 5. Using parcels as indicators requires *a priori* examination of item-level data to ensure that the items constitute a unidimensional scale (Bandalos, 2008; Little et al., 2013; Marsh et al.,

⁴Given that the dataset included multiple schools, we examined whether cluster effects were present by examining intraclass correlation coefficients (ICC). All ICCs were relatively small in size (mean ICC = .04, range = .02 to .05). Because all ICCs were rather small and given that the present data serve illustrative purposes only, we did not include cluster effects in the analyses. Substantive researchers should take appropriate steps to control for cluster effects in multilevel datasets by using appropriate statistical techniques (e.g., type = complex in Mplus, or multilevel modeling methods).

2013). Detailed item-level analyses were conducted for the current data set, and results supported the unidimensionality of the parcels. Structural parameter results were essentially equivalent for models using items as indicators compared to parcels as indicators (detailed item-level analyses are reported in Burns et al., 2014). In the present application, we used mother and father ratings of ADHD-HI and ADHD-IN at wave 3 when children were in second-grade, one year from the initial assessment ($n = 618$).

Data from the present study have previously been used to examine convergent and discriminant validity of ADHD (Burns et al., 2014), and have shown strong levels of convergent and discriminant validity between mother and father ratings of ADHD-HI and ADHD-IN (latent correlations between mother and father ratings range from .60 to .83, with highest correlations across mothers and fathers measuring the same trait). Further, data used in the present study have been used to examine MI across mothers', fathers', and teachers' ratings of ADHD symptoms, and have shown strong MI (Geiser et al., 2014), which is adequate to examine mean differences. The novelty of the current application is that the assumption of sample homogeneity is relaxed, allowing latent means and convergent/discriminant validity coefficients to be examined separately across latent classes.

The present MTMM data set is well-suited to illustrate our new MTMM mixture modeling approach because it is known from the previous substantive literature that a subgroup of approximately 6 – 10% of children have clinically significant levels of ADHD (American Psychiatric Association [APA], 2013; Visser et al., 2014). The proposed MTMM mixture approach allows us to examine potential validity differences across subgroups with clinically relevant versus normative symptom levels (and potentially other unknown subgroups). That is, convergent and discriminant validity might differ in children with clinically relevant symptom levels as compared to children with low or moderate ADHD symptom levels.

Modeling Approach

Given the complexity of the MTMM mixture modeling approach and the many possible alternative models that could be fit, we developed a three-step model-fitting procedure to identify the most appropriate model for the data. The three-step procedure involves several logical steps to test and rule out alternative models, starting with the simplest possible single-class, single-trait models (Step 1). Step 2 tests for the number of classes needed to properly account for population heterogeneity. Step 3 combines the single-trait models into multi-trait models. To illustrate the three-step procedure, we provide a decision tree (Figure 2) with all possible decisions that could have been made in the modeling approach and describe each step in detail below.

Step 1—In Step 1, our goal was to evaluate single-class, single-trait, multi-method (STMM) models for each trait (IN and HI), separately. First, we evaluated a parsimonious STMM model without indicator-specific effects and with configural MI across methods (Figure 1A). If the model without indicator-specific (IS) factors fit well, then we would evaluate even more parsimonious versions of this model that included MI constraints across methods. If, on the contrary, the model without IS factors did not fit well, then we would evaluate a (less parsimonious) model with additional IS factors to account for IS effects

(Figure 1B; i.e., ζ_{ik} in Equation 3). Subsequent steps included evaluating different degrees of MI (configural, weak, strong, and strict) across methods within the same trait to identify the most parsimonious baseline model for subsequent analyses with multiple latent classes.

Step 2—In Step 2, our goal was to evaluate multi-class STMM models for each trait separately, and to examine variance and mean equality within traits across classes. We used the best-fitting STMM models from Step 1 and tested whether and how many latent classes were needed to adequately account for population heterogeneity. Furthermore, we conducted tests of MI across classes for each trait separately. MI is a necessary requirement for interpreting mean differences, and is thus important for interpreting trait $\times$ method interactions. Since MI and class enumeration are both evaluated by examining differences in model fit, MI and number of classes are inherently confounded in the MTMM mixture approach. Therefore, to properly determine the number of classes and level of MI, all levels of MI must be evaluated for all 2-, 3-, 4-, etc. class models. We stopped evaluating additional classes when an entire set of models (e.g., all 4-class models) fit worse than the previously best-fitting model with one less class (e.g., the best-fitting 3-class model).

After the number of classes and level of MI were determined, we evaluated structural (latent variable) parameter MI (i.e., latent means, variances, and correlations) across classes for each construct separately. The first model evaluated equal factor variances across c classes. If such a model did not fit, then a second model would evaluate equal factor variances across $c - 1$ classes, and so on. This step was important to determine potential differences in the latent means, variances, and covariances/correlations across classes for each trait. With these models, we could determine whether means and monotrait-heteromethod correlations (convergent validities) differed across latent classes.

Step 3—The final step in our MTMM mixture analysis was to combine the best fitting multiclass STMM models into an appropriate two-trait MTMM mixture model and to evaluate correlation and covariance equality across classes. This step determined if differences in convergent and discriminant validity were present across classes. In this step, we also examined whether the combination of multiple traits into a single model made it necessary to extract more classes than extracted in the previous STMM models.

All models were fit using robust maximum likelihood estimation in Mplus version 7.11 (Muthén & Muthén, 1998–2012). We used at least 5,000 sets of start values in an attempt to replicate the overall best loglikelihood value for each model. In cases in which the best loglikelihood value could not be replicated at least once, we increased the number of sets of start values up to a maximum of 100,000 (for single-trait models) and 200,000 (for multitrait models) until replication of the best loglikelihood value was achieved.⁵ In cases in which replication of the best loglikelihood value could not be achieved even with 100,000 (for single-trait models) and 200,000 (for multitrait models) sets of start values, we assumed that the corresponding model may not be well-defined for the present data. Cases in which this problem occurred are mentioned in the results section. Note that this problem did not occur

⁵The exact number of sets of start values used can be found in the table notes of Tables 2, 3, and 4.

for any of the models that we accepted as final models and for which detailed results are presented below.

Results

Step 1: Single-class single-trait baseline models

In this step, our goal was to identify well-fitting single-class, STMM baseline models separately for HI and IN. Table 1 shows the fit statistics for different models. The most basic STMM model tested was a model of configural MI without IS factors (Model 1 in Table 1). In the present example, this model did not fit well. Therefore, we evaluated a model with IS factors (see Figure 2 for the comprehensive stepwise modeling procedure). As seen in Table 1, a configural MI model with IS factors (Model 2) showed a much better fit for both HI and IN than Model 1. For both constructs, the IS factors had substantial variances and standardized loadings $.31$ ($p < .05$), showing that the indicators were not perfectly homogenous and that indicator-specificity was shared to some extent across mother and father reports. Because Model 2 fit significantly better than Model 1 and because the IS factors explained a substantial amount of variance, we included IS factors in all subsequent models.

In Step 1, we also tested for MI and structural parameter equality across methods. As shown in Table 1, a model of strict MI (Model 5) fit well for both HI and IN based on chi-square tests. Model 5 also showed the lowest AIC and BIC coefficients for both HI and IN. Parameter estimates obtained from Model 5 showed that mother and father ratings were significantly correlated for both HI ($.81$) and IN ($.77$), indicating strong levels of convergent validity for both constructs. For HI, results indicated relatively low means (mothers: $M = 1.11$, fathers: $M = 1.08$ compared to a possible range from 0 to 5), high true score variances (mothers: $var = 1.11$, fathers: $var = 1.00$), much lower yet significant IS variances (range: $[0.11; 0.15]$), and even lower error variances (range: $[.08; .10]$), indicating that most of the variability in the indicators reflected shared (common) true score variance. For IN, similar results were found with regard to means (mothers: $M = 0.96$, fathers: $M = 1.01$), factor variances (mothers: $var = 0.73$, fathers: $var = 0.79$), IS variances (range: $[0.10; 0.12]$), and error variances (range: $[.07; .09]$). Given its good model fit, we used the strict MI Model 5 as the basis for all multi-class models (that is, in the following steps, we assumed strict MI across mother and father ratings within each class in all multi-class models).

Step 2: Determine the number of classes, test for MI across classes, and test for structural parameter MI in multi-class STMM models

Tables 2 and 3 give an overview of the models evaluated in this step as well as their relative model fit in terms of AIC and BIC coefficients. Each construct was evaluated separately. Since MI is crucial to interpret mean differences across classes, we examined MI across classes first.

Two-class models—Models 6a and 6b were the 2-class versions of Model 5 with configural MI across the two latent classes. According to the AIC and BIC values, adding a second class improved the model fit compared to the single-class Model 5 for both HI and

IN. This indicated the presence of population heterogeneity in the data (i.e., it made sense to consider more than just one class). Therefore, we continued with the two-class models and tested for stricter forms of MI across the two classes. The best-fitting 2-class HI solution was one in which the factor loadings were non-invariant across the two classes. For IN, the best-fitting 2-class solution was one with strong MI across the two classes.

Three-class models—To examine whether more than two classes were needed, we examined solutions with an additional latent class. As can be seen from Tables 2 and 3, the 3-class solutions generally fit better than the two-class solutions. We first fit a model with configural MI across three classes, and then tested for MI across three classes. Of all three-class models considered for HI, the best-fitting solution was a three-class model with weak MI and equal error variances across two of the three classes (Model 14a in Table 2). Of all three-class models considered for IN, the best-fitting solution was a three-class model with strong MI across three classes and equal error variances across two classes (Model 16b in Table 3).

Four- or more class models—After determining that a 3-class solution fit better than a 2-class solution for both HI and IN, we continued with 4-class models. As with the 2- and 3-class models, we evaluated MI across the four classes for each construct separately. For HI, the best-fitting solution was a 4-class model with strong MI across all classes and equal error variances across three classes (Model 19a in Table 2).⁶ We also examined 5-class models for HI. None of the examined 5-class solutions fit better than previous 3- or 4-class solutions for this construct. The four-class solution fit the data better than 2, 3, and 5-class solutions for HI. For IN, none of the examined 4-class models fit better than the 3-class Model 16b, so we did not examine IN models with more than four classes. In summary, the AIC and BIC coefficients supported a 3-class solution for IN and a 4-class solution for HI, both with strong MI and partial strict MI across classes.

Test for structural parameter MI across classes—The final part of Step 2 was to determine whether structural parameters differed across latent classes for HI and IN in the best-fitting model. Since at least strong MI was present across all classes for both constructs, means could be meaningfully compared across classes, and trait $\times$ method interactions were considered.

Evaluating variance equality: For both HI and IN, all classes had equal factor loadings, so it was appropriate to test for the equality of factor variances across all classes. For HI, results showed that true score variances were equal across two of the four classes (Model 22a). For IN, results showed that true score variances were not equal across any of the three classes. For both HI and IN, results showed that IS factor variances were not equal across any of the classes.

⁶The best loglikelihood values for some four-class models for HI could not be replicated even with 100,000 sets of start values. Models for which this problem occurred are marked in Table 2. The model fit results for these models should be interpreted with caution, and no estimated results from these models are discussed in the text. This did not concern the best-fitting HI model (Model 22a), for which detailed outcomes are reported in the text and for which the best loglikelihood value could be replicated. Given the large difference in AIC and BIC values between Model 22a and the other 4-class models, we think that our decision to accept Model 22a as the best model for HI was sensible.

Evaluating mean equality: Intercepts and factor loadings were equal across classes for both HI and IN (strong MI), indicating that means could be meaningfully compared across classes. Results showed that means differed across classes for both HI and IN.

Evaluating correlation equality: For HI, results showed that correlations were equal across classes with the highest and lowest mean levels; further, correlations were equal across the two classes with moderate mean levels. Otherwise, correlations differed across classes. For IN, correlations differed across classes.

Best-fitting single-trait multimethod multi-class models—For HI, the best-fitting multi-class model contained four classes with strong MI across all classes, equal error variances across three classes, and equal true score variances across two classes (Model 22a in Table 2). In Model 22a, the average class assignment probabilities for the most likely latent class were relatively high (range: [.77; .94]), indicating that individuals were reliably classified in this solution. Class 1 (7%) had high mean levels (mothers: $M = 2.57$, fathers: $M = 2.55$), moderate true score variances (mothers: $var = 0.37$, fathers: $var = 0.37$), high IS variances (range: [.42; .59]), and a perfect correlation between the true scores for mother and father ratings ($\phi = 1.03^7$, 95% CI: [1.00, 1.05]), indicating perfect convergent validity in this class. Class 2 (35%) had low mean levels (mothers: $M = 0.85$, fathers: $M = 0.86$), moderate true score variances (mothers: $var = 0.37$, fathers: $var = 0.37$), low IS variances (range: [.06; .10]), and a high correlation between mother and father ratings ($\phi = .71$).

Class 3 (24%) had the lowest mean levels (mothers: $M = 0.19$, fathers: $M = 0.18$), very small true score variances (mothers: $var = 0.02$, fathers: $var = 0.01$), small IS variances (range: [0.00; 0.03]), and an estimated true score correlation between mother and father ratings of $\phi = 1.35$. Estimates of latent correlations above 1.0 in confirmatory factor models represent improper estimates, as they are outside of the range of possible values. Improper estimates may be due to sampling fluctuations or model misspecification such as an overfactorization (Chen et al., 2001; van Driel, 1978). In the present study, we examined whether improper correlation estimates differed significantly from 1.0 by computing their 95% confidence intervals (CIs) as suggested by Chen et al. (2001) and van Driel (1978). The above correlation estimate did not differ significantly from 1.0, 95% CI: [.81, 1.89]. For this reason, we believe that the correlation estimate may have simply reflected a perfect positive association (perfect convergent validity) in Class 3 in the population.⁸ Moreover, the close-to-zero latent variances in this class may have caused some instability in the estimation of the latent correlation as indicated by its large standard error and a wide CI. We discuss the issue of improper correlation estimates in MTMM mixture models in more detail in the Discussion section. Class 4 (34%) had moderate mean levels (mothers: $M = 1.75$, fathers: M

⁷Some estimated correlations in the final solutions exceeded 1.0, but none was statistically significantly larger than 1.0 (as shown by reporting the 95% confidence interval estimates, recommended by Chen et al. [2001] and van Driel [1978]). Correlation estimates exceeding 1.0 that do not differ significantly from 1.0 are often due to sampling fluctuations (van Driel, 1978). We therefore assumed that the corresponding estimates represented perfect correlations between the corresponding factors. However, sometimes improper solutions may be due to model misspecifications (Chen et al., 2001) or small sample sizes (Boomsma, 1985). A more in depth analysis of improper solutions encountered in the present study are discussed in the Discussion section.

⁸Two estimates exceeded 1.0 by a relatively large amount ($\phi = 1.35$ and $\phi = 1.39$). In both of these cases, the variances for variables were close to zero, and neither estimate was statistically significantly different from 1.0. We discuss lack of variability within classes and its influence on improper correlation estimates more thoroughly in the Discussion section.

= 1.67), high true score variances (mothers: $var = 1.28$, fathers: $var = 0.97$), moderate IS variances (range: [.12; .21]), and the lowest correlation between mother and father ratings ($\phi = .57$).

Correlations were significantly different from one another when comparing Classes 1 with 2 ($\phi_{diff} = .32$; $p < .05$), 1 with 4 ($\phi_{diff} = .46$; $p < .05$), 2 with 3 ($\phi_{diff} = .64$; $p < .05$), and 3 with 4 ($\phi_{diff} = .78$; $p < .05$). Correlations were not significantly different from one another when comparing Classes 1 with 3 ($\phi_{diff} = .32$; $p = .23$) or Classes 2 with 4 ($\phi_{diff} = .14$; $p = .29$). These results show that there is a trait $\times$ method interaction. Classes 1 and 3 share the same level of convergent validity and Classes 2 and 4 share the same level of convergent validity. Classes 1 and 3 represent the extreme groups, with the highest and lowest mean levels and perfect correlations (i.e., perfect convergent validity) between mother and father ratings. The moderate trait value groups (Classes 2 and 4) also shared the same level of convergent validity which was lower than convergent validity in Classes 1 and 3.

For IN, the best-fitting model was a 3-class solution with strong MI across all classes, equal error variances across two classes, and no equality constraints on structural (latent variable) parameters across classes (Model 16b in Table 3). Average class assignment probabilities for the most likely latent class were high (range: [.86; .94]), indicating that individuals were reliably classified in this solution. Class 1 (13%) had the lowest mean levels (mothers: $M = 0.04$, fathers: $M = 0.03$), a true score variance of zero, and a very small IS variance (range: [0.00; 0.04]). Given that the true score variance was estimated to be zero in this class, no correlation between mother and father ratings could be computed. We suspect that Class 1 represented a class of highly homogenous, symptom-free individuals. Class 2 (48%) had low mean levels (mothers: $M = 0.67$, fathers: $M = 0.69$), moderate true score variances (mothers: $var = 0.20$, fathers: $var = 0.21$), low IS variances (range: [0.07; 0.08]), and a moderate correlation between mother and father ratings ($\phi = .58$). Class 3 (39%) had the highest mean levels (mothers: $M = 1.63$, fathers: $M = 1.74$), highest true score variances (mothers: $var = .80$, fathers: $var = .78$), moderate IS variances (range: [.16; .22]), and a moderate correlation between mother and father ratings ($\phi = .59$). Classes 2 and 3 showed virtually identical convergent validities with a $\phi_{diff} = .03$, $p = .84$.

Step 3: Combine the best-fitting single-trait models to create an MTMM mixture model

The best-fitting single-trait multi-class models from Step 2 were combined to analyze both constructs (HI, IN) simultaneously. A simultaneous analysis enables researchers to also examine potential differences in discriminant validity coefficients across classes (in addition to differences with respect to convergent validities). Given that the best-fitting solution for HI was a 4-class model, whereas the best-fitting solution for IN was a 3-class model, we estimated MTMM mixture models using 3-classes (Model 29), 4-classes (Model 30), and 5-classes (Model 31). We include the 5-class models to test whether combining the two single-trait models into one multi-trait model required an extraction of additional latent classes. Model fit indices revealed that a 4-class model (Model 30; BIC = 10036; AIC = 9469) fit better than a 3-class model (Model 29; BIC = 10250; AIC = 9820) and a 5-class model (Model 31; BIC = 10271; AIC = 9568). Average class assignment probabilities for the most

likely latent class were high for Model 30 (range: [.88; .95]), indicating that individuals were reliably classified in this solution.

Testing covariance/correlation equality across classes—We evaluated the equality of all six convergent and discriminant validity correlations across the four classes.

Comparing a single correlation across four classes requires six tests. Therefore, a total of 36 equality constraints were evaluated. Table 4 shows the different pairs of correlations that were evaluated for statistical differences between classes. Results indicate that most latent correlations were not significantly different across classes; those that were significantly different were monotrait-heteromethod correlations (convergent validity coefficients).

Evaluating trait $\times$ method interactions—Based on fit indices, we determined that the best fitting combined model was Model 30, a four-class MTMM mixture model with strong MI across classes (model entropy = 0.85). A path diagram for Model 30 with the final model parameter estimates is presented in Figure 3. The results indicated the presence of a “no symptoms” class (Class 1), a “no HI/low IN symptoms” class (Class 2), a “low symptoms” class (Class 3), and a “moderate symptoms” class (Class 4).

The no symptoms class (Class 1, 18.5%) had very low means and variances for both HI and IN, and showed perfect convergent validity as indicated by monotrait-heteromethod correlations of $\phi = 1.39$ for HI (95% CI: [.86, 1.92]) and $\phi = 1.0$ for IN (95% CI: [.67, 1.35]). Again, the latent correlations did not differ significantly from 1.0 as shown by their confidence intervals, indicating that these estimates might simply reflect sampling fluctuations of underlying perfect correlations at the population level and/or some instability in the estimation due to small latent variances within class. The heterotrait-monomethod and heterotrait-heteromethod correlations were much smaller and nonsignificant (ϕ range: [.21; .76]). The no HI/low IN symptoms class (Class 2, 8.7%) had very low means for HI and low means for IN. Interestingly, in line with Class 1, this class showed high monotrait-heteromethod correlations for HI ($\phi = .89$), but not for IN ($\phi = .31$), and showed similar heterotrait-monomethod and heterotrait-heteromethod correlations as Class 1 (ϕ range: [−.11; .89]). The low symptoms class (Class 3, 29.8%) had low means for both HI and IN. Class 3 also showed evidence for convergent validity with high monotrait-heteromethod correlations (ϕ range: [.76; .85]) and had relatively low heterotrait-monomethod and heterotrait-heteromethod correlations (ϕ range: [.38; .49]). The moderate symptoms class (Class 4, 43%) was the largest class, and showed moderate mean levels for both HI and IN and larger variances for both HI and IN than the previous classes. Furthermore, Class 4 showed high monotrait-heteromethod correlations (ϕ range: [.61; .70]) and low heterotrait-monomethod and heterotrait-heteromethod correlations (ϕ range: [.22; .48]).

Comparisons of latent correlations across classes (see Table 4 for the detailed comparisons) revealed that four monotrait-heteromethod correlations differed significantly across classes. The correlation between IN mother with IN father reports as well as between HI mother with HI father reports differed across Classes 1 and 4. The correlations between IN mother and IN father reports also differed across Classes 1 and 2 and Classes 2 and 3. None of the heterotrait-monomethod or heterotrait-heteromethod correlations differed across classes. In summary, the results revealed the presence of a trait-method interaction: We found that the

convergent validity of ADHD symptom reports between mothers and fathers was larger for target children with no or very low symptoms as opposed to children with moderate symptoms.

Discussion

Mixture distribution analyses are useful tools to test whether multiple distinct subpopulations are present within a sample or population. The current paper presented a new MTMM mixture approach that allows testing 1) whether a sample is homogeneous or consists of multiple subpopulations, 2) whether means, variances, and covariances/correlations are equal across latent classes, and 3) whether convergent and discriminant validity coefficients differ across classes. We applied the new method to a dataset using mother's and father's ratings of ADHD symptoms, and found that four latent classes were needed to account for population heterogeneity. We found that four convergent validity correlations differed across latent classes, indicating the presence of a trait $\times$ method interaction. In the present application, this interaction showed that convergent validity was higher for individuals who showed no or very few ADHD symptoms as compared to low or moderate symptom groups. A reason for this could be that individuals who show no symptoms at any time are easy to rate as "symptom free" by multiple reporters. In contrast, individuals who show some symptoms sometimes (or in some contexts but not others) may not be consistently rated by different reporters (i.e., only one of the two reporters may "catch" the occurrence of symptoms, leading to lower correlations among raters).

In the presented application, results found four classes with distinct trait levels: no symptoms, no HI/low IN symptoms, low symptoms, and moderate symptoms. Not only can the MTMM mixture method detect differential trait mean patterns, but also differential method effects, which is the novelty of the method. After finding that several distinct latent classes were needed to properly account for population heterogeneity in the present application, we examined differences in method effects (convergent validities) across classes and found some of them to be different across classes as well. Latent monotrait-heteromethod correlations showed adequate convergent validity for all but one class (low symptoms) comprising of less than 10% of the sample. A difference between this class and other classes was that it had relatively high IN means, but low HI means. Further, this class showed adequate convergent validity for HI but not for IN. Perhaps IN symptoms are more difficult to accurately evaluate without the presence of HI symptoms. Mean and variance discrepancies, combined with the lack of validity in this class support this conclusion.

Examining validity across classes would be futile if classes were not distinct and separable. It is therefore important to determine whether classes are clearly distinct and separable. Classification probabilities in the present application were above .88 for all classes, indicating that the classes were well-separated and assignment of individuals was highly reliable. Without adequate class separation, it would be unclear if individuals belonged to one subgroup or another subgroup.

The MTMM mixture model advances MTMM researchers' tool box in that it allows them to test a strong assumption that is implicitly made (but rarely tested) by most traditional

MTMM approaches: The assumption of homogenous methods effects across the entire range of trait values. Even MTMM methods which allow testing trait $\times$ method interactions (Cole & Maxwell, 1985; Cole, Truglio, & Peeke, 1997; Pohl & Steyer, 2010; Pohl, Steyer, & Kraus, 2008) do not allow evaluating the assumption of population homogeneity across a priori unknown subgroups. As shown in the present paper, this assumption is rather restrictive and likely violated in many practical applications.

The MTMM mixture model also adds to the many methods created and used to analyze mixture distributions (e.g., Courvoisier et al., 2007; Lubke & Muthén, 2005, 2007; McLachlan & Peel, 2000; Muthén, 2001a; Muthén, 2001b), including a new method that combined mixture distribution methods with MTMM analysis to examine different survey responses (Oberski et al., 2015) and a method that used mixture distribution methods to examine criterion validity (Raykov et al., 2015). The novelty of the method is that it combines mixture distribution analysis and MTMM models to examine convergent and discriminant validity in previously unknown subpopulations. In the presented application, a basic CFA-MTMM measurement model (Marsh & Hocevar, 1988) with indicator-specific residual factors, comprising of two traits measured by two methods was combined with mixture distribution analysis. Future applications should apply mixture distribution techniques to more sophisticated MTMM models with method factors, such as, for example, the models presented by Eid (2000), Eid, Lischetzke, Nussbeck, & Trierweiler (2003), Geiser, Eid, West, Lischetzke, and Nussbeck (2012), Pohl and Steyer (2010), Widaman (1985), or include more traits and/or methods to expand the applicability of this approach.

The MTMM mixture approach may be useful in general for refining validation research. Evaluating validity across subgroups or latent classes is an important step in validity research if inhomogeneity is present within a sample, as has been shown in this paper and also Raykov et al. (2015). If a sample is not homogeneous but treated as such, the results of an MTMM analysis or criterion validity study may be biased.

Limitations

The proposed method requires the estimation of many individual models to determine the most appropriate class model. One reason why many individual models must be evaluated is so that researchers can avoid potential confounds between MI constraints and the number of classes to be retained. The modeling approach therefore requires that all combinations of MI be tested for the 2-, 3-, 4-, etc. class models. In addition, many sets of start values should be tried for each model to ensure that the best loglikelihood value can be replicated (Bauer & Curran, 2003). In our experience, multiclass MTMM models often require a large number of starting value sets before replication of the best loglikelihood value is achieved so that a researcher can have confidence in the model results. Taken together, this can take a lot of computation time even with modern computers, especially for models with multiple traits. The implementation of this method can be challenging in the sense that it may not always be easy to determine the final best-fitting model. We recommend that researchers using the MTMM mixture approach follow the flow chart presented in Figure 2 to guide the model selection procedure for their own data.

The number of starting values required for our models to replicate the best loglikelihood value exceeded those found in previous simulation studies on class enumeration (e.g., Lubke & Muthen, 2005, 2007; Nylund, Asparouhov, & Muthen, 2007). This may be explained by peculiarities of the example data set used or the increased complexity of the models estimated in the current paper. Unlike previous simulation studies on class enumeration, our MTMM mixture approach used a more complex factor structure than factor mixture models used in previous simulation studies. Future simulation research should examine this issue in greater detail for the present models using a variety of parameter constellations.

Another difficulty can be caused by very small (or zero) true score variance in some classes. This can occur when classes are highly homogenous (e.g., children with no symptoms) and may cause convergence problems or improper parameter estimates (e.g., negative latent variance estimates, latent correlations exceeding 1.0). In some of our analyses, true score factors had little or no variance within class, meaning that all individuals within the corresponding class were (almost) perfectly homogeneous with regard to that factor. Having latent variables with zero or very small variances can make estimates of latent correlations undefined or unstable and can result in improper latent correlation estimates (e.g., correlations > 1.0 as found in some of the analyses reported in this paper).

By definition, Campbell and Fiske's (1959) correlation-based MTMM analyses relies on the presence of sufficient variance in the test scores in order to determine convergent and discriminant validities. This is a potential limitation of all MTMM analyses, but especially of the MTMM mixture model which may very well reveal a latent class with little or no true score variance, as in the present application. We assumed that the lack of variance in one of the classes in the application indicated perfect inter-individual homogeneity within this class, such that one or more latent variables became constants within the class. Obviously, a constant cannot covary with other variables, making the assessment of convergent and discriminant validity via correlation analysis impossible.

In other instances, very low variance estimates may have led to unstable latent correlation estimates with point estimates greater than 1.0. Such improper solutions are troublesome, because they are impossible in theory and their interpretation is unclear. Past research has shown that other causes of improper parameter estimates can be sampling fluctuations or model misspecifications (Chen et al., 2001; van Driel, 1978). Both Chen et al. and van Driel suggest reporting the 95% confidence intervals of improper estimates to detect whether such estimates are due to sampling fluctuations or model misspecification. When confidence intervals included proper parameter values, both Chen and van Driel concluded that improper estimates are due to sampling fluctuations.

In the present application, we assumed that at least some of the improper latent correlation estimates were attributable to sampling fluctuations and/or instability in estimation, partly due to low class-specific latent variances and/or small class sizes that caused the related latent correlation parameters to be unstable (as shown, in part, by large standard errors associated with the offending parameter estimates). Although all of the corresponding 95% CI intervals included values < 1.0, these offending parameters need to be interpreted with caution. In some cases, they may simply indicate perfect convergent validity within a given

latent class. In other cases, they might indicate model misspecifications (e.g., over- or under-extraction of latent classes). Future simulation research could help clarify these issues. Simulation studies allow researchers to study correctly specified models and thus can be used to rule out the possibility of model misspecifications as a cause of improper solutions. Given the current lack of absolute fit statistics in CFA mixture modeling, misspecifications are harder to detect with actual empirical data, because the true data-generating model remains unknown.

Conclusion

The creation of the latent class MTMM model (Oberski et al., 2015), the presented MTMM mixture model, and the approach for evaluating criterion validity across classes (Raykov et al., 2015) raise awareness that population homogeneity is frequently assumed in validation research, but often untested, leading to potentially biased results. With the availability of factor mixture analysis, testing for population heterogeneity is relatively straightforward and might become standard practice in the evaluation of MTMM data. We hope that readers will find our step-by-step modeling approach to be useful as a guide to implementing MTMM mixture analyses with their own data.

Acknowledgments

Funding

Data collected for research was supported by the Ministry of Economy and Competitiveness grant PSI2011-23254 (Spanish Government). Christian Geiser's research was supported by a grant from the National Institutes on Drug Abuse of the National Institutes of Health under award number R01 DA034770-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5. Arlington, VA: American Psychiatric Publishing; 2013.
- Bagozzi RP, Yi Y. Testing hypotheses about methods, traits, and communalities in the direct-product model. *Applied Psychological Measurement*. 1992; 16(4):373–380.
- Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*. 2003; 8(3):338–363. [PubMed: 14596495]
- Boomsma A. Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*. 1985; 50(2):229–242.
- Browne M. The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical & Statistical Psychology*. 1984; 37(1):1–21. [PubMed: 6733052]
- Burns, GL., Lee, S. Child and Adolescent Disruptive Behavior Inventory–Parent Version 5.0. Pullman, WA: Author; 2011.
- Burns GL, Servera M, Bernard MdM, Carrillo JM, Geiser C. Ratings of ADHD symptoms and academic impairment by mothers, fathers, teachers, and aides: Construct validity within and across settings as well as occasions. *Psychological Assessment*. 2014; 26(4):1247–1258. <http://dx.doi.org/10.1037/pas0000008>. [PubMed: 24932644]
- Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*. 1989; 105(3):456–466.

- Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*. 1959; 56(2):81–105. [PubMed: 13634291]
- Chen F, Bollen KA, Paxton P, Curran PJ, Kirby JB. Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*. 2001; 29(4):468–508.
- Cole DA, Maxwell SE. Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research*. 1985; 20(4):389–417. DOI: 10.1207/s15327906mbr2004_3 [PubMed: 26821990]
- Cole DA, Truglio R, Peeke L. Relation between symptoms of anxiety and depression in children: A multitrait-multimethod-multigroup assessment. *Journal of Consulting and Clinical Psychology*. 1997; 65(1):110–119. [PubMed: 9103740]
- Collins, LM., Lanza, ST. Latent class and latent transition analysis with applications in the social, behavioral, and health sciences. New York: Wiley; 2010.
- Courvoisier DS, Eid M, Nussbeck FW. Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods*. 2007; 12(1):80–104. [PubMed: 17402813]
- Eid M. A multitrait-multimethod model with minimal assumptions. *Psychometrika*. 2000; 65(2):241–261.
- Eid, M. Methodological approaches for analyzing multimethod data. In: Eid, M., Diener, E., editors. *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association; 2006. p. 223-230. p. 283-299.
- Eid, M., Lischetzke, T., Nussbeck, FW. Structural Equation Models for Multitrait-Multimethod Data. In: Eid, M., Diener, E., editors. *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association; 2006. p. 283-299.
- Eid M, Lischetzke T, Nussbeck FW, Trierweiler LI. Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C (M-1) model. *Psychological Methods*. 2003; 8(1):38. [PubMed: 12741672]
- Fiske DW, Campbell DT. Citations do not solve problems. *Psychological Bulletin*. 1992; 112(3):393–395.
- Fiske DW. Convergent-discriminant validation in measurements and research strategies. *New Directions for Methodology of Social & Behavioral Science*. 1982; 12:77–92.
- Geiser C, Burns GL, Servera M. Testing for measurement invariance and latent mean differences across methods: interesting incremental information from multitrait-multimethod studies. *Frontiers in Psychology*. 2014; 5:1–19. DOI: 10.3389/fpsyg.2014.01216 [PubMed: 24474945]
- Geiser C, Eid M, West SG, Lischetzke T, Nussbeck FW. A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling: A Multidisciplinary Journal*. 2012; 19(3):409–436.
- Kenny DA, Kashy DA. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*. 1992; 112(1):165–172.
- Lazarsfeld, PF., Henry, NW. Latent structure analysis. Boston: Houghton Mifflin; 1968.
- Little TD, Rhemtulla M, Gibson KI, Schoemann AM. Why the item versus parcels controversy needn't be one. *Psychological Methods*. 2013; 18:285–300. [PubMed: 23834418]
- Lubke GH, Muthén B. Investigating population heterogeneity with factor mixture models. *Psychological Methods*. 2005; 10(1):21–39. [PubMed: 15810867]
- Lubke GH, Muthén B. Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*. 2007; 14(1):26–47.
- Marsh HW, Hocevar D. A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*. 1988; 73:107–117.
- Marsh HW, Lüdtke O, Nagengast B, Morin AJS, von Davier M. Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecifications with item parcels in CFA models. *Psychological Methods*. 2013; 18:257–284. [PubMed: 23834417]
- McCutcheon, AL. Latent class analysis [Sage university papers series on quantitative applications in the social sciences, series no. 07-064]. Thousand Oaks, CA: Sage; 1987.

- McLachlan, G.J., Peel, D. Finite mixture models. New York: Wiley; 2000.
- Muthén, B.O. Latent variable mixture modeling. In: Marcoulides, G.A., Schumacker, R.E., editors. New developments and techniques in structural equation modeling. London: Lawrence Erlbaum Associates; 2001a. p. 1-33.
- Muthén, B.O. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling. In: Collins, L.M., Sayer, A.G., Collins, L.M., Sayer, A.G., editors. New methods for the analysis of change. Washington, DC: American Psychological Association; 2001b. p. 291-322.
- Muthén, L.K., Muthén, B.O. Mplus User's Guide. 7. Los Angeles, CA: Muthén and Muthén; 1998–2012.
- Muthén B, Muthén LK. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*. 2000; 24(6):882–891.
- Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*. 2007; 14(4):535–569.
- Oberski DL, Hagenaars JAP, Saris WE. The latent class multitrait-multimethod model. *Psychological Methods*. 2015; 20(4):422–443. [PubMed: 26651987]
- Podsakoff PM, MacKenzie SB, Podsakoff NP. Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*. 2012; 63:539–569.
- Pohl S, Steyer R, Kraus K. Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2008; 171(1):41–63.
- Pohl S, Steyer R. Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*. 2010; 45(1):45–72. DOI: 10.1080/00273170903504729 [PubMed: 26789084]
- Raykov T, Marcoulides GA, Li T. Evaluation of measurement instrument criterion validity in finite mixture settings. *Educational and Psychological Measurement*. 2015:1–19.
- Schoot, Rvd, Lugtig, P., Hox, J. A checklist for testing measurement invariance. *European Journal of Developmental Psychology*. 2012; 9(4):486–492. DOI: 10.1080/17405629.2012.686740
- Steyer R, Eid M, Schwenkmezger P. Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*. 1997; 2(1):21–33.
- van Driel OP. On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*. 1978; 43(2):225–243.
- Vermunt, J.K., Magidson, J. Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.L., editors. *Applied latent class analysis*. Cambridge, UK: Cambridge University Press; 2002. p. 89-106.
- Widaman KF. Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*. 1985; 9(1):1–26.
- Widaman KF, Reise SP. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*. 1997:281–324.
- Wothke W, Browne MW. The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika*. 1990; 55(2):255–262.
- Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM, Perou R, Blumberg SJ. Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2014; 53:34–46. [PubMed: 24342384]
- Willcutt EG. The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A meta-analytic review. *Neurotherapeutics*. 2012; 9:490–499. [PubMed: 22976615]

Appendix A. Mplus Input File for Model 30

```

Title: Four-class MTMM Mixture Model for HI and IN
      Strong invariance & partial strict invariance across classes
Variable:
Names are HI1momT3 HI2momT3 HI3momT3 HI1dadT3 HI2dadT3 HI3dadT3
      IN1momT3 IN2momT3 IN3momT3 IN1dadT3 IN2dadT3 IN3dadT3;
classes = c(4);
Analysis:
Type = mixture;
starts = 200000 20000;
Model:
%overall%
! Factor loadings set equal across classes
HIMOM by HI1momT3@1
      HI2momT3 (lambda2a)
      HI3momT3 (lambda3a);
HIDAD by HI1dadT3@1
      HI2dadT3 (lambda2a)
      HI3dadT3 (lambda3a);
INMOM by IN1momT3@1
      IN2momT3 (lambda2)
      IN3momT3 (lambda3);
INDAD by IN1dadT3@1
      IN2dadT3 (lambda2)
      IN3dadT3 (lambda3);
! Intercepts set equal across classes
[HI1momT3@0 HI1dadT3@0];
[HI2momT3 HI2dadT3] (alpha2a);
[HI3momT3 HI3dadT3] (alpha3a);
[IN1momT3@0 IN1dadT3@0];
[IN2momT3 IN2dadT3] (alpha2);
[IN3momT3 IN3dadT3] (alpha3);
!Error variances set equal across classes
HI1momT3 HI1dadT3 (eps1a);
HI2momT3 HI2dadT3 (eps2a);
HI3momT3 HI3dadT3 (eps3a);
IN1momT3 IN1dadT3 (eps1);
IN2momT3 IN2dadT3 (eps2);
IN3momT3 IN3dadT3 (eps3);
! IS factor loadings set to 1
HI2 by HI2momT3@1 HI2dadT3@1;
HI3 by HI3momT3@1 HI3dadT3@1;
IN2 by IN2momT3@1 IN2dadT3@1;

```

```

IN3 by IN3momT3@1 IN3dadT3@1;
! IS factors do not have means
[HI2@0 HI3@0 IN2@0 IN3@0];
! IS factors not correlated with true score factors
HI2 HI3 with HIMOM@0 HIDAD@0;
IN2 IN3 with INMOM@0 INDAD@0;
%C#1%
[HI1momT3@0 HI1dadT3@0 HI2@0 HI3@0 HIMOM HIDAD]
[IN1momT3@0 IN1dadT3@0 IN2@0 IN3@0 INMOM INDAD];
HI2 HI3 with HIMOM@0 HIDAD@0;
IN2 IN3 with INMOM@0 INDAD@0;
HI2 HI3 IN2 IN3;
HIMOM (VHIml);
HIDAD (VHIIdl);
INMOM (VINml);
INDAD (VINIdl);
!trait correlations
HIMOM with HIDAD (HIm_HIdl);
INMOM with INDAD (INm_INdl);
HIMOM with INMOM (HIm_INml);
HIDAD with INDAD (HId_INdl);
HIMOM with INDAD (HIm_INdl);
HIDAD with INMOM (HId_INml);
!IS correlations
HI2 with HI3; IN2 with IN3; HI2 with IN2;
HI3 with IN3; HI3 with IN2; HI2 with IN3;
%C#2%
[HIMOM HIDAD INMOM INDAD];
HI2 HI3 IN2 IN3;
HIMOM (VHIml); ! this was found equal to class 1 variance
HIDAD (VHIIdl); ! this was found equal to class 1 variance
INMOM (VINm2);
INDAD (VINId2);
!trait correlations
HIMOM with HIDAD (HIm_HId2);
INMOM with INDAD (INm_INd2);
HIMOM with INMOM (HIm_INm2);
HIDAD with INDAD (HId_INd2);
HIMOM with INDAD (HIm_INd2);
HIDAD with INMOM (HId_INm2);
!IS correlations
HI2 with HI3; IN2 with IN3; HI2 with IN2;
HI3 with IN3; HI3 with IN2; HI2 with IN3;
%C#3%

```

```

[HIMOM HIDAD INMOM INDAD];
HI2 HI3 IN2 IN3;
HIMOM (VHIm3);
HIDAD (VHId3);
INMOM (VINm3);
INDAD (VIND3);

!Error variances different for IN
IN1momT3 (eps4);
IN2momT3 (eps5);
IN3momT3 (eps6);
IN1dadT3 (eps4);
IN2dadT3 (eps5);
IN3dadT3 (eps6);

!trait correlations
HIMOM with HIDAD (HIm_HId3);
INMOM with INDAD (INm_INd3);
HIMOM with INMOM (HIm_INm3);
HIDAD with INDAD (HId_INd3);
HIMOM with INDAD (HIm_INd3);
HIDAD with INMOM (HId_INm3);

!IS correlations
HI2 with HI3; IN2 with IN3; HI2 with IN2;
HI3 with IN3; HI3 with IN2; HI2 with IN3;
%C#4%

[HIMOM HIDAD INMOM INDAD];
HI2 HI3 IN2 IN3;
HIMOM (VHIm4);
HIDAD (VHId4);
INMOM (VINm4);
INDAD (VIND4);

!Error variances different for HI
HI1momT3 (eps10a);
HI2momT3 (eps11a);
HI3momT3 (eps12a);
HI1dadT3 (eps10a);
HI2dadT3 (eps11a);
HI3dadT3 (eps12a);

!Error variances different for IN
IN1momT3 (eps7);
IN2momT3 (eps8);
IN3momT3 (eps9);
IN1dadT3 (eps7);
IN2dadT3 (eps8);
IN3dadT3 (eps9);

```

```

!trait correlations
HIMOM with HIDAD (HIm_HId4);
INMOM with INDAD (INm_INd4);
HIMOM with INMOM (HIm_INm4);
HIDAD with INDAD (HId_IND4);
HIMOM with INDAD (HIm_INd4);
HIDAD with INMOM (HId_INm4);

!IS correlations
HI2 with HI3; IN2 with IN3; HI2 with IN2;
HI3 with IN3; HI3 with IN2; HI2 with IN3;

Model Constraint:

! Monotrait Heteromethod HImom HIdad
New(Dhi12, Dhi13, Dhi14, Dhi23, Dhi24, Dhi34);
Dhi12= (HIm_HId1/((VHIm1**0.5)*(VHId1**0.5)))-(HIm_HId2/
((VHIm1**0.5)*(VHId1**0.5)));
Dhi13= (HIm_HId1/((VHIm1**0.5)*(VHId1**0.5)))-(HIm_HId3/
((VHIm3**0.5)*(VHId3**0.5)));
Dhi14= (HIm_HId1/((VHIm1**0.5)*(VHId1**0.5)))-(HIm_HId4/
((VHIm4**0.5)*(VHId4**0.5)));
Dhi23= (HIm_HId2/((VHIm1**0.5)*(VHId1**0.5)))-(HIm_HId3/
((VHIm3**0.5)*(VHId3**0.5)));
Dhi24= (HIm_HId2/((VHIm1**0.5)*(VHId1**0.5)))-(HIm_HId4/
((VHIm4**0.5)*(VHId4**0.5)));
Dhi34= (HIm_HId3/((VHIm3**0.5)*(VHId3**0.5)))-(HIm_HId4/
((VHIm4**0.5)*(VHId4**0.5)));

! Monotrait Heteromethod INmom INdad
New(Din12, Din13, Din14, Din23, Din24, Din34);
Din12= (INm_INd1/((VINm1**0.5)*(VIND1**0.5)))-(INm_INd2/
((VINm2**0.5)*(VIND2**0.5)));
Din13= (INm_INd1/((VINm1**0.5)*(VIND1**0.5)))-(INm_INd3/
((VINm3**0.5)*(VIND3**0.5)));
Din14= (INm_INd1/((VINm1**0.5)*(VIND1**0.5)))-(INm_INd4/
((VINm4**0.5)*(VIND4**0.5)));
Din23= (INm_INd2/((VINm2**0.5)*(VIND2**0.5)))-(INm_INd3/
((VINm3**0.5)*(VIND3**0.5)));
Din24= (INm_INd2/((VINm2**0.5)*(VIND2**0.5)))-(INm_INd4/
((VINm4**0.5)*(VIND4**0.5)));
Din34= (INm_INd3/((VINm3**0.5)*(VIND3**0.5)))-(INm_INd4/
((VINm4**0.5)*(VIND4**0.5)));

! Heterotrait Monomethod HImom INmom
New(Dmo12, Dmo13, Dmo14, Dmo23, Dmo24, Dmo34);
Dmo12= (HIm_INm1/((VHIm1**0.5)*(VINm1**0.5)))-(HIm_INm2/
((VHIm1**0.5)*(VINm2**0.5)));
Dmo13= (HIm_INm1/((VHIm1**0.5)*(VINm1**0.5)))-(HIm_INm3/

```

```

((VHIm3**0.5)*(VINm3**0.5)));
Dmo14= (HIm_INm1/((VHIm1**0.5)*(VINm1**0.5)))-(HIm_INm4/
((VHIm4**0.5)*(VINm4**0.5)));
Dmo23= (HIm_INm2/((VHIm1**0.5)*(VINm2**0.5)))-(HIm_INm3/
((VHIm3**0.5)*(VINm3**0.5)));
Dmo24= (HIm_INm2/((VHIm1**0.5)*(VINm2**0.5)))-(HIm_INm4/
((VHIm4**0.5)*(VINm4**0.5)));
Dmo34= (HIm_INm3/((VHIm3**0.5)*(VINm3**0.5)))-(HIm_INm4/
((VHIm4**0.5)*(VINm4**0.5)));
! Heterotrait Monomethod HIdad INdad
New(Dda12, Dda13, Dda14, Dda23, Dda24, Dda34);
Dda12= (HId_INd1/((VHId1**0.5)*(VIND1**0.5)))-(HId_INd2/
((VHId1**0.5)*(VIND2**0.5)));
Dda13= (HId_INd1/((VHId1**0.5)*(VIND1**0.5)))-(HId_INd3/
((VHId3**0.5)*(VIND3**0.5)));
Dda14= (HId_INd1/((VHId1**0.5)*(VIND1**0.5)))-(HId_INd4/
((VHId4**0.5)*(VIND4**0.5)));
Dda23= (HId_INd2/((VHId1**0.5)*(VIND2**0.5)))-(HId_INd3/
((VHId3**0.5)*(VIND3**0.5)));
Dda24= (HId_INd2/((VHId1**0.5)*(VIND2**0.5)))-(HId_INd4/
((VHId4**0.5)*(VIND4**0.5)));
Dda34= (HId_INd3/((VHId3**0.5)*(VIND3**0.5)))-(HId_INd4/
((VHId4**0.5)*(VIND4**0.5)));
! Heterotrait Heteromethod HIdad INmom
New(HdIm12, HdIm13, HdIm14, HdIm23, HdIm24, HdIm34);
HdIm12= (HId_INm1/((VHId1**0.5)*(VINm1**0.5)))-(HId_INm2/
((VHId1**0.5)*(VINm2**0.5)));
HdIm13= (HId_INm1/((VHId1**0.5)*(VINm1**0.5)))-(HId_INm3/
((VHId3**0.5)*(VINm3**0.5)));
HdIm14= (HId_INm1/((VHId1**0.5)*(VINm1**0.5)))-(HId_INm4/
((VHId4**0.5)*(VINm4**0.5)));
HdIm23= (HId_INm2/((VHId1**0.5)*(VINm2**0.5)))-(HId_INm3/
((VHId3**0.5)*(VINm3**0.5)));
HdIm24= (HId_INm2/((VHId1**0.5)*(VINm2**0.5)))-(HId_INm4/
((VHId4**0.5)*(VINm4**0.5)));
HdIm34= (HId_INm3/((VHId3**0.5)*(VINm3**0.5)))-(HId_INm4/
((VHId4**0.5)*(VINm4**0.5)));
! Heterotrait Heteromethod HIMom INdad
New(HmId12, HmId13, HmId14, HmId23, HmId24, HmId34);
HmId12= (HIm_INd1/((VHIm1**0.5)*(VIND1**0.5)))-(HIm_INd2/
((VHIm1**0.5)*(VIND2**0.5)));
HmId13= (HIm_INd1/((VHIm1**0.5)*(VIND1**0.5)))-(HIm_INd3/
((VHIm3**0.5)*(VIND3**0.5)));
HmId14= (HIm_INd1/((VHIm1**0.5)*(VIND1**0.5)))-(HIm_INd4/

```



```

((VHIm4**0.5)*(VIND4**0.5)));
HmId23= (HIm_INd2/((VHIm1**0.5)*(VIND2**0.5)))-(HIm_INd3/
((VHIm3**0.5)*(VIND3**0.5)));
HmId24= (HIm_INd2/((VHIm1**0.5)*(VIND2**0.5)))-(HIm_INd4/
((VHIm4**0.5)*(VIND4**0.5)));
HmId34= (HIm_INd3/((VHIm3**0.5)*(VIND3**0.5)))-(HIm_INd4/
((VHIm4**0.5)*(VIND4**0.5)));

```

Figure 1.

Figure 2.

Figure 3.

Table 1

Single-Class Model Fit Information

	HI						IN					
	χ^2	df	p	χ^2_{diff}	df _{diff}	p _{diff}	AIC	BIC	χ^2	df	p	χ^2_{diff}
Model 1: Configural MI (no IS factors)	252.50	8	<.001				6347	6432	172.40	8	<.001	
Model 2: Configural MI (IS factors)	12.44	5	.03	240.05	3	.00	6113	6211	4.92	5	.43	167.79
Model 3: Weak MI	13.01	7	.07	0.56	2	.77	6110	6198	6.73	7	.46	1.82
Model 4: Strong MI	16.35	9	.06	3.35	2	.19	6109	6189	8.00	9	.53	1.27
Model 5: Strict MI	21.78	12	.04	5.43	3	.17	6109	6175	8.75	12	.72	0.76

Note. HI = hyperactivity/impulsivity; IN = inattention; df = degrees of freedom, AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; MI = measurement invariance; IS = indicator-specific. Bold face indicates best-fitting models.

Table 2

Measurement Invariance and Structural Parameter Equality across Multi-Class Models for HI: Model Fit Information

Model Description	AIC	BIC
2-class models		
Model 6a: Configural MI across classes	5174	5311
Model 7a: Weak MI across classes	5222	5350
Model 8a: Strong MI across classes	5223	5343
Model 9a: Strict MI across classes	5374	5480
3-class models		
Model 10a: Configural MI across 3 classes	5063	5271
Model 11a: Weak MI across 3 classes	5095	5285
Model 12a: Weak MI across 2 classes	5061	5261
Model 13a: Strong MI across 2 classes	5096	5286
Model 14a: Weak MI across 2 classes + Equal error variances across 2 classes	5010	5197
4-class models		
Model 15a: Configural MI across 4 classes	5095	5373
Model 16a: Weak MI across 4 classes	5068 ^a	5321 ^a
Model 17a: Strong MI across 4 classes	5055 ^a	5281 ^a
Model 18a: Strict MI across 4 classes	5137 ^a	5323 ^a
Model 19a: Strong MI across 4 classes + Equal error variances across 3 classes	4987 ^a	5186 ^a
Structural Parameters		
Model 20a: Model 19 + Equal factor variances across 4 classes	5439	5612
Model 21a: Model 19 + Equal factor variances across 3 classes	5058	5240
Model 22a: Model 19 + Equal factor variances across 2 classes	4916	5106
Model 23a: Model 22 + Equal IS variances across 4 classes	5088	5252
Model 24a: Model 22 + Equal IS variances across 3 classes	5016	5189
Model 25a: Model 22 + Equal IS variances across 2 classes	4949	5130
Model 26a: Model 22 + Equal factor means across 4 classes	6153	6316
Model 27a: Model 22 + Equal factor means across 3 classes	5085	5258
Model 28a: Model 22 + Equal factor means across 2 classes	5000	5182

Note. HI = hyperactivity/impulsivity; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; MI = measurement invariance; IS = indicator-specific factor. 5-class models were also tested to determine whether another class was needed. All 5-class models that were tested showed worse fit (lowest BIC = 5346), and all 5-class models contained a class with a proportion of 0%.

^aFor Models 16a, 17a, and 19a, the best loglikelihood values could not be replicated after 100,000 start values.

Therefore, the AIC and BIC values need to be interpreted with caution for these models. All other models had replicated loglikelihood values after 10,000 (for 2 & 3 class models) or 100,000 (for 4 class models) sets of start values. Bold face indicates the best-fitting model.

Table 3

Measurement Invariance & Structural Parameter Equality across Multi-Class Models for IN: Model Fit Information

Model Description	AIC	BIC
2-class models		
Model 6b: Configural MI across classes	4904	5042
Model 7b: Weak MI across classes	4906	5034
Model 8b: Strong MI across classes	4904	5023
Model 9b: Strict MI across classes	5118	5225
3-class models		
Model 10b: Configural MI across 3 classes	4812	5020
Model 11b: Weak MI across 3 classes	4834	5024
Model 12b: Weak MI across 2 classes	4809	5008
Model 13b: Strong MI across 2 classes	NC	
Model 14b: Strong MI across 3 classes	4835	5008
Model 15b: Strict MI across 3 classes	4932	5078
Model 16b: Strong MI across 3 classes + Equal error variances across 2 classes	4695	4854
4-class models		
Model 17b: Configural MI across 4 classes	4744	5023
Model 18b: Weak MI across 4 classes	4777	5030
Model 19b: Weak MI across 3 classes	4781	5043
Model 20b: Weak MI across 2 classes	4806	5076
Structural parameters		
Model 21b: Model 16 + Equal factor variances across 3 classes	4991	5133
Model 22b: Model 16 + Equal factor variances across 2 classes	4784	4934
Model 23b: Model 16 + Equal IS variances across 3 classes	4846	4979
Model 24b: Model 16 + Equal IS variances across 2 classes	4825	4966
Model 25b: Model 16 + Equal factor means across 3 classes	5669	5802
Model 26b: Model 16 + Equal factor means across 2 classes	4852	4994

Note. HI = hyperactivity/impulsivity; IN = inattention; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; MI = measurement invariance; NC = no convergence. Models with 4 classes were tested to determine whether another class was present. All 4-class models that were tested showed worse fit (lowest BIC = 4744). All models had replicated loglikelihood values after 10,000 (for 2-class models) and 50,000 (for 3-class models) sets of start values. Bold face indicates the best-fitting model.

Table 4

Correlation Differences across Classes in the Four-Class MTMM Mixture Model

Correlated Variables	Classes compared	ϕ_{diff}	SE	<i>p</i>
HI mother with HI father	Class 1 – Class 2	.50	.31	.10
	Class 1 – Class 3	.63	.34	.07
	Class 1 – Class 4	.69*	.28	.02
	Class 2 – Class 3	.13	.45	.78
	Class 2 – Class 4	.19	.40	.65
	Class 3 – Class 4	.06	.11	.60
IN mother with IN father	Class 1 – Class 2	.69*	.24	.00
	Class 1 – Class 3	.16	.19	.41
	Class 1 – Class 4	.39*	.17	.02
	Class 2 – Class 3	–.53*	.22	.02
	Class 2 – Class 4	–.30	.18	.09
	Class 3 – Class 4	.24	.16	.13
HI mother with IN mother	Class 1 – Class 2	.07	.31	.83
	Class 1 – Class 3	–.28	.34	.40
	Class 1 – Class 4	–.27	.34	.41
	Class 2 – Class 3	–.22	.26	.41
	Class 2 – Class 4	–.21	.25	.41
	Class 3 – Class 4	.01	.11	.93
HI father with IN father	Class 1 – Class 2	–.07	.27	.78
	Class 1 – Class 3	.27	.31	.39
	Class 1 – Class 4	.33	.31	.29
	Class 2 – Class 3	.34	.27	.20
	Class 2 – Class 4	.41	.26	.12
	Class 3 – Class 4	.06	.11	.56
HI father with IN mother	Class 1 – Class 2	–.12	.44	.79
	Class 1 – Class 3	–.15	.36	.67
	Class 1 – Class 4	–.05	.34	.89
	Class 2 – Class 3	–.04	.36	.92
	Class 2 – Class 4	.07	.35	.84
	Class 3 – Class 4	.11	.13	.41
HI mother with IN father	Class 1 – Class 2	.54	.31	.08
	Class 1 – Class 3	.06	.40	.88
	Class 1 – Class 4	.21	.35	.54
	Class 2 – Class 3	–.48	.38	.20
	Class 2 – Class 4	–.33	.32	.30
	Class 3 – Class 4	.15	.13	.24

*Note.** $p < .05$. HI = hyperactivity/impulsivity; IN = inattention. ϕ_{diff} = correlation difference value.