Teacher's Corner:

Applying and Interpreting Mixture Distribution Latent State-Trait Models

Kaylee Litson, Carly Thornhill, and Christian Geiser

Utah State University

G. Leonard Burns

Washington State University

Mateu Servera

University of the Balearic Islands

Author Note

Kaylee Litson, Department of Psychology, Utah State University; Carly Thornhill, Department of Mathematics & Statistics, 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.

This research was in part supported by the Ministry of Economy and Competitiveness grant PSI2011-23254 (Spanish Government).

Correspondence concerning this article should be addressed to Kaylee Litson, Department of Psychology, Utah State University, Logan, UT 84322. Online supplemental materials for this article can be accessed at https://osf.io/kj9eg/.

Email: kaylee.litson@gmail.com

Abstract

Latent state-trait (LST) models are commonly applied to determine the extent to which observed variables reflect trait-like versus state-like constructs. Mixture distribution LST (M-LST) models (Courvoisier, Eid, & Nussbeck, 2007) relax the assumption of population homogeneity made in traditional LST models, allowing researchers to identify subpopulations (latent classes) with differing trait- and state-like attributes. Applications of M-LST models are scarce, presumably because of the analysis complexity. We present a step-by-step tutorial for evaluating M-LST models based on an application to mother, father, and teacher reports of children's inattention (N = 811). In the application, we found three latent classes for mother and father reports and four classes for teacher reports. All reporter solutions contained classes with very low, low, and moderate levels of inattention. The teacher solution also contained a class with high inattention. Comparable mother and father (but not teacher) classes exhibited similar levels of trait and state variance.

Keywords: latent state-trait, mixture distribution modeling, consistency, occasion-specificity, longitudinal modeling, latent classes

Teacher's Corner:

Applying and Interpreting Mixture Distribution Latent State-Trait Models

Longitudinal data analysis is increasingly applied in psychology and social science research, and numerous structural equation modeling (SEM) approaches to longitudinal data analysis have been proposed (e.g., Bollen & Curran, 2006; Cole & Maxwell, 2003; Grimm, Mazza, & Mazzocco, 2016; McArdle, 1986, 2009; Steyer, Mayer, Geiser, & Cole, 2015). Latent variable SEM techniques are advantageous for the analysis of longitudinal data because of their capabilities to flexibly model change and stability across time. Further, SEM techniques allow testing underlying model assumptions, evaluating change and stability using multiple indicators, and correcting for random measurement error.

One question that researchers address with longitudinal data is whether psychological attributes (e.g., depression, anxiety, well-being, happiness, emotion, impulsivity) reflect stable, trait-like constructs or rather variable, state-like constructs. Latent state-trait (LST) models (Steyer, Ferring, & Schmitt, 1992; Steyer et al., 2015; Steyer, Schmitt, & Eid, 1999) are widely used to evaluate the stable trait- and variable state-like nature of psychological attributes across occasions and are increasingly applied in the social sciences (Geiser & Lockhart, 2012; Prenoveau, 2016).

LST models decompose observed score variance into different components: trait variance, occasion-specific variance, and random measurement error variance (Steyer et al., 2015). The trait component reflects intra-individual stability (i.e., consistency) across time. The occasion-specific component reflects momentary deviations of individuals' true scores from their trait levels within each time point and characterizes effects of situations as well as person × situation interactions. Random measurement error is unsystematic variability in the measurement due to neither trait nor occasion-specific components. These three variance components are

fundamental for identifying the extent to which a variable is stable and trait-like versus more fluctuating and state-like. A variable containing more trait variance than occasion-specific variance is considered trait-like. In contrast, a variable containing more occasion-specific variance than trait variance is considered state-like. More complex LST models also allow identifying method (indicator-specific) variance components (Courvoisier, Nussbeck, Eid, Geiser, & Cole, 2008; Geiser & Lockhart, 2012).

Conventional LST models assume that all individuals in a sample stem from a single homogenous population in which a single set of LST parameters (e.g., the trait, occasion-specific, and measurement error variances) applies to all individuals. Courvoisier, Eid, and Nussbeck (2007) demonstrated that the assumption of population homogeneity can be violated in practice because of the presence of unknown subpopulations (latent classes) that show different trait means and/or that differ with regard to trait consistency, occasion-specificity, and reliability of the measures. When applied to a heterogeneous population, conventional (single-class) LST models may lead to inaccurate or misleading results about the true trait consistency and occasion-specificity of a particular construct for different individuals.

For example, if individuals with lower levels of anxiety show more consistency in their symptoms across time than individuals with higher levels of anxiety – a potential violation of population homogeneity – then resulting trait and occasion-specific variance components from a single-class LST model would not reflect such heterogeneity. Should population heterogeneity truly exist in the data, parameter estimates and resulting conclusions about the trait- and state-like nature of the attribute could be biased. Hypothesizing that anxiety is more consistent and trait-like for individuals with lower anxiety levels than individuals with higher anxiety levels is theoretically plausible, yet such a hypothesis cannot be tested using a single-class LST model (unless groups of individuals with low versus high anxiety were known beforehand). These

situations call for more advanced LST models since LST models assuming population homogeneity would be inappropriate.

Mixture distribution LST (M-LST) models (Courvoisier et al., 2007) relax the assumption of population homogeneity and allow identifying subpopulations (latent classes) across which some or all LST model parameters differ. This allows researchers to identify subgroups of individuals who differ, for example, in their (1) mean trait values, (2) trait variances, (3) occasion-specific variances, and/or (4) measurement error variances (unreliability of measurement). The M-LST model is an extension of standard LST models and can be used to examine whether there are different, previously unknown subpopulations (latent classes) that differ with regard to consistency and variability.

M-LST models are a special case of general factor mixture models (Lubke & Muthén, 2005; 2007; Muthén, 2001). Although factor mixture modeling is widely used in other areas of longitudinal data analysis (e.g., growth mixture modeling; Muthén & Muthén, 2000), it has not been frequently applied to the analysis of state and trait components in social science constructs. The only application of M-LST models that we know of is the one presented in the original Courvoisier et al. (2007) article. We suspect that the lack of use of the M-LST approach by applied researchers may be due to the complex nature of M-LST models, which are not trivial in their application.

The purpose of this paper is to present a step-by-step tutorial for applying M-LST models and interpreting the resulting output. Before discussing M-LST models, we provide a more in-depth description of a prototypical single-class LST model. Subsequently, we show how the single-class LST model is extended to an M-LST model. Third, we discuss the application of the M-LST approach to a data set on children's levels of inattention, using the software package

M*plus* 8 (Muthén & Muthén, 1998-2017). Finally, we discuss some of the advantages and drawbacks of applying the M-LST approach.

LST Models

LST models are longitudinal models that can be used to partition observed variables into consistent trait, fluctuating occasion-specific, and measurement error components (e.g., Cole, Martin, & Steiger, 2005; Eid, Holtmann, Santangelo, & Ebner-Priemer, 2017; Geiser & Lockhart, 2012; Prenoveau, 2016; Schermelleh-Engel, Keith, Moosbrugger, & Hodapp, 2004; Steyer et al., 1992; 2015). These different components are used to determine the proportion of variability that is due to trait influences (stable dispositions), occasion-specific influences (situation and person-situation interactions), and measurement error influences, showing the extent to which each observed variable is trait-like versus state-like. To apply LST models, multiple observed variables (e.g., indicators, items) must each be measured at multiple (at least two) measurement occasions.[1] More complex LST models with autoregressive effects (e.g., Cole et al., 2005; Eid et al., 2017; Kenny & Zautra, 1995; Prenoveau, 2016) require more than two measurement occasions to be identified.

The basic decomposition of observed variables in LST theory is closely related to concepts of classical test theory (Lord & Novick, 1968; Novick, 1966). According to LST theory, each observed variable Y_{it} can be decomposed into a latent state (true score) variable τ_{it} and a measurement error variable ε_{it}:

$$Y_{it} = \tau_{it} + \varepsilon_{it} . \quad (1)$$

[1] Kenny and Zautra (1995) presented a single-indicator LST model. In principle, the M-LST approach that we illustrate in this article could also be applied to Kenny and Zautra's model. In this article, we focus on multiple-indicator LST models, as these have been shown to result in fewer estimation problems compared to the Kenny and Zautra approach (Cole et al., 2005).

The subscripts *i* and *t* indicate the *i*th indicator ($i = 1, \ldots, m$) and the *t*th time point ($t = 1, \ldots k$). The latent state variable τ_{it} represents systematic sources of score variability due to person (trait) and occasion (situation and/or person-situation interaction) effects, whereas the error variable ε_{it} represents unsystematic influences of measurement error.

In conventional (single-class) LST models, the latent state variables τ_{it} are further decomposed into trait and occasion-specific (state) residual variables. Here, we present the multitrait-multistate (MTMS) model (Eid, 1996) as a prototypical LST model (see Figure 1). Due to differences in content, item wording, method effects, or other differences among indicators, the Y_{it} variables may not measure a single homogenous trait factor. The MTMS model in Figure 1 therefore uses indicator-specific trait factors to account for indicator heterogeneity.

Formally, the MTMS model decomposes the latent state variables τ_{it} into indicator-specific trait T_i and occasion-specific residual O_t components: $\tau_{it} = T_i + \gamma_i O_t$, where γ_i is a constant time-invariant scaling (factor loading) parameter. Substituting this decomposition into the basic LST Equation 1 shows that each observed variable can be partitioned into an indicator-specific trait factor T_i, an occasion-specific residual factor O_t, and a measurement error variable ε_{it}:

$$Y_{it} = T_i + \gamma_i O_t + \varepsilon_{it}. \quad (2)$$

The indicator-specific trait factors T_i represent the temporally stable aspects of a given observed variable, whereas the occasion-specific residual factors O_t reflect systematic deviations from the trait level due to the situation and/or person × situation interaction effects at time *t* that are shared across all indicators measured at the same time point. Being defined as residual

variables, the O_t factors have means of zero by definition. Therefore, latent means are only estimated for the trait factors T_i.

The trait factors are allowed to correlate with one another, but we assume in this article that trait factors do not correlate with occasion-residual factors or measurement error variables. We also assume that occasion residual factors are uncorrelated with each other and with all measurement error variables and that error variables are uncorrelated with each other. Not all of these restrictions are required, but they simplify the presentation of the M-LST approach (for a detailed discussion of LST models that relax some of the independence assumptions made here, see Eid et al., 2017). The factor loadings γ_i are typically assumed to be time-invariant for the same indicator to establish measurement equivalence across time.

In summary, the single-class MTMS model estimates the following parameters: *m* trait factor means (where *m* indicates the total number of observed variables per occasion), *m* trait factor variances, $m \cdot (m - 1)/2$ trait factor covariances, *k* occasion-specific residual factor variances (where *k* indicates the total number of measurement occasions), $m - 1$ occasion-specific factor loadings γ_i (one loading per occasion factor is fixed to one for identification and loadings are assumed to be time-invariant), and $m \cdot k$ measurement error variances.

The single-class MTMS model in Figure 1 is a good starting point for analyzing M-LST models because it often shows a decent fit in practical applications. Moreover, Geiser and Lockhart (2012) found that the MTMS model performed well in simulations with different levels of indicator heterogeneity. We therefore use the MTMS model as the baseline model to demonstrate the M-LST approach in the present paper. If the MTMS model does not fit well in an empirical application, one reason may be that there are autoregressive effects between adjacent occasion residual factors, for example, due to a short time lag between measurement

occasions. First-order autoregressive effects can be included for the occasion residual factors to reflect such effects of short-term stability (Cole et al., 2005; Eid et al., 2017; Prenoveau, 2016). Another possible cause of misfit in the basic MTMS model may be that trait changes occurred across time. Trait means are assumed to remain stable and unchanging across time in the MTMS model. If it is likely that trait changes occurred in addition to a state variability process, researchers should empirically evaluate extended LST models that also include trait-change components (Eid & Hoffmann, 1998; Geiser et al., 2015; 2017; Steyer et al., 2015). The general procedures discussed below can be adapted for use with such more complex models.

Extending Single-Class LST Models to M-LST Models

Single-class LST models are suitable when individuals in a sample come from a single, homogeneous population. However, in the presence of population heterogeneity, a single-class LST model could lead to inaccurate or misleading results about the stable trait-like and fluctuating state-like nature of a psychological attribute. M-LST models relax the assumption of population homogeneity by allowing for several latent classes across which some or all model parameters may differ (McLachlan & Peel, 2000; Muthén, 2001).

Within the M-LST framework, latent means, variances, and other model parameters may vary across latent classes *C*, leading to different latent variable distributions across classes. Mathematically, if $C > 1$, each within-class model is estimated jointly using a mixture distribution such that,

$$f(Y) = \sum_{c=1}^{C} \pi_c f_c(Y;\, \theta_c) \ , \tag{3}$$

where *Y* is the vector of observed variables, π_c is the relative class size parameter, and θ_c is the vector of model parameters within the *c*th latent class (for more details, see McLachlan & Peel, 2000). This mixture distribution equation suggests that observed variables are a function of a set

of model parameters, θ_c, with values specific to a given latent class *c*. Consequently, this mixture distribution equation indicates that any model parameter, such as trait means, occasion-specific variances, and factor loadings may be class-specific.

Within each class, the LST model and underlying assumptions are expected to hold, as is expressed by the following model equation:

$$Y_{it} = T_{ic} + \gamma_{ic} O_{tc} + \varepsilon_{itc} . \quad (4)$$

Equation 4 is identical to the single-class LST model Equation 2 except for the addition of the subscript *c*. The subscript *c* indicates that parameters can now be class-specific (e.g., parameters such as the trait factor variances may differ across unknown subgroups *c*). The trait factor T_{ic}, occasion-specific factor O_{tc}, and error variable ε_{itc} can be interpreted within each class as they would in a single-class model.

The relative class size parameter π_c is a probability parameter that indicates the proportion of individuals who are expected to fall within a given class. The class size parameters sum to 1 across the *C* classes: $\sum_{c=1}^{C} \pi_c = 1$. Therefore, the classes are mutually exclusive and exhaustive and there are only *C* – 1 independent class size parameters to estimate.

The class in which an individual is placed is determined by their posterior probability within each class. A posterior probability value is assigned to each individual for each latent class, and the class for which each individual has the greatest posterior probability is the class to which each individual is assigned.

All MTMS model parameters (i.e., the trait means, trait variances, trait covariances, occasion-specific variances, occasion-specific factor loadings, and measurement error variances) in an M-LST analysis may vary across classes. One important goal of an M-LST analysis is to

empirically evaluate which parameter estimates are class-specific versus class-invariant. We demonstrate this in our tutorial section below.

Calculating LST model effect size coefficients. Because we assume the latent trait, occasion-specific, and error variables in the MTMS model to be uncorrelated, the observed variable variances can be additively decomposed into trait, occasion-specific state residual, and measurement error variance within each class. It is also possible to determine the amount of observed variance in each variable that is due to trait components (*consistency*) versus occasion-specific state residual components (*occasion-specificity*).

The consistency coefficient Con represents the proportion of observed variance that is due to the stable trait component:

$$Con = \frac{Var(T_{ic})}{Var(T_{ic}) + \gamma_{ic}^2 Var(O_{tc}) + Var(\varepsilon_{itc})}. \quad (5)$$

The occasion-specificity coefficient $OSpe$ represents the proportion of observed variance that is due to momentary deviations from the trait level:

$$OSpe = \frac{\gamma_{ic}^2 Var(O_{tc})}{Var(T_{ic}) + \gamma_{ic}^2 Var(O_{tc}) + Var(\varepsilon_{itc})}. \quad (6)$$

The reliability coefficient Rel represents the proportion of observed variance that is due to either consistency or occasion-specificity – the two systematic sources of variance – and not due to measurement error:

$$Rel = \frac{Var(T_{ic}) + \gamma_{ic}^2 Var(O_{tc})}{Var(T_{ic}) + \gamma_{ic}^2 Var(O_{tc}) + Var(\varepsilon_{itc})}. \quad (7)$$

In summary, the consistency and occasion-specificity coefficients represent systematic proportions of variance (i.e., portions of true score variance; variance that is not due to measurement error) that sum to reliability. In practice, these coefficients are often used to

quantify the degree of stability (trait effects), variability (situational influences and person-situation interactions) as well as reliability (overall precision) of the measures. Greater levels of consistency indicate that a measure reflects a more trait-like construct. Greater levels of occasion-specificity indicate that a measure reflects a more state-like construct. Greater levels of reliability indicate that a measure contains less measurement error. An advantage of M-LST models is that they can be used to identify subpopulations that differ with regard to their levels of consistency, occasion-specificity, and/or reliability. Using the M-LST framework, the coefficients can be computed separately for each measure and each class.

M-LST Tutorial

Empirical Example

We now present an illustrative application of an M-LST analysis to parent and teacher reports of children's inattention. The inattention construct represents a subset of symptoms of the larger attention deficit hyperactivity disorder (ADHD) and is characterized by age-inappropriate behaviors, including difficulty listening, failing to pay attention to details in various settings, difficulty organizing tasks, failing to finish tasks, and becoming easily distracted (American Psychiatric Association, 2013).

Sample. Data on inattention were gathered from first-grade children from 30 elementary schools across the Balearic Islands and Madrid, Spain. Children's levels of inattention were evaluated by mothers, fathers, and teachers across three waves of assessment. Overall, $N = 811$ children had at least partial data at one of the three time points. For mother reports, $n = 801$; for father reports, $n = 728$; and for teacher reports, $n = 790$. The sample consisted of 54% boys with the average age of the children at the first assessment being 7 years. Children were excluded from the study if they had an official diagnosis by a school or health official of a learning or behavior disorder at the initial assessment. The study had low levels of missingness (93%, 92%

and 89% of participants had partially complete data at the first, second, and third assessments). For the purposes of the present tutorial, we ignored the nested structure of the data. In an actual substantive application, researchers should account for the clustering of observations by using multilevel or other appropriate modeling techniques.

Measure. The measure of inattention used in this tutorial is a nine-item ADHD-inattention subscale of the Child and Adolescent Disruptive Behavior Inventory (CADBI; Burns & Lee, 2010, 2011). Parents and teachers were asked to evaluate children's symptoms on a 6-point Likert scale, where *0 = nearly occurs none of the time (e.g., 2 or fewer times per month)* and *5 = nearly occurs all the time (e.g., many times per day).* Items were combined to create three composite, continuous parcels, each containing three items, with a composite score ranging from 0 to 5 with 16 possible values (0, 0.33, 0.67, 1, … 5; Burns et al., 2014). Because the MTMS model assumes indicators are continuous, researchers should use continuous indicators. If indicators are item-level, we recommend following appropriate methods for parceling item-level data (see Little, Cunningham, Shahar, & Widaman, 2002; Little, Rhemtulla, Gibson, Schoemann, 2013) to create continuous indicators.[2] Wave 1 was collected toward the end of spring semester of the first grade, wave 2 was collected six weeks later, and wave 3 was collected 10.5 months later at the end of the second grade.

Data were positively skewed, such that the sample contained more children with lower levels of inattention than children with higher levels of inattention. Full information maximum likelihood estimation (Enders, 2010) with robust standard errors (MLR) was used to include all available data and to account for non-normality.

[2] LST models have been developed to account for polytomous indicators (Eid, 1996). These models have not yet been applied to a mixture distribution framework.

This dataset is well-suited to illustrate the M-LST approach for various reasons. First, trait levels of inattention among young children have been shown to be relatively stable across short time spans, such as one-year (Faraone et al, 2006; Willcutt et al., 2012). The M-LST approach assumes that trait scores do not change across time for each latent subgroup. This is because in each class, an MTMS model is specified. The MTMS model assumes stability of means across time.[3]

Second, the dataset contains three waves of data and three indicators of inattention in each wave, which fulfills the requirement of having multiple measurement occasions with multiple indicators within each measurement occasion. Third, the data is from a large community-based sample, where some children in the sample are expected to have higher levels of inattention than others (i.e., some children were expected to have inattention scores in the clinical range on the ADHD-inattention symptom dimension while other children were expected to have inattention scores in the moderate range and others were expected to have inattention scores in the very low range, thus no inattention problems). This provides the opportunity to uncover subgroups of children with different symptom levels of inattention, as well as determine whether levels of consistency and occasion-specificity differ or remain the same across these subgroups. Finally, reports of inattention were available from three different methods (mother, father, and teacher reports), which allowed us to examine the replicability of latent classes across different methods.

Modeling Approach

[3] In cases in which constructs show mean change across time, an extended model with a trait-change component would have to be specified in some or all classes. Such hybrid models are beyond the scope of the present tutorial, but have been presented, for example, by Eid & Hoffmann (1998); Geiser et al. (2017); and Steyer et al. (2015) for the single-class case.

Applying M-LST models requires comparing a large number of nested models. Given the complexity of the approach, we present a three-step procedure to evaluate M-LST models. Step 1 is the simplest step and involves estimating a well-fitting and preferably parsimonious single-class LST model. Step 2 serves to determine the number of classes needed to account for population heterogeneity (if any). In addition, Step 2 is used to determine which LST parameters should be assumed to be class-specific versus class-invariant. Step 3 is a replication step, in which additional M-LST models are evaluated across different methods (e.g., sources; in the present project, different methods refer to father and teacher report) to ensure replicability of the class structures found in Step 2. If different methods are not available in a given study, another possibility to examine the replicability of the findings could be to use a second, independent sample for cross-validation.

Table 1 provides a general outline of all possible steps in the modeling approach. Below, we describe each step in detail and also discuss troubleshooting within each step where applicable.

Step 1: Fitting a single-class MTMS model. In Step 1, we evaluated a single-class MTMS model to determine an appropriate baseline model for the M-LST analyses using mother reports of inattention.[4] We chose mother reports because mothers theoretically spend the most time with children as compared to fathers and teachers. Furthermore, mothers evaluated relatively more children (n = 801) than either fathers (n = 728) or teachers (n = 790) in this study. Three manifest indicators (m = 3) were each measured across three occasions (k = 3) per trait. We fit an MTMS model corresponding to Equation 2 (see also Figure 1) to the data. The M*plus* syntax and data for this model is provided in online supplemental materials Appendix A.

[4] Father and teacher models will be discussed in Step 3.

The online supplemental materials can be found at https://osf.io/kj9eg/. This model showed good to excellent fit according to conventional fit statistics, $\chi^2(31, N = 801) = 42.7$, $p = .08$, BIC = 10,285 as well as equivalence testing approaches, $CFI_t = .99$, $RMSEA_t = .04$ (Marcoulides & Yuan, 2017; Yuan et al., 2016), which are inferential rather than descriptive methods for assessing model fit. Results suggest that an LST process accurately described the longitudinal process of the inattention construct in this sample. We therefore used the MTMS model as the baseline model in subsequent M-LST analyses.

Step 2: Fitting the data to multi-class MTMS models. In Step 2, we evaluated various multi-class MTMS models to 1) determine the number of classes to properly account for population heterogeneity (if any), and 2) simultaneously determine which parameters differed across latent classes. The simultaneous aspect of estimating multi-class models required a rather large and comprehensive set of nested analysis models that include various numbers of latent classes ($c = \{2, \ldots, C\}$) as well as various constraints to parameter estimates (i.e., trait means, trait covariances, trait variances, occasion-specific variances, and error variances) across classes.

Nested multi-class models were compared using primarily Bayesian Information Criteria (BIC), which is commonly used when evaluating which mixture model has the best relative fit (Lubke & Muthén, 2005; Lubke & Luningham, 2017; McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007). BIC was chosen as the model fit criterion due to its asymptotic property of correctly selecting the true model if the true model is amongst the set of specified models (Vrieze, 2012). Further, BIC appropriately selects the correct number of classes in sets of more general factor mixture models (Nylund, Asparouhov, & Muthén, 2007) which are related to the present M-LST approach.

Our approach for comparing different models was as follows. We began with a 2-class version of the best-fitting single-class MTMS model. In the first 2-class model, we allowed the trait factor means to vary freely across classes. We constrained all other parameters in this model to be equal across classes. This relatively parsimonious 2-class model differed from the single-class model only in its estimation of two underlying subpopulations with different *means*.

In the second 2-class model, we additionally allowed the trait factor covariances to vary across classes to determine whether the relationship among the indicator-specific traits differed across classes. In the third, fourth, and fifth 2-class models, we additionally allowed the trait factor variances, occasion-specific factor variances, and error variances to vary across classes, respectively. We followed this same sequence for evaluating 3- and 4-class models.

We recommend terminating the model fitting procedure when the best-fitting *c*-class model fits worse than all $c - 1$ class models, unless there is a theory-driven reason to continue estimating additional classes. Further, researchers may consider stopping the model fitting approach if entire sets of models (e.g., all 3-class models) become unstable (i.e., when the best loglikelihood value cannot be replicated for multiple sets of starting values, when models do not converge after a large number of iterations, or when parameter estimates become uninterpretable or have large standard errors).

Step 2a: Fitting 2-class M-LST models. The first multi-class model we fit was a 2-class MTMS model with all parameters constrained equal across classes except for the trait factor means, which were allowed to differ across classes. This model fit the data better than the single-class model in terms of BIC (see Table 2), illustrating heterogeneity within the sample, at least with regard to the trait factor means. Next, we fit a model where both the trait means and the trait covariances were allowed to be class-specific. This model fit better than the model with only class-specific trait means, indicating that the relationship among the indicator-specific traits also

differed across classes. We then continued to estimate 2-class models with class-specific trait factor variances, occasion-specific factor variances, and error variances, respectively.

The best fitting 2-class solution in our application was Model 6. In Model 6, trait means, trait covariances, trait variances, occasion-specific variances, and error variances were all class-specific. Such a model suggests that there are two distinct subpopulations that differ with regard to their average inattention trait levels as well as relative amounts of trait, occasion-specific, and error variance. However, there may be more than two underlying subpopulations in the data. Therefore, we also evaluated models with three classes.

Step 2b: Fitting 3-class M-LST models. We evaluated the same sequence of models for the 3-class solutions. We first evaluated a 3-class model with all parameters constrained equal across classes except for the trait means. This model fit better than the 2-class model with class-specific trait means but did not fit better than any of the other 2-class models. In order to determine whether any of the remaining 3-class models fit the data better than the 2-class models, we continued evaluating 3-class models with class-specific parameter estimates of trait covariances, trait variances, occasion-specific variances, and error variances, respectively.

For the 3-class models using mother reports, the best-fitting model was Model 11, which freely estimated trait means, trait covariances, trait variances, occasion-specific variances, and error variances across classes. Model 11 fit better than all 2-class models, suggesting the presence of at least three distinct subpopulations in the data, all with unique trait factor means and covariances as well as unique trait, occasion-specific, and error variances.

Step 2c: Fitting 4-class M-LST models. Next, we evaluated 4-class M-LST models in the same manner as the 2- and 3-class models. Some of the 4-class solutions did not show proper convergence. In addition, the best loglikelihood value did not replicate for the least restrictive 4-class model even with 15,000 sets of random starting values. Solutions with loglikelihood values

that cannot be replicated should not be interpreted (Bauer & Curran, 2003), as such solutions are likely to represent a local likelihood maximum. Local likelihood solutions may not be trustworthy and may return invalid parameter estimates (for a discussion on proper, yet local solutions, see Li, Harring, & MacReady, 2014). When the best loglikelihood value cannot be replicated for a model, this may also be a sign that too many classes are being extracted and that a simpler class solution is preferable.

Collectively, the estimation problems encountered for some of the 4-class solutions may indicate that a fourth class was not needed for the given data. This interpretation was supported by the fact that none of the 4-class models that showed proper convergence fit better than the best-fitting 3-class model. We therefore report detailed outcomes for the best fitting 3-class model, which was Model 11.[5]

Best-fitting model estimates. The parameter estimates for Model 11 revealed the following classes: one class with very low inattention trait means and non-significant trait and occasion-specific variances (12%), a low inattention trait means class with small but significant trait and occasion-specific variances (57%), and a moderate inattention trait means class with moderate and significant trait and occasion-specific variances (31%; see Table 3). Furthermore, the output revealed a negative occasion-specific variance estimate in the very low trait means class. The occasion-specific variance estimate in question was very close to zero (–.001) and non-significant (p = .601). We therefore assumed this value was truly 0 (Chen, Bollen, Paxton, Curran, & Kirby, 2001), and evaluated Model 11a, which constrained this value to 0. Model 11a

[5] It is possible that a single parameter (e.g., one trait mean or one error variance) differs across classes in the M-LST approach. Further, partial measurement invariance (e.g., Byrne, Shavelson, & Muthén, 1989; Lubke & Neal, 2008) is possible with the M-LST approach. We have not presented a step to examine differences of a single parameter across classes nor a step to examine partial measurement invariance, but it is possible to examine such differences using the present approach. We recommend examining such models only if there is a theoretical or practical reason to do so.

fit slightly better than Model 11. A thorough examination of parameter estimates revealed no practical differences between Models 11 and 11a with regard to class structure or parameter estimates. M*plus* syntax for Model 11a is in online supplemental materials Appendix B.

Model entropy and classification probabilities. Before discussing the parameter estimates of Model 11a, we examined whether this solution contained well separated classes by inspecting model entropy and classification probabilities. Larger values of model entropy, typically values greater than 0.8, indicate well-separated classes (Celeux & Soromenho, 1996). Diagonal classification probabilities approaching 1.0 also support class separation, and provide evidence that observations (i.e., individuals) are placed into their most likely latent class with high certainty. [6] For Model 11a, model entropy was .82, indicating that classes were well separated. Classification probabilities for the most likely class membership were .98 (very low means class), .94 (low means class), and .86 (moderate means class), providing further evidence that the classes were well separated and individuals placed in appropriate latent classes. We thus proceeded to interpret the parameter estimates for Model 11a.

Consistency, occasion-specificity, and reliability estimates. Consistency and occasion-specificity coefficients were calculated to determine whether the construct was more trait- or state-like within each of the classes. Reliability was calculated to evaluate the amount of variance that was not due to measurement error. Estimates of consistency, occasion-specificity, and reliability are shown in Table 3.

In the very low means class (12%), the average reliability estimate was .22, indicating that most variance (.78) in this class was due to random measurement error. A more thorough investigation of results suggested that this class was essentially homogeneous with mean values

[6] For additional classification diagnostics that could be reported using a mixture modeling approach, see Masyn (2013).

close to zero and very small true state (i.e., trait and occasion-specific factor) variance estimates. In fact, all systematic variance components in this class were statistically non-significant. Although consistency and occasion-specificity estimates from this class are shown in Table 3, these estimates should be interpreted with caution as there was very little systematic variance in this class—indicating very high class homogeneity. The consistency, occasion-specificity, and reliability coefficients as defined in LST theory depend on the presence of a non-zero amount of true score variance. Therefore, these coefficients are uninterpretable in a perfectly or close-to-perfectly homogenous subpopulations as the one found here.

In the low means class (57%), the average reliability estimate was .73, and all variances in this class were statistically significant. The average consistency estimate was .40, whereas the average occasion-specificity estimate was .33. Approximately 55% of the true state variance (.55 = .40 / .73) was due to trait effects, whereas 45% of the true state variance (.45 = .33 / .73) was due to occasion-specific effects. These results suggest that mother reports of children's inattention levels reflected a slightly more trait-like than state-like construct in this class.

The moderate means class (31%) showed an average reliability estimate of .84, indicating that this class had the most systematic variance of the three estimated classes. Average consistency was .46 and average occasion-specificity was .38. Both of these values were slightly higher than the low means class estimates of consistency and occasion-specificity. However, in relative terms, approximately 55% of the true state variance was due to trait influences, whereas 45% of the true state variance was due to occasion-specific influences, which mimics the results from the low means class.

In summary, two noteworthy findings emerged from these results. First, the very low means class consisted of a highly homogeneous group of individuals with no significant true inter-individual differences (no variability in the true scores between individuals). The only

source of variability in this class was random measurement error. Second, the moderate means class had a slightly higher amount of reliability, consistency, and occasion-specificity than the low means class, but the relative proportion of variance due to consistency and occasion-specificity was equal across the two classes. These results suggest that, in both the low and moderate means classes, inattention was more "trait-like" than "state-like."

Step 3. Replication using multiple methods. Given that mixture modeling is in part an exploratory method, conclusions drawn from M-LST models should be replicated with data from other observers or independent samples. Replication is a necessary confirmatory step in the M-LST approach, as it is in other mixture modeling approaches (e.g., Lubke & Luningham, 2017). The aim of Step 3 was therefore to replicate the results from Steps 1 and 2 using father and teacher reports of inattention to ensure we obtained similar class structures with similar parameter estimates. Using data from different reporters evaluating the same participants enabled us to cross-tabulate class membership to examine whether participants would be classified similarly across reporters. If a researcher does not have access to multiple sources, an independent sample of participants should be used to replicate the results.

Replication of step 1. The same approach to evaluating mixture LST models was applied to both father and teacher reports of inattention. First, a single-class LST model was fit to both father and teacher reports, resulting in adequate to excellent model fit using conventional fit statistics and fit statistics derived from equivalence testing methods (Marcoulides & Yuan, 2017; Yuan et al., 2016) [fathers: $\chi^2(31, N = 728) = 62.3$, $p = .001$, BIC = 8,615, $RMSEA_t = .05$, $CFI_t = .98$; teachers: $\chi^2(31, N = 790) = 82.3$, $p < .001$, BIC = 10,334, $RMSEA_t = .06$, $CFI_t = .98$].

Replication of step 2. Next, 2-class, 3-class, and 4-class models were evaluated for father and teacher reports of inattention in the same manner as mother reports of inattention. Applicable models that were evaluated for father and teacher reports can be found in Table 2.

Father report results and best-fitting model estimates. For father reports, the model that resulted in the best relative fit was Model 27: a 3-class solution with class-specific trait means, trait covariances, trait variances, occasion-specific variances, and error variances. This was the same model structure that was found for mother reports.

However, unlike the best-fitting model for mother reports, Model 27 contained one very small class (3%) with high means and significantly negative variance estimates. We therefore could not use Model 27 results and instead re-evaluated the model fixing the negative variance estimates to zero. While this high means class was rather interesting from a substantive point of view (i.e., this may represent individuals with clinically significant levels of inattention), significant negative variance estimates are improper parameter estimates. In this smallest class, one trait variance and one occasion-specific residual variance were estimated to be negative and statistically significant. Two error variances were also negative, but not significantly so.

We speculated that the negative variance estimates may indicate the presence of a class of highly homogenous individuals with zero trait and zero systematic occasion-specific variance. Thus, instead of using Model 27 as the final model, we evaluated two variations of Model 27: one in which the trait variance estimates were constrained to 0 in one class (Model 27a), and one in which both the trait and occasion-specific variances were constrained to 0 in one class (Model 27b). Model 27a resulted in the best relative fit and did not produce improper parameter estimates. Thus, we concluded that the final best-fitting M-LST model using father reports was Model 27a.

The parameter estimates in Model 27a showed a relatively similar 3-class solution relative to the best-fitting mother report model: a very low means class (15%) with mostly non-significant trait and occasion-specific variances, a low means class (39%) with small yet significant occasion-specific variances (but zero trait variances), and a moderate means class (46%) with moderate and significant trait and occasion-specific variances (see Table 3). This solution contained a different set of classes than Model 27 in that it did not contain a class with high inattention means.

In Model 27a, entropy was .80, indicating that the classes were well separated. Classification probabilities for the most likely class membership were .96 (very low means class), .90 (low means class), and .90 (moderate means class), providing further evidence that classes were well separated and that individuals were placed into their most likely latent class with high certainty.

In the very low means class (15%), only one trait variance and four error variances were statistically significant. Average reliability was .46, indicating that slightly more than half of variance (.54) in this class was due to random measurement error (see Table 3). Due to the lack of significant trait and occasion-specific variance estimates, we exercised caution when interpreting the average consistency (.24) and occasion-specificity (.22). Similar to the very low means class for mother reports, this class was highly homogeneous with levels of inattention that were practically zero and essentially no systematic variability.

The low means class (39%) was the class with trait variances constrained to zero. All systematic, reliable variance in this class was due to occasion-specificity only. Average reliability (and therefore occasion-specificity) was estimated as .53. Reliability was slightly lower in the father than in the mother low means class.

Results from the moderate means class (46%) showed statistically significant trait, occasion-specific, and error variances. This class had the highest reliability, with an average estimate of .84, mimicking the results from the mothers' moderate means class. Average consistency was .50 while average occasion-specificity was .34. The percentage of reliable variance due to consistency was 60%, which was slightly higher than the percentage of reliable consistency (55%) for the equivalent mother class.

Overall, the father report solution showed a similar class structure compared to the mother report solution: a very low means class that was essentially homogeneous and contained no or very little systematic variability, a low means class with moderately low reliability, and a moderate means class with the highest relative reliability estimates and slightly higher levels of consistency than occasion-specificity. The father report solution, however, also contained no trait variance in the low means class, which was different relative to the mother report solution.

Teacher report results and best-fitting model estimates. For teacher reports, all 2-, 3-, and 4-class models with at least class-specific trait variances showed estimation problems. Specifically, these models did not terminate normally even with 15,000 sets of random starting values. Error messages indicated that there may not have been enough variability to estimate 2-, 3-, or 4-classes while simultaneously estimating class-specific variances.

To further examine this issue, we hypothesized that some of the latent classes were essentially homogeneous with regard to trait variance, similar to what we found for mother and father reports. We evaluated this hypothesis by constraining trait variances to zero within latent classes.

In the first model variation, we evaluated the 2-class model that first showed estimation problems, Model 36, and added a constraint that set the trait variances to zero in one class (Model 36a). Because occasion-specific variances were still constrained equal across the two

classes, we did not evaluate a model that constrained both trait and occasion-specific variances to zero. Model 36a terminated normally and fit better than all other 2-class models.

In the second set of model variations, we evaluated the 3-class model that first showed estimation problems, Model 39, and constrained the trait variances to zero in one class (Model 39a). Model 39a terminated normally, but the loglikelihood value was not replicated even with 15,000 sets of random starting values. We therefore evaluated a second model variation that constrained the trait variances to zero in two latent classes (Model 39b). Model 39b terminated normally and showed a better fit than all other 3-class models.

We finally evaluated the 4-class model that first showed estimation problems, Model 42, with the constraint of trait variances set to zero in one class (Model 42a). This model did not converge. We therefore evaluated a second model variation that constrained the trait variances to zero in two latent classes (Model 42b). This model also did not converge. We then evaluated a third model variation that constrained the trait variances to zero in three latent classes (Model 42c). This model terminated normally and had the best relative fit of all teacher M-LST models.

Due to the estimation problems we encountered with many of the 4-class models, we did not examine more complex 4-class or 5-class models. Thus, the best-fitting teacher model that converged to a proper solution was Model 42c, a 4-class model with trait means and trait covariances freely estimated across classes, occasion-specific and error variances set equal across classes, and trait variances set to zero in all but one class.

In Model 42c, entropy was .80, indicating that classes were well separated. Classification probabilities for most likely class membership for were .96 (very low means class), .72 (low means class), .86 (moderate means class), and .94 (high means class). The very low, moderate, and high means classes had high classification probabilities while the low means class had a

lower classification probability, indicating that this class may not be as clearly defined as the other classes. Overall, these results provide evidence that classes were mostly well separated.

Unlike the mother and father report solutions, the best fitting teacher model was a 4-class solution that contained a very low means class (50%), a low means class (16%), a moderate means class (32%), and a high means class (3%). The very low, low, and high means classes constrained trait variance estimates to 0. These three classes therefore all contained the same average estimates of reliability (.61) and occasion-specificity (.61).

The moderate means class for teachers showed higher indicator reliability (.93) than either the mother or father moderate means class. This class also showed higher average levels of consistency (.73) and lower average levels of occasion specificity (.20) than either the mother or father moderate means class. Relatively speaking, 78% of the true state variance in this class was due to trait influences, whereas only 22% was due to occasion-specific influences.

Summarizing and Comparing the Best-Fitting Model Results across Informants

Table 3 summarizes the class solutions for the three types of informants (for the entire set of unstandardized parameter estimates, see online supplemental materials Appendix C). Three classes with similar trait means and trait variances emerged across informants: (1) a class with very low trait means and non-significant (or constrained to 0) trait variances, (2) a class with low trait means and low (or constrained to 0) trait variances, and (3) a class with moderate trait means and moderate trait variances. The M-LST solution for teacher reports also contained a class with high trait means, and trait variances constrained to 0.

Most children in the estimated mother report solution fell into either the low means class (57%) or moderate means class (31%), while the fewest children were assigned to the very low means class (12%). Similar to the mother report solution, most children in the father report models fell into either the low means class (39%) or the moderate means class (46%), whereas

the fewest children were assigned to the very low means class (15%). Unlike the mother and father report solutions, most children in the teacher report solution fell into the very low trait means class (50%), and fewer children were assigned to classes with higher levels of inattention (see Table 3). Notably, the teacher solution was the only final solution to estimate a fourth class that contained a very small percentage of children (3%) with high inattention trait means. Results seem to indicate that the estimated M-LST solutions between mothers and fathers were relatively similar. In contrast, the estimated M-LST solutions for parents versus teachers seemed less comparable, as indicated by differences in the estimated number of latent classes as well as the class sizes.

Cross-tabulation of predicted class membership. To determine whether the M-LST solutions for mother, father, and teacher reports showed a significant amount of convergent validity, we estimated class membership for all individuals based on their most likely class assignment in each of the three solutions. Table 4 shows the results from a cross-tabulation analysis. Results showed a strong and highly significant association of the class membership between mother and fathers, χ^2_{MF} (4, N = 724) = 361.5, p < .001, Cramér's V = .50, mothers and teachers, χ^2_{MT} (6, N = 780) = 119.8, p < .001, Cramér's V = .28, and fathers and teachers, χ^2_{FT} (6, N = 712) = 125.4, p < .001, Cramér's V = .30. Further examination of the cross-tabulation results led to the following conclusions:

1. Children who were assigned to the mothers' very low symptoms class were likely to be assigned to the fathers' or teachers' very low symptoms classes.
2. Children who were assigned to the fathers' very low symptoms class were likely to be assigned to the teachers' very low symptom class.

3. Children who were assigned to the teachers' moderate symptoms class were likely to be assigned to the fathers' moderate symptoms class.
4. Children who were assigned to the teachers' high symptoms class were likely to be assigned to the mothers' and fathers' moderate symptoms classes.
5. Children who were assigned to the fathers' very low symptoms class were *unlikely* to be assigned to the mothers' or teachers' moderate symptoms classes.
6. Children who were assigned to the fathers' moderate symptoms class were *unlikely* to be assigned to the mothers' very low symptoms class.
7. Children who were assigned to the teachers' high symptoms class were *unlikely* to be assigned to the mothers' or fathers' low symptoms classes.
8. Children who were assigned to the teachers' low, moderate, or high symptoms classes were *unlikely* to be assigned to the mothers' or fathers' very low symptoms classes.

These results seem to support the notion that mother and father solutions are quite comparable. The results also support the notion that parent and teacher solutions are relatively comparable, despite the fact that one additional class emerged based on teacher reports.

Comparing reliability, consistency, and occasion-specificity across reporters. Across mother, father, and teacher solutions, reliability, consistency, and occasion-specificity contained some notable similarities. In the very low means classes, both the mother and father solutions showed very low reliability and uninterpretable consistency and occasion-specificity. In the low means classes, father and teacher solutions contained somewhat low levels of reliability and zero trait variance due to necessary model constraints. In the moderate means class, mother, father, and teacher solutions all contained relatively high levels of reliability, the mother and father solutions contained similar levels of consistency and occasion-specificity, and all three solutions showed that inattention was more trait- than state-like.

Although results were for the most part similar across reporters, there was one notable difference across the mother, father, and teacher solutions with regard to consistency, occasion-specificity, and reliability estimates. Namely, the teacher moderate means class had relatively higher levels of consistency than either the father or mother moderate means class.

Conclusions from the application. Overall, results supported the replicability of the M-LST solution across methods. However, there were also notable differences between the three different solutions, particularly differences in (1) the number of estimated latent classes between teacher and mother/father solutions, (2) the class probabilities of the very low means class between teacher and mother/father solutions, and (3) the consistency and occasion-specificity of the moderate means class between the teacher and mother/father solutions. Although the specific estimates between mother and teacher solutions seemingly differed, the overall placement of individuals within different classes was relatively consistent across mother and teacher solutions. Further, results were replicated to a large extent between mothers and fathers.

Discussion

In order to evaluate the trait- and state-like aspects of psychological attributes, researchers often employ LST models (Cole et al., 2005; Courvoisier et al., 2007; Geiser & Lockhart, 2012; Prenoveau, 2016; Schermelleh-Engel et al, 2004) which are derived from LST theory (Steyer et al., 1992; Steyer et al., 1999; Steyer et al., 2015). An extension of LST models to heterogeneous populations are M-LST models. M-LST models (Courvoisier et al., 2007) allow researchers to evaluate differences in trait means, trait variances, occasion-specific variances, and error variances across previously unknown latent subpopulations. Thus, M-LST models have the potential to uncover subgroups of individuals who differ with regard to consistency, occasion-specificity, and reliability. Other mixture models that evaluate longitudinal processes are readily used by applied researchers (e.g., growth mixture models; Bauer & Curran,

2008; Muthen & Muthen, 2000), but M-LST models, to our knowledge, have not been applied beyond their initial presentation by Courvoisier et al. (2007).

The M-LST approach requires estimating many latent variable models and comparing model fit indices, which can be cumbersome and requires knowledge of how to implement latent variable models using appropriate software (e.g., M*plus*). In this article, we described a multi-step approach that facilitates the application of the M-LST approach. We also provide M*plus* syntax in online supplemental materials Appendices A and B that researchers can use in their own applications.

We provided a step-by-step modeling procedure to evaluate M-LST models with guidelines for applying M-LST models, troubleshooting M-LST models, replicating M-LST models, and a general structure for reporting M-LST model results beyond what was provided in the original Courvoisier et al. (2007) article. In our application of the M-LST approach, we found that the guidelines provided by Courvoisier et al. (2007) were relatively clear, but did not instruct on 1) comparing nested models to evaluate class-specific versus class-invariant parameters, 2) how to model essentially homogeneous subpopulations, 3) what steps to take if a model does not converge, and 4) how to meaningfully replicate results. We have addressed these topics throughout the modeling approach to more directly guide researchers in their application of the M-LST approach.

To illustrate the step-by-step procedure, we applied the M-LST approach to a dataset containing mother, father, and teacher reports of children's levels of inattention. We found that the M-LST solution could be replicated well across mother and father solutions. The best fitting teacher solution showed some differences in class sizes as well as differences in the relative amounts of reliability, consistency, and occasion-specificity when compared to the mother report solution. A cross-tabulation analysis of class membership showed a significant association

between the class assignments for different informants, indicating that class membership in the mother solution was related to class membership in the father and teacher solutions. In spite of the differences between solutions (especially between mother and teacher solutions), individuals were likely to be placed into similar classes across reporters. If we had found highly discrepant results across our replications, we would not have trusted that our results showed a true mixture solution, and we recommend not interpreting models for which solutions cannot be replicated.

Troubleshooting

Some challenges researchers may face when using the M-LST approach include non-replicated log-likelihood values, obtaining output that includes improper parameter estimates, or encountering models that do not converge to a solution at all. We ran into each of these challenges in our example of the M-LST approach, more often with a larger number of classes containing class-specific parameters. We propose a few strategies researchers may use to address these challenges.

Loglikelihood non-replication. Many models estimated for the present tutorial required several additional runs due to non-replicated loglikelihood values. Models cannot be meaningfully interpreted (and should thus not be included in the comparison procedure) without a replication of the best loglikelihood value (Bauer & Curran, 2003). For most of the 2- and 3-class models that we evaluated, the loglikelihood value was replicated after 1,000 or 5,000 starts. However, with the more complex 3-class and many of the 4-class models, the loglikelihood value was often not replicated even after 15,000 starts.

Non-replicated loglikelihood values are often a result of too few start values or an unidentifiable model. We recommend researchers increase the number of starts up to 15,000 (the maximum used in this paper), which is easily done in M*plus* using the starts command. If the

best loglikelihood value is still not replicated, this may indicate that the model is not well-defined for the data at hand and that simpler models should be used.

Improper parameter estimates. Should an improper parameter estimate occur, we encourage researchers to examine their data, as well as the theory driving their research, to determine why such a result might have occurred. Improper solutions can occur for various reasons, including sampling fluctuations, empirical underidentification, and model misspecification (Chen et al., 2001). Improper estimates should be appropriately addressed in M-LST models. We implemented model constraints with mother, father, and teacher models to estimate solutions that contained proper parameter estimates. Specifically, when we encountered negative variance estimates, we evaluated the potential cause of the improper estimate and whether the estimate was significant or non-significant. For non-significant negative estimates, we simply constrained that specific variance estimate to zero. For significant negative variance estimates, we determined whether the negative variance estimate was due to within-class homogeneity. It is likely that some classes will contain no trait or systematic (trait + occasion-specific) variance because of perfect within-class homogeneity. We encourage researchers to examine both the resulting parameter estimates in addition to theory to guide how best to handle improper parameter estimates.

Non-Convergence. Model misspecification may lead to models not converging. Should a model not converge, this may indicate that the model is over-parameterized or otherwise misspecified. We recommend incrementally simplifying the model by reducing the number of parameters if appropriate. If even a relatively parsimonious M-LST model does not converge, researchers should consider the possibility that the data may not be well-suited for the M-LST approach. Perhaps there is no substantial population heterogeneity to model.

Conclusion

With the rise of mixture modeling approaches in addition to the more prominent use of LST models in the social science literature (Geiser & Lockhart, 2012), it seems reasonable that researchers would ask questions that only M-LST models can answer. M-LST models are unique in their ability to determine whether trait- and state-like influences differ across unknown subgroups. The application of M-LST models does not come without challenges, many of which we address in this tutorial. We present this tutorial not as a perfect example of M-LST analysis, but rather as an instructive guide to aide researchers in applying M-LST analyses to their own data. We hope that readers will find this tutorial and our stepwise modeling approach helpful in applying M-LST models to their own data.

References

American Psychiatric Association. (2013). Neurodevelopmental Disorders. *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338-363.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. John Wiley & Sons.

Burns, G. L., & Lee, S. (2010). Child and Adolescent Disruptive Behavior Inventory–Teacher Version 5.0. *Pullman, WA: Author.*

Burns, G. L., & Lee, S. (2011). Child and Adolescent Disruptive Behavior Inventory–Parent Version 5.0. *Pullman, WA: Author.*

Burns, G. L., Servera, M., del Mar Bernad, M., Carrillo, J. M., & Geiser, C. (2014). 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*, *26*(4), 1247-1258.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468-508.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 558-577.

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*(1), 3-20.

Courvoisier, D. S., Eid, M., & Nussbeck, F. W. (2007). Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods, 12*(1), 80-104.

Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C., & Cole, D. A. (2008). Analyzing the convergent and discriminant validity of states and traits: Development and applications of multimethod latent state-trait models. *Psychological Assessment, 20*, 270-280.

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research—Online, 1*, 65–85.

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics, 23*, 193-215.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent state-trait models with autoregressive effects: Insights from LST-R theory. *European Journal of Psychological Assessment, 33*, 285-295.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Publications.

Faraone, S. V., Biederman, J., & Mick, E. (2006). The age-dependent decline of attention deficit hyperactivity disorder: A meta-analysis of follow-up studies. *Psychological Medicine, 36*(2), 159-165.

Geiser, C., Hintz, F. A., Burns, G. L., & Servera, M. (2017). Latent variable modeling of person-situation data. In J. F. Rauthmann, R. Sherman, & D. C. Funder (Eds.), *The Oxford handbook of psychological situations*. New York: Oxford University Press. Advance online publication. doi: 10.1093/oxfordhb/9780190263348.013.15

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A., & Koch, T. (2015). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavioral Research, 47*, 172-203.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods, 17*(2), 255-283.

Grimm, K. J., Mazza, G. L., & Mazzocco, M. M. (2016). Advances in methods for assessing longitudinal change. *Educational Psychologist*, *51*(3-4), 342-353.

Kenny, D. A. & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*, 52-59.

Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research, 18*, 416-441.

Li, M., Harring, J. R., & Macready, G. B. (2014). Investigating the feasibility of using Mplus in the estimation of growth mixture models. *Journal of Modern Applied Statistical Methods*, *13*(1), 31.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151-173.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285-300.

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

Lubke, G. H., & Luningham, J. (2017). Fitting latent variable mixture models. *Behaviour Research and Therapy, 98*, 91-102.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21-39.

Lubke, G. H., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(1), 26-47.

Lubke, G. H., & Neal, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43*(4), 592-620.

Marcoulides, K. M., & Yuan, K. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(1), 148-153. doi: 10.1080/10705511.2016.1225260

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. A. Little (Ed.), *The Oxford handbook of quantitative methods* (pp. 551–611). New York, NY: Oxford University Press.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavioral Genetics, 16*(1), 163–200.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882-891.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569.

Prenoveau, J. M. (2016). Specifying and interpreting latent state-trait models with autoregression: An illustration. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 731-349.

Schermelleh-Engel, K., Keith, N., Moosbrugger, H., & Hodapp, V. (2004). Decomposing person and occasion-specific effects: An extension of latent state-trait theory to hierarchical LST models. *Psychological Methods, 9*, 198-219.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*(2), 79–98.

Steyer, R., Mayer, A., Geiser, C, & Cole, D. A. (2015). A theory of states and traits—revised. *Annual Review of Clinical Psychology, 11*, 71-98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389-408.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, *17*(2), 228–243. http://doi.org/10.1037/a0027127

Willcutt, E. G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A meta-analytic review. *Neurotherapeutics, 9*, 490-499.

Yuan, K., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319-330. doi: 10.1080/10705511.2015.1065414

Table 1. Applying Mixture LST Models: A Step-by-Step Guide

Step	Description
1	Determine a well-fitting single-class LST model.
2	Evaluate *c*-class (where $c = \{2, \ldots, C\}$) versions of the best-fitting single-class LST model. Simultaneously determine whether estimates of trait means, trait covariances, trait variances, occasion-specific variances, and/or error variances differ across classes.
2a	Evaluate 2-class models.
2b	Evaluate 3-class models.
2c	Evaluate additional *c*-class models. End the Step 2 when no *c*-class model fits better than the best-fitting *c*-1-class model.
2d	Troubleshooting: Re-evaluate any model which did not have a replicated loglikelihood value, did not converge, or contained improper parameter estimates.
3	Apply Steps 1 and 2 using a different method measuring the same construct.
3a	Compare the best-fitting models from Steps 2 and 3.
3b	Cross-tabulate most likely class membership from best-fitting models.

Note. This modeling approach is recommended in the application of M-LST models unless researchers have reason to evaluate different models. We do not present here how to find the most appropriate single-class LST model and refer interested readers to Steyer et al. (2015) for an overview of LST modeling approaches.

Table 2. M-LST Model Fit Information

Model #	Model Description	BIC
	Mother Reports	
1	1-class model	10285
	2-class models	
2	- Trait means vary	10103
3	- Trait means and correlations vary	9978
4	- Trait means, correlations, and trait variances vary	9847
5	- Trait means, correlations, trait variances, and occasion-specific variances vary	9507
6	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	9217
	3-class models	
7	- Trait means vary	10024
8	- Trait means and correlations vary	9882
9	- Trait means, correlations, and trait variances vary	9731
10	- Trait means, correlations, trait variances, and occasion-specific variances vary	9168
11	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	8742 [a]
	4-class models	
12	- Trait means vary	9997
13	- Trait means and correlations vary	9851
14	- Trait means, correlations, and trait variances vary	9708

15	- Trait means, correlations, trait variances, and occasion-specific variances vary	9708
16	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	LL
	Troubleshooting: Addressing non-significant negative variance estimates	
11a	- Model 11 with negative occasion-specific variance set to 0 in one class	8735
	Father Reports	
17	1-class model	8615
	2-class models	
18	- Trait means vary	8470
19	- Trait means and correlations vary	8409
20	- Trait means, correlations, and trait variances vary	8288
21	- Trait means, correlations, trait variances, and occasion-specific variances vary	8031
22	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	7840
	3-class models	
23	- Trait means vary	8489
24	- Trait means and correlations vary	8366
25	- Trait means, correlations, and trait variances vary	8207

26	- Trait means, correlations, trait variances, and occasion-specific variances vary	8057
27	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	7861[a]
	4-class models	
28	- Trait means vary	8407
29	- Trait means and correlations vary	8356
30	- Trait means, correlations, and trait variances vary	8206
31	- Trait means, correlations, trait variances, and occasion-specific variances vary	LL
32	- Trait means, correlations, trait variances, and occasion-specific variances, and error variances vary	LL
	Troubleshooting: Addressing significant negative variance estimates	
27a	- Model 27 with trait variances set to 0 in one class	7619
27b	- Model 27 with trait and occasion-specific variances set to 0 in one class	7798
	Teacher Reports	
33	1-class model	10334
	2-class models	
34	- Trait means vary	9945
35	- Trait means and correlations vary	9906
36	- Trait means, correlations, and trait variances vary	DNT*

	3-class models	
37	- Trait means vary	9739
38	- Trait means and correlations vary	LL
39	- Trait means, correlations, and trait variances vary	DNT
	4-class models	
40	- Trait means vary	9628
41	- Trait means and correlations vary	9570
42	- Trait means, correlations, and trait variances vary	DNT
	Troubleshooting: Addressing no within-class variation	
36a	- Model 36 with trait variances set to 0 in one class	9385
39a	- Model 39 with trait variances set to 0 in one class	LL
39b	- Model 39 with trait variances set to 0 in two classes	9310
42a	- Model 42 with trait variances set to 0 in one class	DNT
42b	- Model 42 with trait variances set to 0 in two classes	DNT
42c	- Model 42 with trait variances set to 0 in three classes	9294

Note. All parameters were constrained equal across classes unless otherwise noted in the Model Description. [a] = This model contained improper estimates and was the best-fitting model, so it was re-evaluated in the troubleshooting section with appropriate model constraints (see text for more details); **Bold** = final best fitting model; DNT = model did not terminate; LL = loglikelihood not replicated after a maximum of 15,000 starts.

Table 3. Consistency, Occasion-Specificity, and Reliability Estimates across Classes for Each Method.

Class Description	%	Trait Means	Trait Consistency *Con*	Occasion-Specificity *OSpe*	Reliability *Rel*
			Mother Report		
Very low means	12%	.07 [.06, .08]	.07 [.03, .14][b]	.15 [.00, .36][b]	.22 [.04, .43]
Low means	57%	.74 [.65, .81]	.40 [.34, .49]	.33 [.27, .40]	.73 [.61, .86]
Moderate means	31%	1.96 [1.75, 2.15]	.46 [.37, .57]	.38 [.31, .45]	.84 [.75, .93]
			Father Report		
Very low means	15%	.11 [.09, .14]	.24 [.03, .50][b]	.22 [.03, .59][b]	.46 [.08, .85]
Low means[a]	39%	.59 [.50, .64]	.00 [.00, .00]	.53 [.40, .70]	.53 [.40, .70]
Moderate means	46%	1.71 [1.56, 1.85]	.50 [.41, .66]	.34 [.21, .44]	.84 [.76, .92]
			Teacher Report		
Very low means[a]	50%	.15 [.07, .19]	.00 [.00, .00]	.61 [.16, .93]	.61 [.16, .93]
Low means[a]	16%	.79 [.54, .98]	.00 [.00, .00]	.61 [.16, .93]	.61 [.16, .93]
Moderate means	32%	1.75 [1.38, 2.11]	.73 [.55, .91]	.20 [.02, .41]	.93 [.89, .97]
High means[a]	3%	4.38 [4.22, 4.60]	.00 [.00, .00]	.61 [.16, .93]	.61 [.16, .93]

Note. Reported values represent the average estimates with the range in brackets. [a] = Estimated trait variances were constrained to 0 within this class. [b] = Variance estimates to calculate these values were all or mostly non-significant; caution should be taken when interpreting these results. Model entropy for the mother solution = 0.82. Model entropy for the father solution = 0.80. Model entropy for the teacher solution = .80.

Table 4. Crosstab Results Comparing Mother, Rather, and Teacher M-LST Most Likely Class Membership

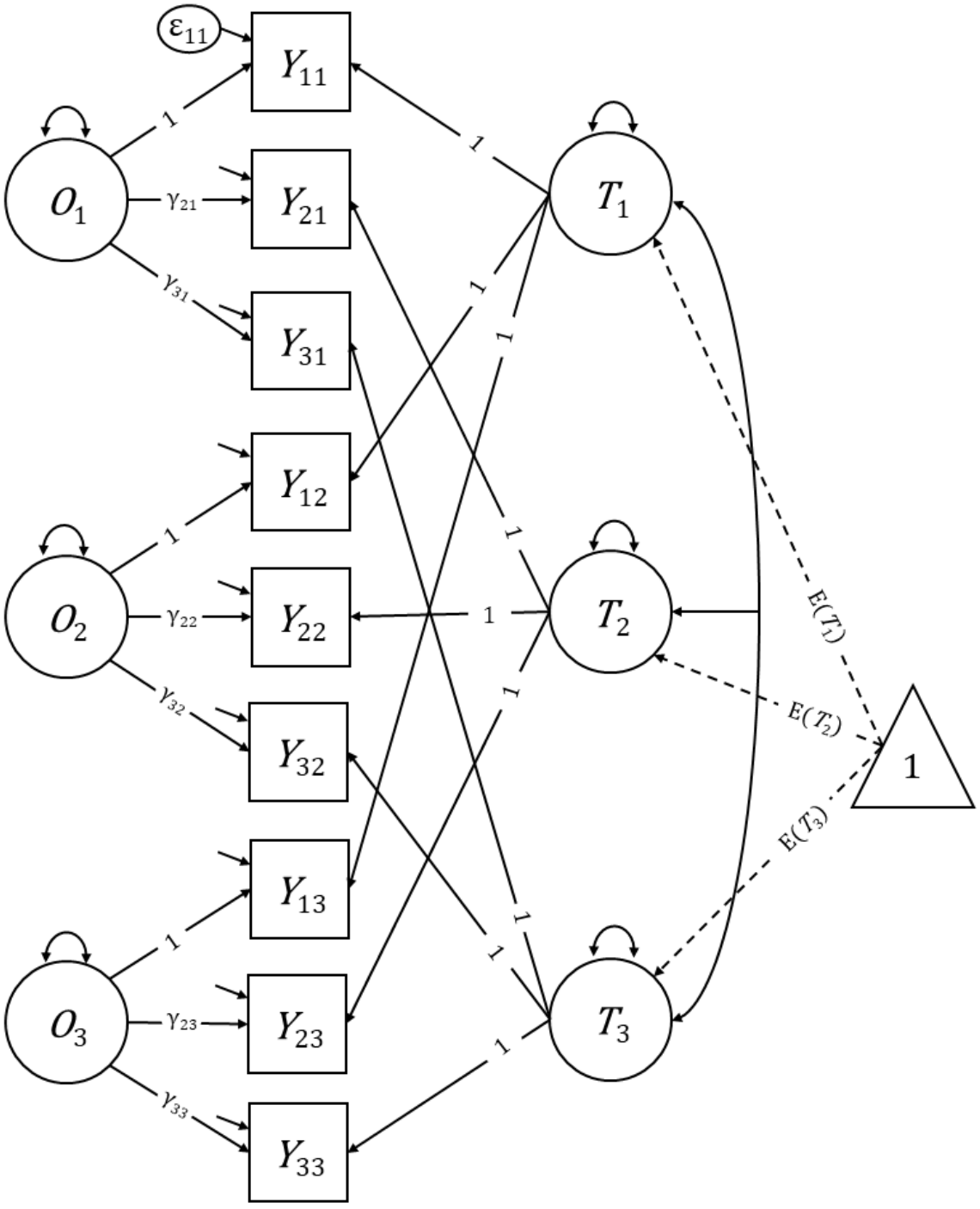
Class Description		Very low means	Low means	Moderate Means	High Means	*Total*
			Fathers			
Very low means		**60 (.67)**	24 (.27)	5 (.06)		*89*
Low means	Mothers	45 (.11)	**233 (.55)**	150 (.35)		*426*
Moderate means		5 (.02)	27 (.13)	**175 (.85)**		*209*
Total		*110*	*284*	*330*		*724*
			Teachers			
Very low means		**74 (.79)**	7 (.07)	13 (.14)	0 (.00)	*94*
Low means	Mothers	**248 (.55)**	74 (.17)	122 (.27)	3 (.01)	*447*
Moderate means		65 (.27)	39 (.16)	**114 (.48)**	21 (.09)	*239*
Total		*387*	*120*	*249*	*24*	*780*
			Teachers			
Very low means		**88 (.84)**	8 (.08)	9 (.09)	0 (.00)	*105*
Low means	Fathers	**171 (.61)**	44 (.16)	65 (.23)	2 (.01)	*282*
Moderate means		98 (.30)	57 (.18)	**150 (.46)**	20 (.06)	*325*
Total		*357*	*109*	*224*	*22*	*712*

Note. Values represent the raw number of estimated individuals within each latent class. Values in parentheses represent the proportion of individuals per row. **Bolded** values indicate the cell with the highest proportion per row.

Figure Captions

Figure 1. Single-Class Multitrait-Multistate Latent State-Trait Model. The model includes three indicators measured across three occasions, though more or less indicators and occasions can be implemented in practice.

Figure 1


ε_{11}
Y_{11}
Y_{21}
Y_{31}
Y_{12}
Y_{22}
Y_{32}
Y_{13}
Y_{23}
Y_{33}
O_1
O_2
O_3
γ_{21}
γ_{31}
γ_{22}
γ_{32}
γ_{23}
γ_{33}
T_1
T_2
T_3
$E(T_1)$
$E(T_2)$
$E(T_3)$
1