Structural Equation Modeling of Multiple-Indicator Multimethod-Multioccasion Data: A Primer

Christian Geiser and Fred Hintz

Utah State University

G. Leonard Burns

Washington State University

Mateu Servera

University of Balearic Islands

Author Note

Christian Geiser and Fred Hintz, Department of Psychology, Utah State University. G. Leonard Burns, Washington State University. Mateu Servera, Department of Psychology, University of Balearic Islands.

This 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, and two grants from the Ministry of Economy and Competitiveness of Spanish Government under award number PSI2011-23254 and PSI2014-52605-R (AEI/FEDER, UE). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Citation

Geiser, C., Hintz, F., Burns, G. L., & Servera, M. (in press). Structural equation modeling of multiple-indicator multimethod-multioccasion data: A primer. *Personality and Individual Differences*.

Abstract

We provide a tutorial on how to analyze multiple-indicator multi-method (MM) longitudinal (multi-occasion, MO) data. Multiple-indicator MM-MO data presents specific challenges due to (1) different types of method effects, (2) longitudinal and cross-method measurement equivalence (ME) testing, (3) the question as to which process characterizes the longitudinal course of the construct under study, and (4) the issue of convergent validity versus method-specificity of different methods such as multiple informants. We present different models for multiple-indicator MM-MO data and discuss a modeling strategy that begins with basic single-method longitudinal confirmatory factor models and ends with more sophisticated MM-MO models. Our proposed strategy allows researchers to identify a well-fitting and possibly parsimonious model through a series of model comparisons. We illustrate our proposed MM-MO modeling strategy based on mother and father reports of inattention in a sample of $N = 805$ Spanish children.

Keywords: longitudinal confirmatory factor analysis, multitrait-multimethod, multiple reporters, measurement equivalence, structural equation modeling, multistate model, latent variables, inattention

Introduction

More and more personality researchers collect and study longitudinal data, that is, data from more than one measurement occasion. Such "multi-occasion" (MO) data is useful to examine developmental processes, changes in personality across time, or the effects of experimental manipulations, events, or treatments on personality constructs. In addition, personality researchers have been encouraged to collect data from multiple methods (multiple sources, multiple reporters, multiple informants, observations, indirect measures, physiological measures, etc.) to examine construct validity, potential discrepancies between different methods, and obtain more detailed information on the relevant constructs. The idea of collecting and studying multimethod (MM) data goes back to Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach to construct validation, which has been extremely influential in nearly all areas of psychology (Eid & Diener, 2006).

The statistical modeling of both longitudinal and MM data is complex and comes with specific challenges. Even greater complexities arise when MM and MO designs are combined such that data is obtained from multiple reporters or other methods on each of multiple measurement occasions (so-called MM-MO designs). This is particularly true when multiple observed variables (indicators; e.g., items, tests, or questionnaire subscales) are used at each measurement occasion. The use of multiple indicators has been recommended for MM as well as MO data, because multiple-indicator data allows researchers to test assumptions (i.e., regarding measurement equivalence) that cannot be tested with single-indicator data (e.g., Marsh, 1993; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Cole, Martin, & Steiger, 2005).

Over the past 10 to 15 years, a number of latent variable approaches have been proposed that combine the advantages of longitudinal (MO) studies with the advantages of MM approaches (e.g., Cole & Maxwell, 2003; Courvosier, Nussbeck, Eid, Geiser, & Cole, 2008; Crayen, Geiser,

Scheithauer, & Eid, 2011; Geiser, 2009; Geiser, Eid, Nussbeck, Courvoisier, & Cole, 2010a, b; Grimm, Pianta, & Konold, 2009; Koch, Schultze, Holtmann, Geiser, & Eid, in press). In the present paper, we provide a tutorial on how to analyze multiple-indicator MM-MO data with structural equation modeling. Below we first introduce the data example and then discuss key modeling issues with MM-MO data.

Data Example

Sample. We use data from a longitudinal study on symptom and impairment dimensions with Spanish children (for details, see Bernad, Servera, Becker, & Burns, 2016; Burns, Servera, Bernad, Carillo, & Geiser, 2014; Burns, Becker, Servera, Bernad, & García-Banda, 2017). For the purposes of the present tutorial, we analyzed mother (Method 1) and father reports (Method 2) of $N = 805$ Spanish children's inattention (IN) symptoms over three measurement occasions (i.e., spring semesters of the first, second, and third grade with each assessment separated by 12 months). In order to reduce the level of complexity of the analyses and output, we focus on a *single construct* (IN). Generalizations of the described modeling approaches to more than one construct are discussed in Geiser et al. (2010a, b).

Measure. We used the 9-item IN subscale of the Child and Adolescent Disruptive Behavior Inventory (CADBI; Burns et al., 2014). Two example items of the IN subscale are: "Has difficulty keeping attention focused on homework or home activities (e.g., difficulty remaining focused during homework, chores, conversations, lengthy readings)" and "Fails to give close attention to details or makes careless mistakes in homework or home activities (e.g., overlooks or misses details; work is inaccurate)." The nine IN symptoms were rated on a 6-point scale (i.e., 0 = *almost never* [never or about once per month), 1 = *seldom* [about once per week], 2 = *sometimes* [several times per week], 3 = *often* [about once per day], 4 = *very often* [several times per day], and 5 = *almost always* [many times

per day]). Mothers and fathers were instructed to base their ratings on the child's behavior in the home and community and to make their ratings independently.

To obtain multiple indicators for each type of reporter and measurement occasion, we assigned the nine items to three parcels (with each parcel representing the average of three items) following procedures recommended by Little (2013) to create homogenous parcels. On each measurement occasion, the corresponding parcels for mothers and fathers consisted of the same items. Furthermore, the parcels were composed of the same items on each measurement occasion. Full information maximum likelihood estimation in the Mplus statistical software (version 7.4, Muthén & Muthén, 1998-2010) was used for all analyses. The Mplus input/output files for all models can be found in the supplemental materials.

Specific Modeling Issues With MM-MO Data

There are several peculiarities of multiple-indicator longitudinal MM-MO data that make the statistical modeling of such data challenging. These issues concern different forms of method effects that may arise in MM-MO studies, the question of measurement equivalence (ME) across time and methods, and the question as to whether and how the construct under study changes over time. Below we describe each issue in detail. Subsequently, we illustrate a modeling strategy for addressing each issue based on the data example described above.

Method effects. In longitudinal multiple-indicator MM-MO data, there are at least two potential sources of method effects:

(1) Method effects due to the repeated use of the same measures (observed variables, indicators; e.g., item, questionnaire, or test scores) across time within each method. These types of method effects are often referred to as *indicator-specific effects* (e.g., Marsh &

Grayson, 1994; Raffalovich & Bornstedt, 1987). Indicator-specific effects are due to the fact that different measures often reflect (at least slightly) different facets of a construct or contain other sources of method variance (e.g., due to a specific response format or item wording that is not shared with the remaining indicators). This is the case even for measures that are supposedly homogenous—in longitudinal studies, it often turns out that a measure is more highly correlated with itself over time than with other measures of the same construct due to indicator-specificity (Cole & Maxwell, 2003).

(2) Method effects due to the use of different methods (e.g., different reporters or raters) within each measurement occasion. Different methods often provide unique perspectives on the constructs to be measured or show specific biases. Even when methods use the same or equivalent measures (e.g., if a parent report questionnaire uses the same items as a self-report questionnaire), different methods typically do not show perfect agreement in their ratings. A specific complexity with MM-MO data arises from the fact that when different methods use the same items, indicator-specific effects due to the use of the same measures may be shared across methods.

Both types of methods effects described above need to be addressed when modeling MM-MO data. Ignoring method effects can lead to bias in the estimated parameters of the model. Furthermore, method effects are usually of substantive interest, as they inform researchers about convergent validity and method specificity between different methods of measurement.

Measurement equivalence across time and methods. Another important issue in longitudinal modeling in general concerns the degree of *measurement equivalence* (ME; aka *measurement invariance*) of the indicators across time. ME refers to the extent to which the measures are related in the same way to the underlying latent variables on each measurement occasion (Raykov, 2006). A

sufficient degree of ME is required for a meaningful comparison of, for example, latent means across time.

Widaman and Reise (1997) distinguish between *configural*, *weak*, *strong*, and *strict ME*. *Configural invariance* only requires that the number of factors and the pattern of factor loadings be the same on each occasion of measurement. *Weak ME* (or *metric invariance*) holds if—in addition to configural invariance—the factor loadings remain the same over time. *Strong ME* (*scalar invariance*) additionally requires the observed variable measurement intercepts to be time-invariant. The condition of *strict factorial invariance* is satisfied when the measures additionally show constant measurement error variances across time.

In practical applications, it is often desirable to use measures that show at least strong ME (constant loadings and intercepts). The reason is that strong ME ensures that the latent variables are in the same metric at each time point (i.e., the origin and units of measurement are the same at each measurement occasion under strong ME). Strong ME therefore allows for meaningful comparisons of latent variable variances and means across time.

In MM-MO studies that use equivalent measures (e.g., questionnaires or items) across methods, ME can not only be studied *across time*, but also *across methods* (Geiser, Burns, & Servera, 2014). If at least strong ME is established across methods, this means that the latent variables are measured in the same metric across methods. Therefore, strong ME enables meaningful comparisons of latent means and variances *across methods*. Establishing ME not only across time but also across methods is thus advantageous, because it allows researchers to study the convergent validity of latent mean levels, latent variances (the extent of true individual differences), and latent covariances across methods. This enables researchers to evaluate between-method similarities and differences in more detail beyond just evaluating correlations of the same measures across methods as in the classical Campbell and Fiske

(1959) approach. For example, it can be important for researchers to know whether different methods reflect the same mean levels of a given construct. The additional ME constraints across methods (if tenable) also lead to an even more parsimonious model (fewer loading, intercept, or error variance parameters are estimated in a model that features ME constraints).

Longitudinal process. Researchers often want to find out whether the construct under study is stable or changes across time. Below we present a general longitudinal structural equation model that allows capturing both state variability (short-term reversible fluctuations) and potential trait changes (more long-term and enduring changes; Nesselroade, 1991). For a detailed discussion of more specific longitudinal models that correspond to state variability and/or trait change processes see Geiser, Hintz, Burns, and Servera (in press). In the Discussion section, we refer to relevant extensions of these models to MM data.

Modeling Strategy and Results

In our example multiple-indicator MM-MO data set for the IN construct, we are dealing with a 3 (indicators) × 2 (methods) × 3 (measurement occasions) measurement design. We are thus modeling the means, variances, and covariances of 18 observed variables (see Table 1 regarding the descriptive statistics for all indicators). To simplify the analysis, we recommend beginning with separate analyses for each method (Step 1). After identifying well-fitting and parsimonious models for each method, we proceed to Step 2 in which the models from Step 1 are combined into a common model and more sophisticated MM-MO models are examined.

Step 1: Analyze Separate MO Models For Each Method

It is useful to simplify the analysis of complex MM-MO data by first examining separate longitudinal models for each method (and, in the case of multiple traits, by analyzing separate models

for each trait-method combination). In this way, the homogeneity (unidimensionality) of indicators, the question of measurement equivalence (ME) across time as well as the underlying longitudinal process (e.g., state variability or trait changes) can first be studied separately for each method and well-fitting models can be established for each method.

A useful starting point in Step 1 is to fit a so-called *latent state* (LS) or *multistate model* (Geiser et al., 2010a; Steyer, Ferring, & Schmitt, 1992) separately for each method. Fitting a series of LS models to each method separately allows (1) examining potential method effects due to indicator heterogeneity, (2) testing for ME across time within each method, and (3) studying the type of longitudinal process for each method through examining latent variable parameters such as LS factor means, variances, and covariances/correlations. The LS model is an excellent baseline model for these types of analyses, because contrary to other longitudinal models, it has a simple and saturated (unrestricted) latent variable covariance and mean structure. Therefore, it allows researchers to focus on measurement-related issues first (i.e., the question of indicator homogeneity and ME) before examining more complex latent variable structures.

An example of a single-method LS model for three indicators and three measurement occasions is depicted in Figure 1. In the LS model, there is a single LS variable at each time point. In our example, each LS variable represents true individual differences in the IN construct at a given measurement occasion. The LS variables are measured and identified by (1) using multiple (at least two) indicators at each time point (typically, the same variables are used at each time point as this is critical for establishing ME and comparability across time; in our case, the same three item parcels served as indicators at each time point), (2) fixing the loading of one indicator to 1, and (3) fixing the intercept of the same indicator to 0. Fixing the loadings and intercepts of this so-called reference indicator (the first indicator in Figure 1) allows identifying the variances and means of the LS variables

at each time point. When only two indicators are available at each time point, model identification additionally requires that each LS variable be substantially correlated with at least one other LS variable in the model or with at least one external variable that is added to the model.

In most cases, LS variables have substantial, positive correlations with at least some of the remaining LS variables, thus ensuring identification even with only two indicators. This is because most psychological variables show at least some stability of individual differences across time—this stability is reflected in the LS variable correlations. For more state-like constructs (e.g., hormone levels, mood), such stabilities may be relatively low, however, especially when measurement occasions are widely spaced out (e.g., more than one year interval between measurement occasions). In this case, it would be wise to collect data for at least three indicators on each measurement occasion to avoid potential identification problems.

We recommend starting with an LS model as shown in Figure 1 in Step 1 that places no constraints on measurement parameters (loadings, intercepts, and error variances) other than the reference indicator restrictions needed to identify the model as described above (Model 1.1). Model 1.1 assumes invariance of measurement parameters only for the reference indicator (for which the loadings and intercepts are fixed to the same values on all measurement occasions) and otherwise represents a model of configural ME. (The constraints for the reference indicator serve to identify and conveniently interpret the LS factor means and variances, but do not represent testable ME restrictions.) If Model 1.1 shows a reasonable model fit, this model can be used in subsequent analyses of higher levels of ME (weak, strong, and strict invariance). According to Schermelleh-Engel, Moosbrugger, and Müller (2003), a structural equation model can be considered to show a good approximate fit when the ratio of chi-square statistic to its degrees of freedom (*df*) is ≤ 2 or at least ≤ 3, the root mean square error of approximation (RMSEA) is at or below .05, the comparative fit index (CFI) is at or above .95, and the

standardized root mean square residual (SRMR) is at or below .05 (similarly criteria were proposed by Hu & Bentler, 1999). We furthermore considered the chi-square difference test for nested model comparisons and Akaike's information criterion (AIC) for nested and non-nested model comparisons. The model with the smallest AIC is considered the best-fitting model.

In our example, Model 1.1 did not show a very good fit for either mother or father reports. This was shown especially by a large ratio of chi-square value to *df* of > 5 for both types of reporters. In addition, RMSEA values were larger than the desirable cut-off value of .05 for both mothers and fathers. Although other fit indices (SRMR, CFI) indicated a decent fit, the chi-square values and RMSEA coefficients provided evidence that Model 1.1 was too simplistic for the given data and that additional effects in the data should be modeled.

When Model 1.1 does *not* fit well as in the present example, this is often a sign that indicators are not perfectly homogenous (i.e., that they do not measure exactly the same construct due to method or indicator-specific effects). In this case, a model that adds indicator-specific factors can be examined next (Model 1.2). Model 1.2 is depicted in Figure 2. Model 1.2 addresses method effects due to more or less heterogeneous indicators (differences in item content, wording, response format etc.). This model uses a reference indicator (the first indicator Y_{1t} in the figure) to identify the LS variables at each time point. For this reference indicator, the factor loadings and intercepts are fixed to 1 and 0, respectively, as in Model 1.1. In addition, the second and third indicators Y_{2t} and Y_{3t} load onto additional indicator-specific (method) factors IS_i. The IS_i factors are residual factors that represent specific variance in Indicator 2 and 3 that is (1) not shared with the reference indicator and (2) stable across time (i.e., shared by the same indicator across measurement occasions).

In Model 1.2, the LS factors are thus specific to the reference indicator and have an additional index for Indicator 1 (S_{1t}). The choice of the reference indicator is not completely trivial, especially

when indicators are very heterogeneous and represent rather distinct facets of a construct. In this case, the indicator that most closely reflects the intended construct (i.e., a "marker indicator") should be chosen as reference indicator. For indicators that are essentially homogenous and show only small indicator-specific effects, the choice of the reference indicator is typically (more or less) arbitrary: With homogenous indicators that can be seen as random representations out of the same universe of indicators for a given construct, different choices of the reference indicator are unlikely to produce dramatically different results.

All IS_i factors, being defined as residual factors, have means of zero. Latent means are thus estimated only for the LS factors in Model 1.2. The IS_i factor variances are identified by fixing at least one factor loading to 1 for each IS_i factor. Given that the factors are measured by the same variables at each time point, it is often reasonable (and parsimonious) to fix all IS_i factor loadings to 1 at all time points (as was done in the present example), although this is not necessary for identification as long as there are three or more time points. For only two time points, one loading per IS_i factor can still be estimated freely (if necessary) as long as each IS_i factor is substantially correlated with at least one other factor or variable in the model.

For even slightly heterogeneous indicators, Model 1.2 typically shows a substantially better global model fit than Model 1.1. In our example, the chi-square values were much lower for Model 1.2 for both mothers and fathers (see Table 2). Furthermore, descriptive fit measures also indicated that Model 1.2 should be preferred over Model 1.1. This showed the presence of at least some degree of indicator heterogeneity. The estimates of IS_i factor loadings and variances obtained from Model 1.2 can be examined to study the degree of heterogeneity of indicators in more detail. Squared standardized IS_i factor loadings give the proportion of stable indicator variance that is not shared with the reference indicator and thus are a measure of indicator heterogeneity (method effects at the indicator [parcel]

level), whereas the squared state factor loadings give the proportion of (non-reference) indicator variance that is shared with the reference indicator.

In our example, both the IS_i factor loadings and variances were significantly different from zero. Their sizes (compared to the state factor variances and loadings) were relatively small, however, indicating that indicators were relatively homogenous and did not differ strongly in terms of the underlying constructs. This was expected given that the parcels were constructed to be rather homogenous. Squared standardized loadings revealed that only between 5% and 9% percent of the indicator variance was due to stable indicator specific effects, whereas between 77% and 86% percent of the indicator variance reflected state variance shared with the reference indicator. In summary, even though indicator-specific effects were relatively minor in our example, Model 1.2 exhibited a substantially better fit than Model 1.1. Indicator-specific effects observed in Model 1.2, although small, were statistically significant and not completely negligible. Therefore, we chose to use Model 1.2 rather than Model 1.1 in subsequent analyses of ME for each of the two methods.

Next, we examined ME across time by estimating constrained versions on Model 1.2. Model 1.2 did not place any restrictions on factor loadings, intercepts, or residual variances and can thus be seen as a model of configural ME across time (the same indicators loaded onto the state factors at each measurement occasion). Model 1.3 was a version of Model 1.2 in which the factor loadings for the same indicators were set equal across time (i.e., a model with time-invariant state factor loadings for all indicators). Model 1.3 is nested within Model 1.2 and reflected the assumption of weak ME across time. Nested models can be compared via chi-square difference tests. A *significant* chi-square difference value indicates that the more constrained model (here: Model 1.3) fits significantly worse than the less constrained model (here: Model 1.2). In this case, the additional constraints are not tenable. On the contrary, a *non-significant* chi-square difference value means that there is no

statistically significant decline in fit when making the additional constraints. In this case, the more constrained (and therefore more parsimonious) model is preferred.[1]

In our example, the fit indices (see Table 2) revealed that Model 1.3 did not fit significantly worse than Model 1.2, indicating that the assumption of weak ME across occasions did not have to be rejected for either mother or father reports. We therefore tested for strong ME across time next (Model 1.4: equal loadings *and* intercepts across time for all indicators across occasions). Model 1.4 did not fit significantly worse than Model 1.3 for either mother or father reports, thus allowing us to retain the assumption of strong ME across time for both types of reporters. Model 1.5 added the constraint of equal error variances across time for all indicators (strict ME). This assumption had to be rejected for both mother and father reports according to the chi-square difference test, indicating that the amount of measurement error variance changed in one or more indicators between at least two time points. We therefore based subsequent equality tests of latent variable parameters on the less constrained Model 1.4 rather than Model 1.5. Strong ME as represented by Model 1.4 is sufficient to examine latent variable means, variances, and covariances across time.

Step 1 also serves to identify the longitudinal process reflected in each method. The LS model provides some answers to the question as to whether a construct is characterized by state variability or trait changes in each method. These answers can be obtained by fitting LS model versions that place restrictions on latent variable parameters, that is, LS factor means, variances, and potentially covariances. For example, if an LS model with equal LS factor means across time fit the data well, this would indicate mean stability of the construct within that method, which could point to a process of state variability (no trait changes).

[1] Other authors (e.g., Little, 2013) have proposed more complex criteria for testing invariance constraints in longitudinal studies. Comparing different criteria is beyond the scope of the present tutorial.

In Model 1.6, we tested whether the latent state factors had equal means on all measurement occasions. Model 1.6 thus reflected the assumption of mean stability across time. Model 1.6 was not rejected for mother reports, but did show a significantly worse fit compared to Model 1.4 for father reports. An examination of the parameter estimates in Model 1.4 for father reports revealed that the latent state factor means at Time 1 and 3 were somewhat larger than the state factor mean at Time 2. This small difference in average levels of IN was not reflected in mother reports.

Subsequent models with equal state factor variances were rejected for both mother and father reports indicating that the amount of true individual differences in IN was not constant across time for either mother or father reports. An examination of the parameters estimates obtained in Model 1.6 for mothers revealed a larger state factor variance at Time 1 compared to Time 2 and 3. Model 1.4 for fathers revealed that state factor variances were similar at Time 1 and 3, but higher than the state factor variance at Time 2. An inspection of the LS factor correlations in Model 1.6 (mothers) and Model 1.4 (fathers) revealed substantial stability of individual differences across time according to both types of reporters (LS factor correlation range for mothers: [.72; .79], for fathers: [.68; .82]).

In summary, Step 1 revealed the presence of small indicator-specific effects in both mothers' and fathers' reports of IN in children showing that parcels were essentially, but not perfectly homogenous. Furthermore, Step 1 allowed us to determine the degree of ME across time for each of the two methods. For both mother and father reports of IN, strong ME across time could be assumed, allowing us to perform meaningful comparisons of latent state factor means and variances across time. These structural comparisons revealed that mother reports showed mean stability across time, indicating no average IN symptom change across time. Father reports indicated similar average levels of IN at Time 1 and 3 but a slightly lower level at Time 2. Correlations between LS factors across time

were substantial for both types of reporters indicating considerable stability of individual differences across time.

Step 2: Analyze Combined MM-MO Models

In the next step, we combined the best-fitting single-method LS models from Step 1 into a comprehensive MM-MO models for the IN construct. When equivalent measures are used across methods as in the present example, the level of ME across methods can be examined first in Step 2. As explained previously, establishing ME not only across time, but also across methods, allows for interesting comparisons of latent variable parameters across methods (Geiser et al., 2014) and can lead to a more parsimonious model.

Combined MM LS models also allow researchers to study the convergent validity across methods by estimating correlations of LS factors across methods and by comparing structural parameters (latent means, variances, and covariances) across methods. High LS factor correlations across methods for the same time point indicate agreement between reporters (the presence of convergent validity), whereas low correlations indicate a lack thereof. Furthermore, more sophisticated combined MM models can be estimated later in Step 2 that directly contrast different methods against a reference method. This is useful to obtain coefficients of convergent validity and method specificity for each indicator in the form of proportions of explained variance (R^2). We first discuss the extension to a MM LS model and then present an additional, more sophisticated MM-MO model that allows more detailed comparisons across methods.

MM-LS models. In our example, we combined Model 1.6 for mother reports and Model 1.4 for father reports from Step 1 into a MM LS model (Model 2.1, see Figure 3A). In this model, all LS factors were allowed to correlate and all indicator-specific factors were allowed to correlate, but no

correlations were allowed between any LS factors and indicator-specific factors. This model showed a decent chi-square/*df* ratio of 2.81 as well as good RMSEA, SRMR, and CFI values.

The estimates obtained from Model 2.1 revealed perfect correlations between the indicator-specific factors for mother and father reports for both Parcel 2 and 3 [latent correlation estimates of 1.04 (*SE* = 0.09) and 1.06 (*SE* = 0.10)]. This indicated that parcel-specific effects for Indicator 2 and 3 indicator were perfectly homogenous (unidimensional) across mother and father reports. We therefore tried a more parsimonious model version (Model 2.2) in which the indicator-specific factors for corresponding parcels were collapsed into a single factor across reporters, respectively. That is, the simplified Model 2.2 contained only one indicator-specific factor for Parcel 2 and one indicator-specific factor for Parcel 3 that applied to both types of reporters (see Figure 3B). Given greater parsimony, Model 2.2 had an improved chi-square/*df* ratio of 2.6 and fit better than Model 2.1 according to the AIC. Other fit indices were similarly good for both Models 2.1 and 2.2, so that we decided to continue tests of ME with the more parsimonious Model 2.2.

Model 2.3 tested whether indicator-specific factor loadings were equal across mother and father reports by fixing all indicator-specific factor loadings to 1. The slightly improved chi-square/*df* ratio of 2.55 and non-significant chi-square difference test indicated that this assumption was tenable. Other fit indices also supported the more parsimonious Model 2.3. In Model 2.4, we tested for equal state factor loadings across mother and father reports (reflecting the assumption of weak ME across methods). Model 2.4 had a similar chi-square/*df* ratio of 2.52 and did not fit significantly worse than Model 2.3 according to the chi-square difference test. Model 2.4 also showed a lower AIC value than Model 2.3 and was supported by the remaining fit indices.

Next, we tested for equal intercepts across mother and father reports (Model 2.5; reflecting the assumption of strong ME across methods). This constraint was not rejected by the chi-square

difference test. Although the AIC was slightly higher for Model 2.5 than for Model 2.4, the remaining fit indices (including a chi-square/*df* ratio of 2.52) were similar for both models. We therefore decided to retain the more parsimonious Model 2.5.

In Model 2.6, we set error variances equal across reporters for corresponding parcels (reflecting the assumption of strict ME across methods). The chi-square difference test indicated that this constraint led to a significant decline in fit. We therefore preferred Model 2.5, assuming strong (but not strict) ME across methods. As stated previously, strong ME is sufficient for meaningful comparisons of LS factor means and variances. In the subsequent Model 2.7, we tested for equal LS factor means across methods for Time Points 1 and 3 (Step-1 analyses had already revealed that means for fathers differed slightly between Times 1 and 3 vs. Time 2). Model 2.7 did not fit significantly worse than Model 2.5 according to the chi-square difference test and had a decent fit overall (including a chi-square/*df* ratio of 2.52). This indicated that mother and father reports in general showed convergent validity with regard to mean levels of IN in this sample, with the exception of Time 2, for which father reports indicated a slightly lower level of IN (0.92 vs. 0.97 on the 0 to 5 point response scale). This mean difference was very small however, and likely does not represent a substantively meaningful difference. Note that convergent validity of mean levels does not mean that reporters also showed perfectly correlated LS scores—we address the question of convergent validity with respect to the LS factor correlations in a later section below.

Next, we tested for equal LS factor variances across reporters. Equal LS factor variances across methods would indicate that different methods reflect the same amount of true score variability. Our Step-1 analyses had already revealed that equal state factor variances could not be assumed across time within each method. Therefore, we only tested whether state factor variances were equal across reporters at the same time point (Model 2.8). Model 2.8 did not fit significantly worse than Model 2.7

and had a decent overall fit. This indicated that equal state factor variances across methods could be assumed at each time point. Mother and father reports thus showed convergent validity also with respect to the amount of true individual differences in IN at each time point. Therefore, it made sense to also test for equal covariances of LS factors across reporters (Model 2.9). This assumption was also tenable according to the chi-square difference test and global fit measures. We accepted Model 2.9 as the final combined MM-LS model. The good fit of Model 2.9 indicated that mother and father reports also showed convergent validity with regard to the overtime correlations of LS factors. In other words, mother and father reports reflected the same amount of stability of individual differences in IN across time (state factor correlations for both reporters: Time 1 with Time 2: .80; Time 2 with Time 3: .76; Time 1 with Time 3: .70).

An examination of the correlations between corresponding mother-father LS factors that were measured at the same time point revealed very high convergent validity between mothers and fathers at each measurement occasion. The three correlations were .84 (Time 1), .83 (Time 2), and .79 (Time 3). This indicated that the rank order of individuals with regard to true IN scores was rather similar across reporters at each time point—mother and father consistently showed high agreement in their reports.

CS-C(*M* – 1) models. One way to examine the degree of convergent validity in more detail is by means of a more sophisticated MM-MO model that contrasts the reports of one (or several) methods against a reference method. This allows researchers to determine variance components due to shared construct variance (convergent validity relative to the reference method), method specificity (discrepancies of a given method relative to the reference method), and measurement error (unreliability) in each indicator. This possibility is offered by a model called the correlated states-correlated (methods – 1) or CS-C(*M* – 1) model (Geiser, 2009; Geiser et al., 2010b).

The CS-C(M – 1) model requires researchers to select one method as reference prior to the analysis (Eid et al., 2003). The choice of the reference method is usually based on the following considerations (for a more detailed discussion and guidelines as to the choice of an appropriate reference method, see Geiser, Eid, & Nussbeck, 2008). First, if there is a method that is (1) already known (or hypothesized) to be highly valid or (2) most established or trusted in a given field (so-called "gold-standard" method), then this method is selected as reference (e.g., an objective measure of a construct such as a test or physiological measure versus more subjective measures such as self- and other reports). In other cases, there may be no gold standard available, but a specific contrast between methods (e.g., self- versus other reports of behavior) may be most interesting or relevant to the research problem at hand. In the case of just two methods, the choice of the reference method is relatively arbitrary from the point of view of variance components, as these will give similar answers in either case.

A path diagram of a CS-C(M – 1) model with general state factors for the present design with two methods, three time points, and three indicators per time point is depicted in Figure 4. In the picture, the first method (i.e., mother reports) was selected as reference method. The second method (i.e., father reports) served as non-reference method. The LS factors in Figure 4 are defined by the reference method and therefore reflect true individual differences in IN as reported by mothers. The father report indicators also load onto the reference LS factors and in addition have loadings on method factors M_{mt}. The M_{mt} factors are defined as residual factors with respect to the corresponding state factors S_{11t}. Therefore, the method factors M_{mt} capture systematic variance in the non-reference (father) indicators that is shared by all father report indicators at a given time point, but *not* shared with the mother indicators at the same time point. Variance shared with the mother report indicators

(convergent validity) is reflected in the factor loadings of the father indicators on the mother report state factors.

One loading per method factor is fixed to 1 for identification on each measurement occasion. Method factors, defined as residuals, have means of zero. Therefore, as in previous models, latent means are only estimated for the LS factors. Furthermore, method factors are not allowed to correlate with corresponding state factors that pertain to the same construct and time point. Indicator-specific factors IS_{im} are specified in the same way and have the same meaning as in the previous MM-LS model. These factors reflect indicator-specific effects in mother and father indicators relative to a reference indicator. The IS_{im} factors are not allowed to correlate with either the LS or method factors M_{mt}.

An alternative way to specify a CS-C(M – 1) model is to use *indicator-specific* instead of *general* LS factors. In CS-C(M – 1) model with indicator-specific LS factors (see Figure 5), indicator-specific effects are reflected in the fact that each indicator has its own LS variable. Therefore, separate indicator-specific factors IS_{im} are not required in this model. The CS-C(M – 1) model with indicator-specific LS factors can be a useful alternative to the model with general LS factors when different methods use the same indicators as in the present example. An advantage of the CS-C(M – 1) model with indicator-specific LS factors is that it does not require indicator-specific effects to be homogeneous (unidimensional) across measurement occasions. Instead, indicator-specific LS factors can simply be freely correlated and do not require a unidimensional structure across time. Another advantage is that separate LS factors are specified for each indicator. Latent means, variances, and covariances can thus be estimated separately for each LS factor for the reference method. Therefore, the model is particularly useful when indicators reflect rather distinct facets of a construct for which general state factors are less meaningful.

In the CS-C(*M* – 1) model with indicator-specific state variables, convergent validity and method-specificity can be quantified in terms of the *consistency* and *method-specificity coefficients* (Eid et al., 2003; Geiser et al., 2010b). The consistency coefficient for an observed variable (indicator) is given by the squared standardized state factor loading for that indicator and reflects the proportion of variability in that indicator that is shared with the reference method. The method specificity coefficient is given by the squared standardized method factor loading and reflects the proportion of variability in a non-reference indicator that is due to method effects (systematic variability that is not shared with the reference method). Consistency and method-specificity coefficients add up to the reliability coefficient, which gives the proportion of total reliable (systematic) variability in an indicator. Consistency and method-specificity coefficients can also be calculated for the underlying true score variables. This can be accomplished by dividing the observed variable consistency and method-specificity coefficients for a given indicator by the reliability estimate for that indicator. That way, one can more easily examine which proportion of the true score (overall systematic variance) is shared with the reference method versus method-specific.

In our example, both the CS-C(*M* – 1) model with IS_{im} factors (Model 2.10) and the CS-C(*M* – 1) model with indicator-specific LS factors (Model 2.11) fit the data well (see Table 2). According to the AIC, the latter model had the best fit of all models considered. We therefore present detailed outcomes for Model 2.11. Tables 3 and 4 contain the relevant parameter estimates for this model. Table 5 gives the consistency, method specificity, and reliability coefficients. The CS-C(*M* – 1) model revealed that mother and father reports showed fairly strong convergent validity on all three measurement occasions. This can be seen from the high standardized loadings of the father report indicators on the reference state factors and the comparatively low standardized loadings on the father report method factors. The consistency and method specificity coefficients in Table 5 show that

between two-thirds and three-fourths of the true score variance in father reports was shared with mother reports. Only between a fourth and a third of the true score variance in father reports was method-specific in the sense that it was *not* shared with mother reports.

Correlations of method factors across measurement occasions (range: [.45; .50], see Table 4) indicated moderate stability of method effects (deviations from mother reports) over time. This means that there was a tendency for fathers who overestimated a child's IN symptoms on one measurement occasion to also overestimate this child's IN symptoms on another measurement occasion. However, the stabilities of the method effects were relatively modest, indicating that there was some change in method effects as well.

Summary of the Analysis

In summary, our detailed analyses of longitudinal mother and father report IN data with multiple MM-MO models revealed the following findings:

- Mother and father reports of IN showed high levels of convergent validity in different domains: There was strong ME across reporters indicating that the IN latent variables were measured in the same metric for mothers and fathers. Mother and father reports resulted in identical latent IN means on two out of the three measurement occasions as well as identical latent variances on each measurement occasion. Mother and father reports also resulted in identical stability estimates (latent state correlations) across time. In addition, mother and father reports shared a large portion of true score variance on all three measurement occasions.

- Mother and father reports both revealed the same high level of stability of true individual differences across time. This indicated that IN may be a rather trait- than state-like construct in this age group and population.
- Method effects of father reports (deviations from mother reports) showed moderate stabilities across measurement occasions, indicating that specific father true score variance is somewhat (but far from perfectly) consistent across time.

Outlook: Other Models for MM-MO Data

In subsequent steps, researchers could analyze more specialized MM-MO models. Due to space limitations, we can only briefly review these models, but provide the appropriate references for more detailed descriptions and applications.

State Variability Models. For constructs that represent a longitudinal process of pure state variability (e.g., mood), latent state-trait extension of MM-MO models can be used to determine variance components due to trait, state residual, and measurement error components for different methods as well as for examining the convergent validity of state and trait components across methods. MM-LST models have been presented by Courvoisier et al. (2008); Hintz, Geiser, and Shiffman (2016); Koch et al. (in press); as well as Scherpenzeel and Saris (2007). An application of a MM-LST model to IN can be found in Litson, Geiser, Burns, and Servera (in press).

Change and Growth Models. Investigators are often interested in modeling change over time in more detail. For this purpose, the MM-MO state models presented here can be adapted to include latent change (difference score) variables (Raykov, 1993; Steyer, Eid, Schwenkmezger, 1997). Geiser et al. (2010a) presented an extension of the MM-LS model to an MM change score model. Geiser et al. (2010b) discuss a latent change version of the CS-C(M – 1) model. Both models allow researchers to

examine the convergent validity of change scores. Grimm et al. (2009) presented an extension of the correlated traits-correlated methods (CT-CM) model to a multiple-indicator MM growth curve model. However, their extension uses a single indicator per method rather than multiple indicators for each method. Crayen et al. (2011) discuss a multigroup modeling approach that is based on the CS-C(M – 1) model and can be used when multiple intervention or control groups are compared longitudinally using a multigroup MM-MO design.

Models for more complex MM designs. Koch, Schultze, Eid, and Geiser (2014) presented an extended CS-C(M – 1) model for MM measurement designs that use both structurally different (fixed) and interchangeable (random) methods. A similar extension of Courvoisier et al.'s (2008) MM-LST model is discussed in Koch et al. (in press).

Conclusion

The modeling of multiple-indicator MM-MO data involves a number of specific issues such as examining different types of method effects, testing for ME across time and methods, and studying convergent validity through structural (latent variable) parameters. In the present tutorial, we addressed the question as to how such complex MM-MO data and models can be approached and how a well-fitting and possible parsimonious model can be identified. We hope that our tutorial will be useful to get substantive researchers started on their analysis of MM-MO data.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

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. https://doi.org/10.1037/1082-989X.10.1.3

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*, 558–577.

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

Crayen, C., Geiser, C., Scheithauer, H., & Eid, M. (2011). Evaluating interventions with multimethod data: A structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 497-524.

Eid, M., & Diener, E. (2006). *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple indicator CT-C(*M*–1) model. *Psychological Methods, 8*, 38–60.

Geiser, C. (2009). *Multitrait-Multimethod-Multioccasion Modeling.* München, Germany: AVM.

Geiser, C., Burns, G.L., & Servera, M. (2014). Testing for measurement invariance and latent mean differences across methods: Interesting incremental information from multitrait-multimethod studies. *Frontiers in Psychology*: *Quantitative Psychology and Measurement*, 5:1216. doi:10.3389/fpsyg.2014.01216.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods, 13*(1), 49–57. https://doi.org/10.1037/1082-989X.13.1.49

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010a). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. *Developmental Psychology, 46*(1), 29.

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010b). Multitrait-multimethod change modelling. *AStA Advances in Statistical Analysis, 94*(2), 185-201.

Geiser, C., Hintz, F., Burns, G.L., Servera, M. (in press). Latent variable modeling of person-situation data. In D. C. Funder, J. F. Rauthmann & R. A. Sherman (Eds.), *The Oxford Handbook of Psychological Situations.*

Grimm, K. J., Pianta, R. C., & Konold, T. (2009). Longitudinal multitrait-multimethod models for developmental research. *Multivariate Behavioral Research, 44*(2), 233-258.

Hintz, F., Geiser, C., & Shiffman, S. (2016). *A latent state-trait model for analyzing states, traits, situations, method effects, and their interactions*. Manuscript submitted for publication.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology*, *5*: 311.

Koch, T., Schultze, M., Holtmann, J., Geiser, C., & Eid, M. (in press). A multimethod latent state-trait model for structurally different and interchangeable methods. *Psychometrika*.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford.

Litson, K., Geiser, C., Burns, G. L., & Servera, M. (in press). Trait and state variance in multisource assessments of ADHD and academic impairment. *Journal of Clinical Child and Adolescent Psychology*.

Marsh, H. W. (1993). Multitrait-multimethod analyses: Inferring each trait/method combination with multiple indicators. *Applied Measurement in Education, 6*, 49–81.

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural equation modeling*, *1*, 116–145.

Muthén, L. K. and Muthén, B. O. (1998–2010). *Mplus User's Guide. Fourth Edition*. Los Angeles, CA: Muthén & Muthén.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change. Recent advances, unanswered questions, future directions* (pp. 92-105). Washington, DC: American Psychological Association.

Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance components of factor models: Estimation with longitudinal data. *Sociological Methods & Research, 15*, 385–405.

Raykov, T. (1993). A structural equation model for measuring residualized change and discerning patterns of growth or decline. *Applied Psychological Measurement, 17*(1), 53-71.

Raykov, T. (2006). Examining temporal stability of scale validity in longitudinal studies. *Multivariate Behavioral Research, 41,* 401–426.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research - Online, 8*, 23–74.

Scherpenzeel, A., & Saris, W. E. (2007). Multitrait-multimethod models for longitudinal research. In K. van Montfort, A. Satorra, & H. Oud (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 381–401). New York: Erlbaum.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research - Online, 2*, 21–33.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 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*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* (pp. 281–324). Washington: American Psychological Association.

Table 1. *Correlations, Means, and Standard Deviations for Mother and Father Reports of Inattention (IN)*

	Mother reports									Father reports								
Parcel	T1P1 M	T1P2 M	T1P3 M	T2P1 M	T2P2 M	T2P3 M	T3P1 M	T3P2 M	T3P3 M	T1P1 F	T1P2 F	T1P3 F	T2P1 F	T2P2 F	T2P3 F	T3P1 F	T3P2 F	T3P3 F
T1P1M	—																	
T1P2M	.87	—																
T1P3M	.82	.85	—															
T2P1M	.70	.68	.66	—														
T2P2M	.68	.74	.70	.86	—													
T2P3M	.64	.66	.71	.82	.85	—												
T3P1M	.63	.61	.62	.65	.66	.62	—											
T3P2M	.61	.66	.63	.68	.75	.66	.87	—										
T3P3M	.61	.63	.66	.63	.67	.68	.86	.86	—									
T1P1F	.77	.74	.72	.67	.66	.62	.59	.56	.55	—								
T1P2F	.70	.78	.72	.61	.67	.61	.54	.56	.53	.87	—							
T1P3F	.69	.73	.78	.64	.68	.67	.56	.58	.58	.83	.86	—						
T2P1F	.68	.65	.61	.75	.73	.67	.60	.64	.59	.74	.71	.70	—					
T2P2F	.66	.68	.64	.70	.79	.68	.60	.66	.60	.71	.75	.72	.88	—				
T2P3F	.65	.63	.64	.69	.72	.76	.58	.61	.61	.71	.69	.75	.84	.85	—			
T3P1F	.50	.50	.51	.54	.59	.55	.70	.70	.66	.59	.59	.59	.68	.66	.61	—		
T3P2F	.53	.58	.56	.56	.64	.57	.67	.76	.68	.61	.65	.62	.68	.74	.65	.88	—	
T3P3F	.51	.53	.54	.54	.59	.57	.66	.69	.72	.59	.60	.62	.64	.66	.66	.87	.87	—
M	0.94	1.20	1.10	0.91	1.14	1.06	0.99	1.18	1.12	0.99	1.20	1.13	0.93	1.10	1.05	1.03	1.23	1.14
SD	0.97	1.09	1.05	0.96	1.07	1.03	1.00	1.09	1.06	0.99	1.10	1.06	0.97	1.05	1.02	1.02	1.11	1.07

Note. T = time, P = parcel, M = mother report, F = father report

Table 2. *Goodness of Fit Statistics for Different Models Fit to the Inattention Data*

Model	Description	χ^2	*df*	*p*	RMSEA	SRMR	CFI	AIC	$\Delta\chi^2$	Δdf	$p(\Delta\chi^2)$
				Step 1 (mother report)							
1.1	LS model, configural ME	184.87	24	<.001	.09	0.018	.964	10383			
1.2	Model 1.1 with IS_i factors for Parcel 2 and 3,	24.59	21	.26	.015	0.01	.999	10174			
1.3	Model 1.2 with weak ME	27.23	25	.34	.011	0.013	1	10170	2.64	4	.61
1.4	Model 1.2 with strong ME	35.01	29	.20	.016	0.014	.999	10171	7.77	4	.10
1.5	Model 1.2 with strict ME	50.90	35	.03	.024	0.018	.996	10184	15.97	6	.01
1.6	**Model 1.4 with equal state factor means**	**40.54**	**31**	**.11**	**.02**	**0.019**	**.998**	**10174**	**5.52**	**2**	**.06**
1.7	Model 1.5 with equal state factor variances	47.77	33	.04	.024	0.029	.997	10180	7.23	2	.02
				Step 1 (father report)							
1.1	LS model, configural ME	137.87	24	<.001	.081	.02	.973	8622			
1.2	Model 1.1 with IS_i factors for Parcel 2 and 3,	34.43	21	.033	.030	0.012	.997	8495			
1.3	Model 1.2 with weak ME	42.55	25	.016	.031	0.02	.996	8498	8.12	4	.09
1.4	**Model 1.2 with strong ME**	**47.69**	**29**	**.016**	**.030**	**0.019**	**.996**	**8495**	**5.14**	**4**	**.27**
1.5	Model 1.2 with strict ME	66.19	35	.001	.035	0.022	.993	8510	18.50	6	.005
1.6	Model 1.4 with equal state factor means	59.26	31	.002	.035	0.025	.993	8504	11.58	2	.003
1.7	Model 1.5 with equal state factor variances	56.48	31	.003	.034	0.037	.994	8503	8.79	2	.01
				Step 2 (mother and father reports combined)							
2.1	Models 1.6 (Mothers) and 1.4 (Fathers) combined	326.37	116	<.001	.047	0.018	.979	17235			
2.2	Model 2.1 with general IS_i factors across reporters	335.90	129	<.001	.045	0.018	.980	17224	9.53	13	.73
2.3	Model 2.2 with equal IS_i factor loadings	343.76	135	<.001	.044	0.019	.980	17223	7.86	6	.25

2.4	Model 2.3 with weak ME across reporters	345.58	137	<.001	.043	0.019	.980	17221	1.82	2	.40
2.5	Model 2.4 with strong ME across reporters	350.65	139	<.001	.043	0.019	.979	17222	5.06	2	.08
2.6	Model 2.5 with strict ME across reporters	367.40	148	<.001	.043	0.02	.979	17226	16.74	9	.05
2.7	Model 2.4 with equal LS factor means across reporters [a]	355.73	141	<.001	.043	0.02	.979	17223	5.07	2	.08
2.8	Model 2.7with equal LS factor variances for corresponding time points	358.07	144	<.001	.043	0.021	.979	17219	2.34	3	.50
2.9	**Model 2.8 with equal LS factor covariances across reporters**	**360.11**	**147**	**<.001**	**.042**	**0.022**	**.979**	**17220**	**2.04**	**3**	**.56**
2.10	CS-C(*M* – 1) model with general state and IS_i factors	397.33	137	<.001	.049	0.067	.975	17284			
2.11	**CS-C(*M* – 1) model with indicator-specific state factors**	**210.85**	**91**	**<.001**	**.04**	**0.07**	**.988**	**17123**			

Note. LS = latent state. IS = indicator specific. CS-C(*M* – 1) = correlated state-correlated (methods minus one). RMSEA = root mean squared error of approximation. SRMR = standardized root mean square residual. CFI = comparative fit index. AIC = Akaike's information criterion. Best-fitting models for which detailed outcomes are reported in the text are printed in bold face.

[a] Latent state factor means set equal across reporters on all time points except Time 2.

Table 3. *Parameter Estimates and Standard Errors for Model 2.11*

Parameter	Estimate	*SE*	Standardized estimate
Parcel 1 state factor loadings			
Mother report	1.00 [a, b]	–	.95, .94, .95[c]
Father report	0.93 [b]	0.02	.83, .83, .79 [c]
Parcel 2 state factor loadings			
Mother report	1.00 [a, b]	–	.97, .98, .97 [c]
Father report	0.92[b]	0.02	.84, .83, .80 [c]
Parcel 3 state factor loadings			
Mother report	1.00 [a, b]	–	.95, .95, .96 [c]
Father report	0.93 [b]	0.02	.83, .83, .79 [c]
Method factor loadings			
Parcel 1, father report	1.00 [a, b]	–	.50, .50, .55 [c]
Parcel 2, father report	1.18 [b]	0.04	.50, .50, .55 [c]
Parcel 3, father report	1.04 [b]	0.04	.46, .48, .54 [c]
State factor means			
Parcel 1	.936	0.03	
Parcel 2	1.16	0.04	
Parcel 3	1.07	0.03	
Intercepts			
Mothers and fathers, Time 1 and Time 3	0.00 [a]	–	
Fathers, Time 2 Parcel 1	0.01	0.02	
Fathers, Time 2, Parcel 2	–0.02	0.02	
Fathers, Time 2, Parcel 3	–0.004	0.02	
Factor variances			
State 1 Parcel 1 Mothers	0.85	0.07	
State 1 Parcel 2 Mothers	1.24	0.09	
State 1 Parcel 3 Mothers	1.05	0.08	
State 2 Parcel 1 Mothers	0.74	0.06	
State 2 Parcel 2 Mothers	1.01	0.07	

State 2 Parcel 3 Mothers	0.85	0.06	
State 3 Parcel 1 Mothers	0.76	0.06	
State 3 Parcel 2 Mothers	1.13	0.08	
State 3 Parcel 3 Mothers	0.86	0.07	
Time 1 Father Method Factor	0.27	0.03	
Time 2 Father Method Factor	0.23	0.03	
Time 3 Father Method Factor	0.33	0.05	
Measurement error variances			
Parcel 1, Time 1, Mothers	0.09	0.01	0.10
Parcel 2, Time 1, Mothers	0.07	0.02	0.06
Parcel 3, Time 1, Mothers	0.11	0.02	0.12
Parcel 1, Time 1, Fathers	0.08	0.01	0.07
Parcel 2, Time 1, Fathers	0.09	0.02	0.06
Parcel 3, Time 1, Fathers	0.13	0.02	0.10
Parcel 1, Time 2, Mothers	0.10	0.01	0.12
Parcel 2, Time 2, Mothers	0.05	0.01	0.05
Parcel 3, Time 2, Mothers	0.09	0.01	0.10
Parcel 1, Time 2, Fathers	0.06	0.01	0.06
Parcel 2, Time 2, Fathers	0.07	0.01	0.06
Parcel 3, Time 2, Fathers	0.08	0.01	0.08
Parcel 1, Time 3, Mothers	0.08	0.01	0.10
Parcel 2, Time 3, Mothers	0.08	0.02	0.07
Parcel 3, Time 3, Mothers	0.08	0.01	0.09
Parcel 1, Time 3, Fathers	0.08	0.01	0.07
Parcel 2, Time 3, Fathers	0.08	0.02	0.06
Parcel 3, Time 3, Fathers	0.09	0.02	0.08

Note. Model 2.11 = CS-C(*M* – 1) model with indicator-specific state factors. [a] parameter fixed for identification. [b] parameter set equal across measurement occasions and therefore reported only once. [c] standardized factor loadings were not set equal across time and are reported for each of the three measurement occasions in the order Time 1, Time 2, Time 3. Dashes (–) indicate that a standard error was not computed due to a parameter being fixed rather than freely estimated.

Table 4. *Latent Variable Correlations in Model 2.11*

Factor	S1P1M	S1P2M	S1P3M	S2P1M	S2P2M	S2P3M	S3P1M	S3P2M	S3P3M	Method 1	Method 2	Method 3
S1P1M												
S1P2M	.94											
S1P3M	.92	.93										
S2P1M	.79	.75	.75									
S2P2M	.75	.79	.77	.94								
S2P3M	.74	.73	.79	.91	.92							
S3P1M	.69	.66	.69	.73	.72	.70						
S3P2M	.69	.72	.71	.75	.79	.73	.94					
S3P3M	.69	.68	.73	.72	.72	.75	.94	.93				
Method 1	0[a]	0 [a]	0 [a]	.15	.14	.15	.13	.08	.08			
Method 2	.18	.16	.14	0 [a]	0 [a]	0 [a]	.14	.13	.11	.45		
Method 3	.07	.10	.08	.08	.11	.08	0 [a]	0 [a]	0 [a]	.49	.50	

Note. Model 2.11 = CS-C(M – 1) model with indicator-specific state factors. S = state factor. P = parcel. M = mother report. Method 1 = method factor father reports time 1. Method 2 = method factor father reports time 2. Method 3 = method factor father reports time 3. [a] correlations between state and method factors at the same time point are fixed to zero by definition of the model.

Table 5. *Consistency, Method-Specificity, and Reliability Estimates for Model 2.11*

	Observed Variables			Latent state (true score) variables	
Variable	Consistency	Method specificity	Reliability	Consistency	Method specificity
			Time 1		
Parcel 1, Mothers	.90	—	.90	1.00	.00
Parcel 2, Mothers	.94	—	.94	1.00	.00
Parcel 3, Mothers	.90	—	.90	1.00	.00
Parcel 1, Fathers	.68	.25	.93	.74	.26
Parcel 2, Fathers	.70	.25	.94	.74	.26
Parcel 3, Fathers	.69	.22	.90	.76	.24
			Time 2		
Parcel 1, Mothers	.88	—	.88	1.00	.00
Parcel 2, Mothers	.95	—	.95	1.00	.00
Parcel 3, Mothers	.90	—	.90	1.00	.00
Parcel 1, Fathers	.69	.25	.94	.74	.26
Parcel 2, Fathers	.69	.25	.94	.73	.27
Parcel 3, Fathers	.69	.23	.92	.75	.25
			Time 3		
Parcel 1, Mothers	.91	—	.91	1.00	.00
Parcel 2, Mothers	.93	—	.93	1.00	.00
Parcel 3, Mothers	.91	—	.91	1.00	.00
Parcel 1, Fathers	.62	.31	.93	.67	.33
Parcel 2, Fathers	.64	.31	.94	.68	.32
Parcel 3, Fathers	.63	.29	.92	.68	.32

Note. Mother reports served as reference method. Father reports were contrasted against mother reports. Therefore, method specificities are zero for mother report indicators by definition.

Figures

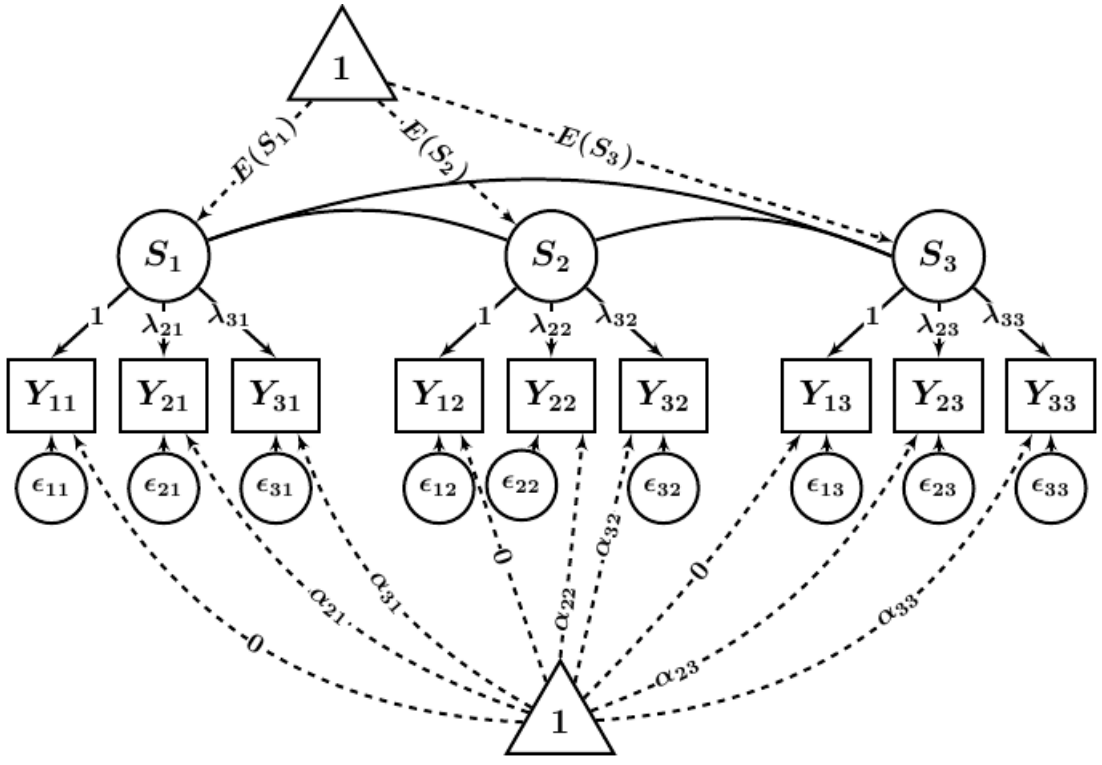

Figure 1. Path diagram of a single-method latent state (LS) model for three indicators and three time points. Y_{it} = observed variable (indicator; i = indicator, t = time point). S_t = common latent state factor. λ_{it} = state factor loading. ε_{it} = measurement error variable. The triangles represent a constant of 1 that is used to add a mean structure to the model [intercepts α_{it} and latent state factor means $E(S_t)$].

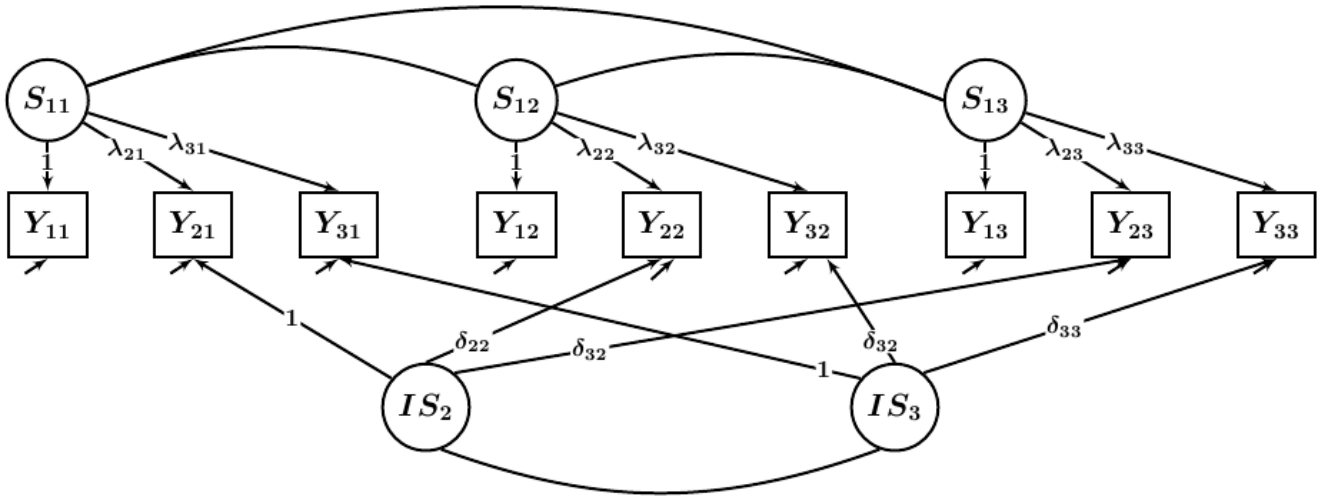

Figure 2. Path diagram of a single-method LS model for three indicators and three time points with indicator specific factors for the second and third indicator. S_{1t} = latent state factor specific to indicator Y_{1t}. IS_i = indicator-specific residual factor. λ_{it} = state factor loading. δ_{it} = indicator-specific factor loading. The mean structure is the same as in Figure 1 and has been omitted from the picture to reduce clutter.

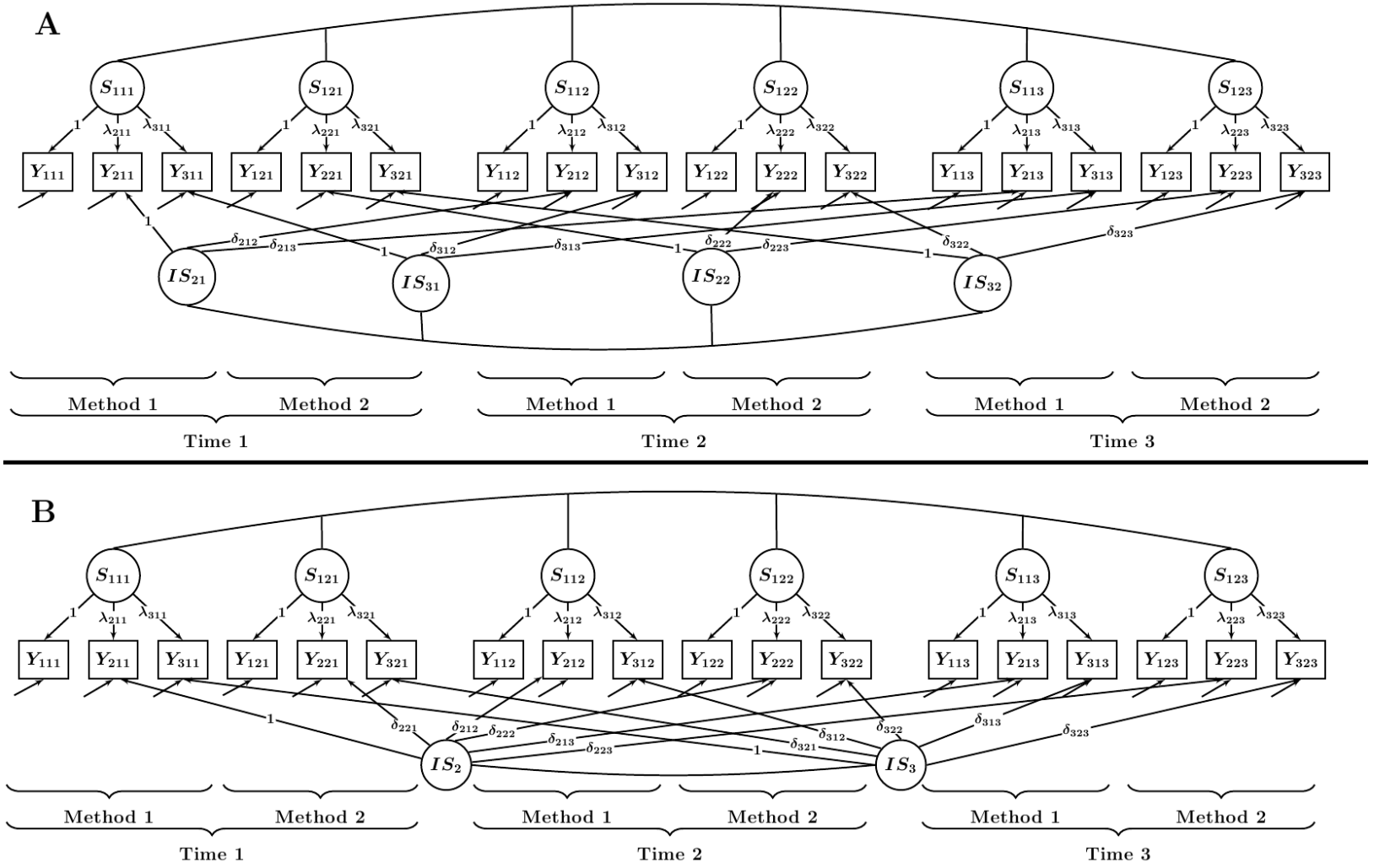

Figure 3. Path diagram of multi-method LS models for three indicators, two methods, and three time points. Indicators are indexed *imt* (*i* = indicator, *m* = method, *t* = time point). A: each method has separate (method-specific) indicator-specific factors IS_{im}. B: both methods share the same (common) indicator-specific factors IS_i. The mean structure has been omitted from the pictures to reduce clutter, but is described in the text and was included in the analysis of each model.

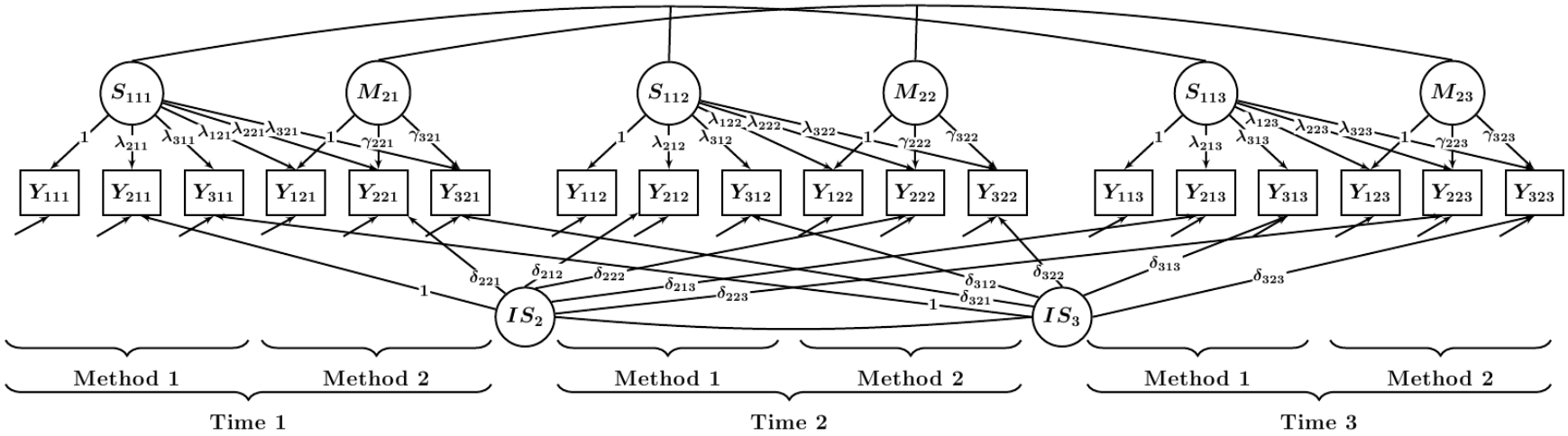

Figure 4. Path diagram of a correlated states, correlated (methods minus one) [CS-C(M – 1)] model with general state factors and indicator-specific factors for three indicators, two methods, and three time points. Method 1 serves as reference method. Method factors M_{2t} are included for the second method to contrast Method 2 against Method 1 on each measurement occasion. Both methods share the same (common) indicator-specific factors IS_i. The mean structure has been omitted from the picture to reduce clutter, but is described in the text and was included in the analysis of the model.

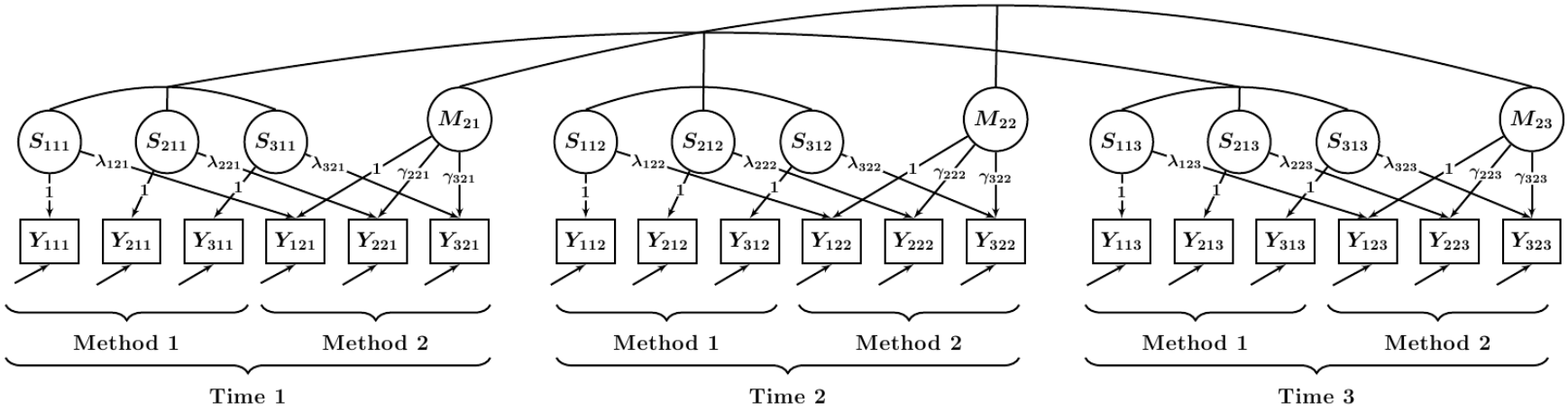

Figure 5. Path diagram of a correlated states, correlated (methods minus one) [CS-C(M – 1)] model with indicator-specific state factors for three indicators, two methods, and three time points. The first method is selected as reference method. Method factors M_{2t} are included for the second method to contrast Method 2 against Method 1 on each measurement occasion. The mean structure has been omitted from the picture to reduce clutter, but is described in the text and was included in the analysis of the model.