frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 17 April 2014
doi: 10.3389/fpsyg.2014.00311

A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods

***Tobias Koch*[1]*, *Martin Schultze*[1], *Michael Eid*[1] and *Christian Geiser*[2]**

[1] *Department of Educational Science and Psychology, Freie Universität Berlin, Berlin, Germany*
[2] *Psychology Department, Utah State University, Logan, UT, USA*

Edited by:
Holmes Finch, Ball State University, USA

Reviewed by:
Thorsten Meiser, University of Mannheim, Germany
Evgueni Borokhovski, Concordia University, Canada

***Correspondence:**
Tobias Koch, Department of Educational Science and Psychology, Freie Universität Berlin, Room J 26/24, Habelschwerdter Allee 45, 14195 Berlin, Germany
e-mail: tkoch@zedat.fu-berlin.de

One of the key interests in the social sciences is the investigation of change and stability of a given attribute. Although numerous models have been proposed in the past for analyzing longitudinal data including multilevel and/or latent variable modeling approaches, only few modeling approaches have been developed for studying the construct validity in longitudinal multitrait-multimethod (MTMM) measurement designs. The aim of the present study was to extend the spectrum of current longitudinal modeling approaches for MTMM analysis. Specifically, a new longitudinal multilevel CFA-MTMM model for measurement designs with structurally different and interchangeable methods (called Latent-State-Combination-Of-Methods model, LS-COM) is presented. Interchangeable methods are methods that are randomly sampled from a set of equivalent methods (e.g., multiple student ratings for teaching quality), whereas structurally different methods are methods that cannot be easily replaced by one another (e.g., teacher, self-ratings, principle ratings). Results of a simulation study indicate that the parameters and standard errors in the LS-COM model are well recovered even in conditions with only five observations per estimated model parameter. The advantages and limitations of the LS-COM model relative to other longitudinal MTMM modeling approaches are discussed.

Keywords: multilevel structural equation modeling, longitudinal modeling, MTMM modeling, multirater data, rater bias, method effects, simulation study

1. INTRODUCTION

An increasing body of research is devoted to longitudinal data analysis examining the change and stability of a given attribute across time (see Singer and Willett, 2003; Khoo et al., 2006). The prominence of longitudinal studies may be explained by the fact that longitudinal measurement designs bear many advantages. Longitudinal measurement designs are more informative than cross-sectional studies, allowing researchers to (1) investigate change and/or variability processes, (2) test the degree of measurement invariance as well as indicator-specific effects, and (3) examine potential causal relationships (see Steyer, 1988, 2005). Over the last decades, many statistical models have been proposed for analyzing longitudinal data including multilevel as well as latent variable modeling approaches (c.f. Little et al., 2000; Singer and Willett, 2003; Rabe-Hesketh and Skrondal, 2004; Steele et al., 2008; Heck et al., 2013). On the other hand, only few attempts have been made to develop appropriate models for longitudinal multitrait-multimethod (MTMM) data (e.g., Kenny and Zautra, 2001; Burns and Haynes, 2006; Courvoisier et al., 2008; Grimm et al., 2009; Geiser et al., 2010; Koch, 2013).

Originally, multitrait-multimethod (MTMM) analysis was developed for scrutinizing the construct validity of social science measures (Campbell and Fiske, 1959). According to Campbell and Fiske (1959) at least two traits (e.g., empathy and aggression) and two methods (e.g., student reports and teacher reports) are required for investigating the degree of convergent and discriminant validity among different measures. Convergent validity refers to the associations (correlations) between two methods measuring the same trait (e.g., the correlation between empathy measured via student and teacher reports). Discriminant validity refers to the question of whether and to which extent methods are able to differentiate between different traits (e.g., the correlation between self-reported empathy and self-reported aggression).

Combining the advantages of longitudinal modeling approaches and MTMM modeling approaches can be fruitful. For example, longitudinal MTMM models allow researchers to investigate the construct validity of different measures across time by combining the information provided by multiple methods or reporters in a single model. This is useful because a researcher would otherwise have to estimate separate longitudinal models for each reporter and no information as to the relationship between reporters could be obtained. Moreover, longitudinal MTMM models allow modeling method effects, examining the stability and change of these method effects across time, and scrutinizing potential causes of method effects by including other (manifest of latent) variables in the model.

The purpose of the present work is to extend the range of longitudinal models for the analysis of complex longitudinal MTMM data by presenting a comprehensive modeling framework for different types of methods. Specifically, we present a new multilevel structural equation model for the analysis of longitudinal MTMM

data featuring interchangeable and structurally different methods. The model is called Latent-State-Combination-Of-Methods model (LS-COM) model. The LS-COM model combines the advantages of four modeling approaches, that is, structural equation modeling, multilevel modeling, longitudinal modeling, and MTMM modeling with interchangeable and structurally different methods. In particular, the LS-COM allows researchers to (1) explicitly model measurement error, (2) specify method factors on different measurement levels, (3) analyze the convergent and discriminant validity across multiple occasions, (4) investigate change and stability of construct and methods effects across time, and (5) test important assumptions in longitudinal data analysis such as the degree of measurement invariance. The LS-COM model is formulated based on the principles of stochastic measurement theory (Zimmerman, 1975; Steyer and Eid, 2001), which has the advantage that all latent variables in the model are psychometrically well-defined as random variables with a clear meaning.

The article is structured as follows: First, we review conventional (single-method) models of longitudinal confirmatory factor analysis with a special focus on latent state (LS) models (Steyer et al., 1992). Second, we discuss current extensions of LS-modeling approaches to MTMM designs with structurally different methods. In this regard, we review the correlated state-correlated method minus one [CS-C(M-1)] model by Geiser et al. (2010). Furthermore, we explain the differences between measurement designs with structurally different methods, interchangeable methods, or a combination of both methods. We show that the CS-C(M-1) model is useful for modeling data obtained from longitudinal MTMM measurement designs with structurally different methods, but that this model is not suitable for measurement designs with a combination of structurally different and interchangeable methods. Third, we present the new LS-COM model for longitudinal MTMM designs with structurally different and interchangeable methods. The new LS-COM model fills a gap in the literature, as previous approaches to longitudinal MTMM analysis focused exclusively on structurally different methods. Fourth, we report the results of a Monte Carlo simulations study in which we examined the statistical performance of the LS-COM model. Finally, we discuss the advantages and limitations of the LS-COM model compared to other longitudinal MTMM modeling approaches.

2. LONGITUDINAL CONFIRMATORY FACTOR ANALYSIS

The versatility and flexibility of the CFA framework have inspired the development of different CFA models for longitudinal measurement designs, for example, autoregressive models (Hertzog and Nesselroade, 1987; Jöreskog, 1979a,b; Marsh, 1993; Eid and Hoffmann, 1998), latent state models (Steyer et al., 1992), latent change (difference score) models (Steyer et al., 1997, 2000; McArdle, 1988), latent state-trait models (Steyer et al., 1992, 1999), and latent growth curve models (McArdle and Epstein, 1987; Meredith and Tisak, 1990; Hancock et al., 2001; Bollen and Curran, 2006; Duncan et al., 2006). Most previous longitudinal models have been designed for single method measurement designs (e.g., self-reports) only. Presumably, the simplest CFA model for longitudinal data is the latent state (LS) model, which represents an extension of classical test theory to longitudinal measurement designs (see Steyer et al., 1992; Marsh, 1993; Tisak and Tisak, 2000; Geiser, 2009). The LS model is often used as a baseline model, given that it implies no restrictions with regard to the structural part of the model (see **Figure 1**). Hence, the LS model is often used for testing the measurement model (e.g., the validity of the assumed factor structure, measurement invariance restrictions, correlations of error variables, unidimensionality of the scales on an occasion of measurement). According to latent state theory (see Steyer et al., 1992), each observed variable Y_{il} can be decomposed into a latent state (S_{il}, occasion-specific true score) variable and a measurement error variable ϵ_{il}, where i is the indicator (item or parcel) and l denotes the occasion of measurement:

$$Y_{il} = S_{il} + \epsilon_{il}. \tag{1}$$

The latent state variable S_{il} represents the individual state scores at a particular occasion of measurement, whereas the measurement error variables reflect unsystematic influences due to measurement error. It can be shown that the additive decomposition of the observed variables Y_{il} into a latent state variable S_{il} and a latent measurement error variable ϵ_{il} follow directly, if both latent variables are defined in terms of conditional expectations (see Steyer, 1988, 1989; Steyer et al., 1992). In order to estimate a latent state model, it is assumed that (1) the latent state variables belonging to the same occasion of measurement are linear functions of each other (i.e., congenerity assumption):

$$S_{il} = \alpha_{ii'l} + \lambda_{ii'l} S_{i'l}, \tag{2}$$

and that (2) the measurement error variables [i.e., $Cov(\epsilon_{il}, \epsilon_{il})$ for $(i, l) \neq (i', l')$] are uncorrelated with each other. Equation (2) states that the latent state variables are linear functions of each other and only differ by an additive constant $\alpha_{ii'l}$ and multiplicative constant $\lambda_{ii'l}$. With respect to this assumption, it is possible to show that Equation (2) is equivalent to $S_{il} = \alpha_{il} + \lambda_{il} S_l$. Hence, the general measurement equation of a latent state model with common latent state factors can be written as follows:

$$Y_{il} = \alpha_{il} + \lambda_{il} S_l + \epsilon_{il}. \tag{3}$$

α_{il} is the intercept and λ_{il} is the factor loading parameter pertaining to the latent state factors. As a consequence of the assumptions explained above, the total variance of the observed variables can be decomposed as follows:

$$Var(Y_{il}) = \lambda_{il}^2 Var(S_l) + Var(\epsilon_{il}). \tag{4}$$

The reliability of each observed variable is then given by:

$$Rel(Y_{il}) = \frac{\lambda_{il}^2 Var(S_l)}{Var(Y_{il})}. \tag{5}$$

Figure 1 shows a path diagram of the latent state model for three indicators and three occasions.

The correlations between the latent state factors S_l characterize the stability of interindividual differences on the given attribute

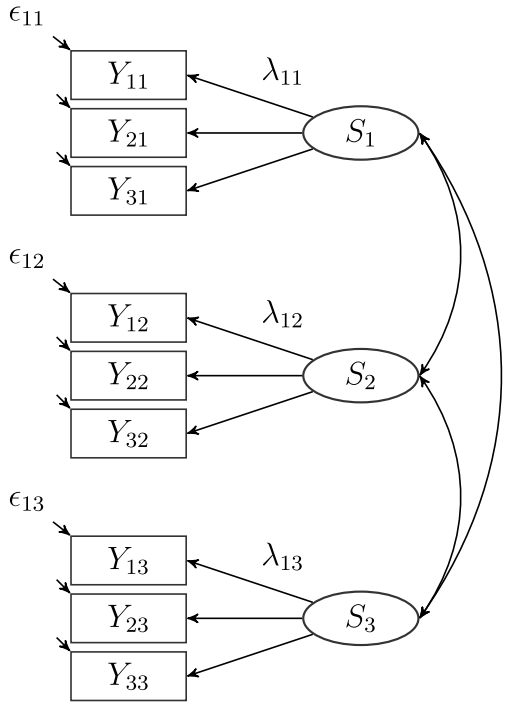

FIGURE 1 | A latent state (LS) model with three indicators ($i = 3$) and three occasions of measurement ($l = 3$), where S_l: latent state factors, λ_{il}: latent state factor loadings, ϵ_{il}: measurement error variables. Intercepts (α_{il}) are not depicted in the figure.

(see **Figure 1**). High correlations reflect that individual differences with regard to a particular attribute (construct) are rather stable over time. Researchers may also investigate mean change of a given construct across time. For meaningful interpretations of latent mean change, we recommend that measurement invariance (MI) should be tested and that researchers should at least establish strong MI (e.g., Meredith, 1993; Widaman and Reise, 1997; Millsap, 2012).

Strong MI can be established by imposing the following restrictions:

1. The intercepts of the observed variables α_{il} have to be set equal across time (i.e., $\alpha_{il} = \alpha_{il'} = \alpha_i$).
2. The factor loading parameters λ_{il} have to be set equal across time (i.e., $\lambda_{il} = \lambda_{il'} = \lambda_i$) and one factor loading parameter on each occasion of measurement has to be fixed to the same value (e.g., $\lambda_1 = 1$).
3. The mean of the first latent state factor has to set to be zero [i.e., $E(S_1) = 0$].
4. The mean of the remaining latent state factors can be freely estimated [i.e., $E(S_l) \neq 0$].

Strong MI is a prerequisite for studying true mean change (Steyer et al., 1997, 2000)[1]. Restrictions 3 and 4 allow examining mean change relative to the first measurement occasion[2]. Although the LS model (as well as other longitudinal CFA models) offers many advantages such as analyzing change and stability of an attribute apart from measurement error influences and testing the degree of measurement invariance, the LS model is limited in terms of incorporating data from multiple raters or methods, because the model does not contain method factors. In order to study the convergent and discriminant validity in longitudinal MTMM designs, more sophisticated models are needed.

3. LONGITUDINAL CFA-MTMM MODELS

According to Eid and Diener (2006) multimethod measurement designs overcome many limitations of single method measurement designs and should therefore be preferred whenever possible. With respect to longitudinal CFA-MTMM models it is possible to (1) investigate the convergent and discriminant validity at each occasion of measurement and across different occasions of measurement, (2) study change and stability of construct and method effects across time, (3) model measurement error, (4) investigate the generalizability of method effects, and (5) test important assumptions such as measurement invariance and/or indicator-specific effects.

Today, MTMM measurement designs are commonly analyzed using confirmatory factor analysis (CFA-MTMM models) with multiple indicators in each trait-method unit (e.g., Widaman, 1985; Marsh and Hocevar, 1988; Marsh, 1989; Wothke, 1995; Dumenci, 2000; Eid, 2000; Eid et al., 2003, 2006). Up to now, only few CFA-MTMM models have been proposed for the analysis of longitudinal data (e.g., Kenny and Zautra, 2001; Burns and Haynes, 2006; Courvoisier et al., 2008; Grimm et al., 2009; Geiser et al., 2010; Koch, 2013).

One exception is the study by Grimm et al. (2009) who recently proposed a longitudinal CFA-MTMM model combining the correlated trait-correlated method (CT-CM) approach (Widaman, 1985; Marsh and Grayson, 1995) and the latent growth curve modeling approach (e.g., McArdle and Epstein, 1987; Meredith and Tisak, 1990). However, results of previous studies have shown that the CT-CM modeling approach is associated with various theoretical and empirical problems (e.g., Marsh, 1989; Kenny and Kashy, 1992; Marsh and Grayson, 1994; Steyer, 1995; Eid, 2000; Geiser et al., in press). In addition, the CFA-MTMM model by Grimm et al. (2009) is limited to single-indicator measurement designs and does not allow specifying trait-specific method factors.

Geiser et al. (2010) developed a longitudinal CFA-MTMM model [called correlated state-correlated method minus one, CS-C(M-1) model] that combines LS theory with the correlated trait-correlated method minus one [CT-C(M-1)] approach (Eid, 2000; Eid et al., 2003, see **Figure 2**). In this model, one method has to be chosen as reference method which all other methods are compared to. The common latent state factor is the state factor of the reference method. Each observed variable of a

[1]For more details on partial (MI) see Byrne et al. (1989) and on approximate measurement invariance see Van De Schoot et al. (2013).

[2]Another possibility is to set the intercept of a reference indicator (e.g., first indicator) to zero on all occasions of measurement and estimate the latent means of the latent state factor on each occasion of measurement.

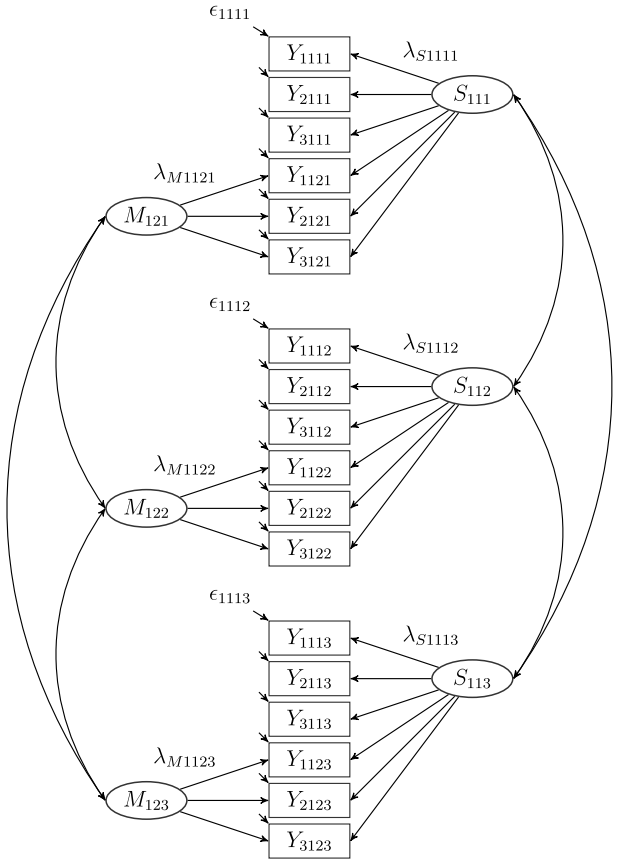

FIGURE 2 | A latent correlated state correlated method minus one [CS-($M-1$)] model with three indicators ($i = 3$), one construct ($j = 1$), two methods ($k = 2$) and three occasions of measurement ($l = 3$), where Y_{ijkl}: observed variables, S_{jkl}: latent state factors, λ_{Sijkl}: state factor loadings, M_{jkl}: latent method factors, λ_{Mijkl}: method factor loadings, ϵ_{ijkl}: measurement error variables. Intercepts (α_{ijkl}) are not depicted in the figure.

non-reference method is decomposed into three parts: (1) a part that is predictable by the common state factor, (2) a part that is method-specific, and (3) measurement error. One advantage of the CS-C(M-1) model is that all latent variables are well-defined with a clear and unambiguous interpretation (Geiser, 2009). The CS-C(M-1) model also overcomes many limitations of previous CFA-MTMM modeling approaches. For example, the CS-C(M-1) model allows specifying trait-specific method factors using multiple indicators per trait-method unit (TMU) and separating the observed variance into trait, method, and measurement error variance. According to the results of simulation studies (Crayen, 2008; Geiser, 2009), the CS-C(M-1) model performs well in many conditions.

However, the CS-C(M-1) model cannot be applied to all possible longitudinal MTMM measurement designs. In particular, the CS-C(M-1) model is not suitable for MTMM measurement designs combining structurally different and interchangeable methods. In the next section, the differences between measurement designs with structurally different and interchangeable methods are explained in greater detail.

4. DIFFERENT TYPES OF METHODS

Eid et al. (2008) clarified that the type of method used in a study is of particular importance for defining appropriate CFA-MTMM models. More specifically, Eid et al. (2008) showed that measurement designs with (a) interchangeable methods, (b) structurally different methods, and (c) a combination of structurally different and interchangeable methods imply different sampling procedures and therefore require different CFA-MTMM models. According to Eid et al. (2008), interchangeable methods are methods that can be randomly sampled from a set of similar methods. Consider, for example, multiple peer ratings of students' empathy. Both, peer ratings and subordinates' ratings can be considered as interchangeable, because they have more or less the same access to the target's behavior (Eid et al., 2008). **Figure 3B** illustrates the sampling procedure for interchangeable methods. According to **Figure 3B**, measurement designs with interchangeable methods imply a multistage sampling procedure (Eid et al., 2008; Koch et al., in press). First, a target (t, e.g., teacher) is randomly chosen from a set of all possible targets ($t \in T$, i.e., all teachers). Second, multiple (e.g., three) students (e.g., Edgar, Emily, and Mark) are randomly sampled from the same target-specific rater set R_t. Therefore, measurement designs with interchangeable methods imply a multilevel data structure (Eid et al., 2008).

In contrast, measurement designs with structurally different methods (see **Figure 3A**) use methods that are not randomly sampled out of a common set of similar methods (raters). For example, structurally different methods such as self-ratings, parent ratings, and the ratings of the class teacher do not stem from the same group of methods, but differ in many ways. As a consequence, measurement designs with structurally different methods can usually be modeled with single-level factor models [e.g., CS-(M-1) model]. An increasing number of studies use a combination of structurally different and interchangeable methods. For example, in organizational psychology it is very common to use self-reports, supervisor reports, and interchangeable colleague reports (so-called 360° feedback designs). In educational and developmental psychology, many researchers use student reports, teacher and parent reports, as well as interchangeable peer reports. All of these designs imply a combination of structurally different and interchangeable methods.

5. THE NEED FOR LONGITUDINAL MULTILEVEL CFA-MTMM MODELS

So far, no appropriate CFA model has been proposed for longitudinal MTMM data combining structurally different and interchangeable methods. Researchers who use such MTMM measurement designs (e.g., longitudinal multisource feedback designs with different types of raters) are therefore forced to either aggregate the interchangeable ratings into a single score or analyze both types of methods (structurally different and interchangeable methods) in separate models. The aggregation of level one units (here interchangeable methods) has been associated with various methodological shortcomings, such as, interpretation problems (e.g., ecological fallacy), loss of information, smaller sample size, larger standard errors, and loss of power (Hox, 2010; Snijders and Bosker, 2011). If both types of methods are analyzed separately, then researchers are not able to integrate (or compare)

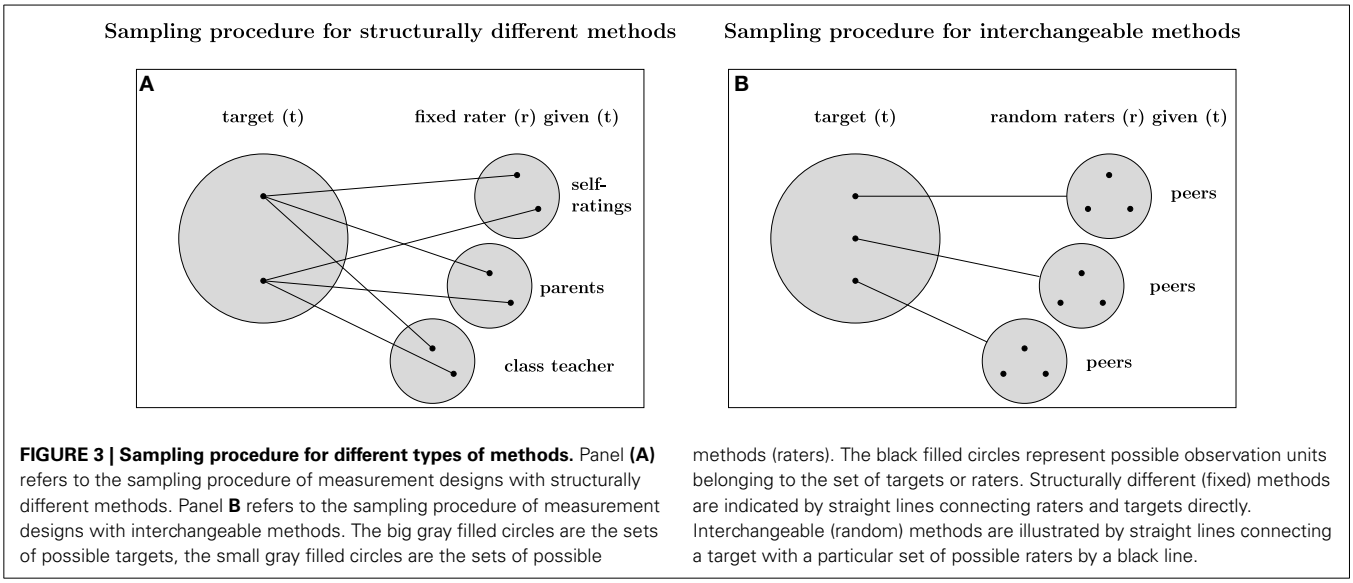

FIGURE 3 | Sampling procedure for different types of methods. Panel **(A)** refers to the sampling procedure of measurement designs with structurally different methods. Panel **B** refers to the sampling procedure of measurement designs with interchangeable methods. The big gray filled circles are the sets of possible targets, the small gray filled circles are the sets of possible methods (raters). The black filled circles represent possible observation units belonging to the set of targets or raters. Structurally different (fixed) methods are indicated by straight lines connecting raters and targets directly. Interchangeable (random) methods are illustrated by straight lines connecting a target with a particular set of possible raters by a black line.

the information of both types of methods (rater groups) in the same model. For example, convergent validity of interchangeable peer reports and self-reports could be examined. Given that many researchers increasingly apply measurement designs with a combination of structurally different and interchangeable methods, there is a need for developing new methods for the analysis of such complex MTMM measurement designs. In the next section, we present the LS-COM model, which integrates LS theory and the CS-C(M-1) modeling approach for a combination of different types of methods. In addition, we present the results of a Monte Carlo simulation study, in which we examined the statistical performance of the LS-COM model under different conditions.

6. THE LATENT STATE COMBINATION OF METHODS (LS-COM) MODEL

The LS-COM allows researchers

1. to scrutinize the degree of measurement invariance across time,
2. to test mean changes of particular constructs,
3. to examine the stability and change of construct and method effects across time,
4. to investigate the psychometric properties (e.g., the convergent and discriminant validity and reliability) of the given measures on each occasion of measurement and across occasions of measurement, and
5. to scrutinize the generalizability of method effects across different methods and/or different constructs.

Similar to the CS-C(M-1) model, we define the LS-COM model in different steps.

6.1. STEP 1: CHOICE OF REFERENCE METHOD AND BASIC DECOMPOSITION

In the first step, a reference or gold-standard method has to be chosen. The remaining methods serve as non-reference methods. The reference method is often a method that is either seen as most valid by the researcher based on theory or prior empirical results or a method that is particularly outstanding or special relative to the other methods (e.g., objective IQ tests versus self-ratings of intelligence). In the LS-COM model either one of the structurally different or the set of interchangeable methods can be chosen as reference method. For the sake of simplicity, we define the LS-COM model for two structurally different methods (method 1 = self-report, method 2 = parent report) and one set of interchangeable methods (method 3 = multiple peer reports for a student). Note that the LS-COM model is not restricted in terms of the number of structurally different methods. Moreover, we chose the first method (a structurally different method, e.g., self-reports) as reference method and assume that there is only one parent report for each target. Pham et al. (2012) as well as Schultze (2012) show how the set of interchangeable methods can be chosen as reference method. The observed variables of each method can be decomposed into a latent state and a latent measurement error variable:

$$\text{Level 2:} \quad Y_{tij1l} = S_{tij1l} + \epsilon_{tij1l}, \quad \text{(structurally different method 1)} \tag{6}$$

$$\text{Level 2:} \quad Y_{tij2l} = S_{tij2l} + \epsilon_{tij2l}, \quad \text{(structurally different method 2)} \tag{7}$$

$$\text{Level 1:} \quad Y_{rtij3l} = S_{rtij3l} + \epsilon_{rtij3l}. \quad \text{(set of interchangeable methods)} \tag{8}$$

The index i represents the indicators, j is the construct, k is the method, and l is the occasion of measurement. In addition, the indices r for rater and t for target are required. The reason is that the interchangeable raters r are nested within different targets t. Hence, the observed variables of the self-reports and parent reports are measured on Level 2 (the target level), whereas the observed variables pertaining to the interchangeable methods (peer reports) are measured on Level 1 (the rater level).

A value of the target-specific latent state variables S_{tijkl} is the true score of target t with respect to indicator i, construct j, method k (i.e., self-report or parent report), and occasion of measurement l. The rater-specific latent state variables S_{rtijkl} reflect the (method-specific) true peer rating of a rater r for a particular target t on indicator i, construct j, and occasion of measurement l. The measurement error variables on both levels are represented by ϵ_{tijkl} (Level 2) and ϵ_{rtijkl} (Level 1). In the Appendix A in Supplementary Material, we show how the latent state and measurement error variables are formally defined in terms of conditional expectations.

6.2. STEP 2: DEFINITION OF RATER-SPECIFIC LATENT METHOD VARIABLES ON LEVEL 1

In the second step, rater-specific (Level 1) latent method variables are defined for the interchangeable methods (i.e., multiple peer reports). This is possible given that multiple peers r rate each target t on different items (indicators: i). Therefore, the rater-specific latent state variables can be decomposed into a rater-unspecific latent state S_{tij3l} variable and a rater-specific method UM_{rtij3l} variable.

$$Y_{rtij3l} = S_{rtij3l} + \epsilon_{rtij3l} \tag{9}$$

$$S_{rtij3l} = S_{tij3l} + UM_{rtij3l} \tag{10}$$

$$Y_{rtij3l} = S_{tij3l} + UM_{rtij3l} + \epsilon_{rtij3l}. \tag{11}$$

A value of the latent state variables S_{tij3l} can be conceived as the expected peer rating of the target t across the true occasion-specific peer ratings for that target. That is, the latent state variables S_{tij3l} can be conceived as the average peer rating and are thus variables on Level 2. A value of the latent unique method variables UM_{rtij3l} is the true occasion-specific deviation of a particular rater from this true mean. Hence, a value of the UM_{rtij3l}-variables represents the over- or underestimation of the true expected peer rating by a particular rater r. Positive values indicate an overestimation, whereas negative values indicate an underestimation of the true expected peer rating by a particular rater. Given that the unique method UM_{rtij3l}-variables are defined as latent residual variables, the general properties of residual variables hold. That means that the unique method UM_{rtij3l}-variables are uncorrelated with the Level 2 latent state S_{tij3l} variables [i.e., $Cor(S_{tij1l}, UM_{rtj3l}) = 0$] and have an expectation (mean) of zero [i.e., $E(UM_{rtj3l}) = 0$]. Moreover, as in classical multilevel (structural equation) models, it is assumed that the Level 1 residuals (here: the UM_{rtij3l}-variables) are independently and identically distributed on Level 1 (i.e., iid-assumption).

6.3. STEP 3: LATENT REGRESSIONS AND DEFINITION OF LATENT METHOD VARIABLES ON LEVEL 2

Given that all latent state variables S_{tijkl} are now measured on the same level (Level 2; the target level), it is possible to contrast the latent state variables pertaining to different types of methods against each other. Following the original CT-C(M-1) approach for structurally different methods (Eid, 2000; Eid et al., 2003, 2008), the latent state variables pertaining to the non-reference methods are regressed on the latent state variables pertaining to the reference method (in this example self-reports):

$$E(S_{tij2l}|S_{tij1l}) = \alpha_{ij2l} + \lambda_{Sij2l}\, S_{tij1l}, \quad \text{(parent reports)} \tag{12}$$

$$E(S_{tij3l}|S_{tij1l}) = \alpha_{ij3l} + \lambda_{Sij3l}\, S_{tij1l}. \quad \text{(peer reports)} \tag{13}$$

The (independent) latent state variable S_{tij1l} in the latent regression analysis denotes the occasion-specific true score measured by the reference method (e.g., self-reports). The residuals of the latent regression analyses are defined as latent method variables. These method variables are also measured on the target level (Level 2). With regard to the structurally different non-reference method (e.g., parent reports), the method variables can be defined as follows:

$$M_{tij2l} \equiv S_{tij2l} - E(S_{tij2l}|S_{tij1l}) = S_{tij2l} - (\alpha_{ij2l} + \lambda_{Sij2l}\, S_{tij1l}). \tag{14}$$

The method variables M_{tij2l} represent that part of the true parent reports that cannot be predicted by the self-reports. In other words, these method variables capture the occasion-specific part of the parent report that cannot be predicted by the self-report. As consequence of the definition of the M_{tij2l}-variables as residual variables the latent method variables are uncorrelated with the latent state variables [i.e., $Cor(S_{tij1l}, M_{tj2l}) = 0$] and have an expectation (mean) of zero [i.e., $E(M_{tj2l}) = 0$]. For the set of interchangeable methods (e.g., peer reports), the method variables can be defined as follows:

$$CM_{tij3l} \equiv S_{tij3l} - E(S_{tij3l}|S_{tij1l}) = S_{tij3l} - (\alpha_{ij3l} + \lambda_{Sij3l} S_{tij1l}). \tag{15}$$

The method variables CM_{tij3l} represent that part of the true expected peer ratings that is not shared with self-report on the same occasion of measurement. The common method variable is called common method variable, given that they represent the perspective of the peers that is shared by all peers, but is not shared with the self-reports on a particular occasion of measurement. By definition the latent common method variables are uncorrelated with the corresponding latent state variables of the reference method [i.e., $Cor(S_{tij1l}, CM_{tj3l}) = 0$] and have an expectation (mean) of zero [i.e., $E(CM_{tj3l}) = 0$]. Moreover, the following correlations are assumed to be zero in the LS-COM model:

$$Cor(S_{tij1l}, UM_{rtj'3l'}) = 0, \tag{16}$$

$$Cor(CM_{tij3l}, UM_{rtj'3l'}) = 0, \tag{17}$$

$$Cor(M_{tij2l}, UM_{rtj'3l'}) = 0, \tag{18}$$

$$Cor(\epsilon_{rt(ijkl)}, \epsilon_{rt(ijkl)'}) = 0, \tag{19}$$

$$Cor(\epsilon_{t(ijkl)}, \epsilon_{t(ijkl)'}) = 0, \tag{20}$$

$$Cor(\epsilon_{rt(ijkl)}, \epsilon_{rt(i'j'k'l')}) = 0. \tag{21}$$

According to Equations (16–18), it is assumed that the Level 1 unique method variables are uncorrelated with all variables on

Level 2 (i.e., latent state, latent common method, and latent method variables). Equations (19–21) imply that all measurement error variables belonging to different indicators, different constructs, different methods, and different occasions of measurement are uncorrelated with each other.

6.4. STEP 4: DEFINITION OF LATENT METHOD FACTORS

In order to define latent method factors, it is assumed that the latent method variables of the same method only differ by multiplicative constants (i.e., λ_{Mij2l}, λ_{CMij3l}, λ_{UMij3l}). According to these assumptions, it is possible to define common latent method factors that are homogeneous across different indicators (i.e., M_{tj2l}, CM_{tj3l}, UM_{rtj3l}):

$$\text{Level 2:} \quad M_{tij2l} = \lambda_{Mij2l} M_{tj2l}, \tag{22}$$

$$\text{Level 2:} \quad CM_{tij3l} = \lambda_{CMij3l} CM_{tj3l}, \tag{23}$$

$$\text{Level 1:} \quad UM_{rtij3l} = \lambda_{UMij3l} UM_{rtj3l}. \tag{24}$$

The above Equations (22–24) state that the method effects are now measured by latent method factors that are common to all indicators.

6.5. STEP 5: DEFINITION OF LATENT STATE FACTORS

Following a similar logic, it is possible to construe a latent state factor S_{tj1l} that is common to all indicators:

$$S_{tij1l} = \alpha_{ij1l} + \lambda_{Sij1l} S_{tj1l}. \tag{25}$$

Overall, the general measurement equation of the LS-COM model for three methods (e.g., $k = 1$ = self-report, $k = 2$ = parent report, $k = 3$ = peer reports) and latent state factors (S_{tj1l}) can be expressed by:

$$Y_{tij1l} = \alpha_{ij1l} + \lambda_{Sij1l} S_{tj1l} + \epsilon_{tij1l}, \tag{26}$$

$$Y_{tij2l} = \alpha_{ij2l} + \lambda_{Sij2l} S_{tj1l} + \lambda_{Mij2l} M_{tj2l} + \epsilon_{tij2l}, \tag{27}$$

$$Y_{rtij3l} = \alpha_{ij3l} + \lambda_{Sij3l} S_{tj1l} + \lambda_{CMij3l} CM_{tj3l} + \lambda_{UMij3l} UM_{rtj3l} + \epsilon_{rtij3l}. \tag{28}$$

Equation (26) states that the reference method (e.g., self-report) indicators are only measured by a latent reference state factor S_{tj1l} with an intercept α_{ij1l} and factor loading parameter λ_{Sij1l} and a latent measurement error variable ϵ_{tij1l}. According to Equation (27) the indicators pertaining to a structurally different non-reference method (e.g., parent reports) are measured by the latent reference state factor S_{tj1l} (with an intercept α_{ij2l} and factor loading parameter λ_{Sij2l}), a latent method factor M_{tj2l} (with a factor loading parameter λ_{Mij2l}), and a measurement error variable ϵ_{tij2l}. Finally, Equation (28) states that the indicators belonging to the interchangeable non-reference method (e.g., peer reports) are measured by the latent reference state factor S_{tj1l} (with a corresponding intercept α_{ij3l} and factor loading parameter $\lambda_{Sij3l} S_{tj1l}$), a latent common method factor CM_{tj3l} at Level 2 and a latent unique method factor UM_{rtj3l} at Level 1 (with corresponding factor loading parameters λ_{CMij3l} and λ_{UMij3l}), as well as a measurement error variable ϵ_{rtij3l}.

7. VARIANCE DECOMPOSITION

Based on the definition of the LS-COM model each indicator's variance can be decomposed as follows:

$$Var(Y_{tij1l}) = \lambda^2_{Sij1l} Var(S_{tj1l}) + Var(\epsilon_{tij1l}), \tag{29}$$

$$Var(Y_{tij2l}) = \lambda^2_{Sij2l} Var(S_{tj1l}) + \lambda^2_{Mij2l} Var(M_{tj2l}) + Var(\epsilon_{tij2l}), \tag{30}$$

$$Var(Y_{rtij3l}) = \lambda^2_{Sij3l} Var(S_{tj1l}) + \lambda^2_{CMij3l} Var(CM_{tj3l}) + \lambda^2_{UMij3l} Var(UM_{rtj3l}) + Var(\epsilon_{rtij3l}). \tag{31}$$

Based on the above variance decomposition (see Equations 29–31), it is possible to define different coefficients for quantifying convergent validity, method-specificity and reliability (see **Table 1**). In contrast to the CS-C(M-1) model, the LS-COM model allows calculating Level 2 and Level 1 variance coefficients, because it contains method factors at both Level 1 (UM_{rtj3l}) and Level 2 (CM_{tj3l}).

In total, four different consistency coefficients [i.e., $Con(S_{tij2l})$, $Con(S_{tij3l})$, $Con(S_{rtij3l})$, and the rater consistency coefficient $RC(S_{rtij3l})$] can be defined. The Level 2 consistency coefficient $Con(S_{tij2l})$ for the indicators pertaining to the structurally different non-reference methods represents the amount of true interindividual differences of the non-reference method (e.g., parent report) that can be explained by the reference method (self-report). The Level 1 consistency coefficient $Con(S_{rtij3l})$ for the indicators pertaining to the interchangeable non-reference methods (e.g., peer reports) reflects the amount of true interindividual differences of the individual peer reports that can be explained by the reference method (here: self-report).

Sometimes researchers rather seek to know whether peers in general agree with the student self-reports. In such cases, they may calculate the Level 2 consistency coefficient $Con(S_{tij3l})$ for the indicators pertaining to the set of interchangeable methods. This consistency coefficient captures the amount of true interindividual differences of the expected peer ratings (the entire peer-group) that can be explained by the reference method (here: self-reports). Moreover, the true rater consistency coefficient $RC(S_{rtij3l})$ is defined as the proportion of true interindividual differences of the peer ratings that are free of measurement error and rater-specific effects. The rater consistency coefficient indicates how much true variance of a non-reference indicator is due to the overall amount of rater agreement (peers and self-ratings) and not due to measurement error influences or individual (rater-specific) influences. The true rater consistency coefficient can also be interpreted as true intra-class correlation. Moreover, three different method-specificity coefficients [i.e., $MS(S_{tij2l})$, $CMS(S_{rtij3l})$, and $UMS(S_{rtij3l})$] can be analyzed. The method specificity coefficients $MS(S_{tij2l})$ indicate the degree or true variance of a non-reference method indicator pertaining to a structurally different method (e.g., parent reports) that is

Table 1 | Variance components of the non-reference method indicators in LS-COM model.

Level	Method	Definition
CONSISTENCY		
Level 2	Struct. different	$Con(S_{tij2l}) = \frac{\lambda^2_{Sij2l} Var(S_{tij1l})}{Var(Y_{tij2l}) - Var(\epsilon_{tij2l})}$
Level 2	Interchangeable	$Con(S_{tij3l}) = \frac{\lambda^2_{Sij3l} Var(S_{tij1l})}{\lambda^2_{Sij3l} Var(S_{tij1l}) + \lambda^2_{CMij3l} Var(CM_{tj3l})}$
Level 1	Interchangeable	$Con(S_{rtij3l}) = \frac{\lambda^2_{Sij3l} Var(S_{tij1l})}{Var(Y_{rtij3l}) - Var(\epsilon_{rtij3l})}$
Level 1	Interchangeable	$RC(S_{rtij3l}) = \frac{\lambda^2_{Sij3l} Var(S_{tij1l}) + \lambda^2_{CMij3l} Var(CM_{tj3l})}{Var(Y_{rtij3l}) - Var(\epsilon_{rtij3l})}$
METHOD SPECIFICITY		
Level 2	Struct. different	$MS(S_{tij2l}) = \frac{\lambda^2_{Mij2l} Var(M_{tj2l})}{Var(Y_{tij2l}) - Var(\epsilon_{tij2l})}$
Level 2	Interchangeable	$CMS(S_{rtij3l}) = \frac{\lambda^2_{CMij3l} Var(CM_{tj3l})}{Var(Y_{rtij3l}) - Var(\epsilon_{rtij3l})}$
Level 1	Interchangeable	$UMS(S_{rtij3l}) = \frac{\lambda^2_{UMij3l} Var(UM_{rtj3l})}{Var(Y_{rtij3l}) - Var(\epsilon_{rtij3l})}$
RELIABILITY		
Level 2	Struct. different	$Rel(Y_{tij2l}) = 1 - \frac{Var(\epsilon_{tij2l})}{Var(Y_{tij2l})}$
Level 1	Interchangeable	$Rel(Y_{rtij3l}) = 1 - \frac{Var(\epsilon_{rtij3l})}{Var(Y_{rtij3l})}$

not determined by the reference method (e.g., self-report). The unique method specificity coefficient $UMS(S_{rtij3l})$ represents the proportion of true variance of a non-reference method indicator pertaining to the interchangeable methods that is neither shared with the self-reports nor with other peers. Hence, this coefficient reflects the unique view of a particular rater on a particular occasion of measurement. The common method specificity coefficient $CMS(S_{rtij3l})$ reflects the proportion of true interindividual differences of the peer ratings that cannot be explained the reference method (here: self-reports), but that is shared by other peers (Eid et al., 2008). Hence, this coefficient can also be interpreted as "rater consensus" with respect to the peer ratings that is not shared with the reference method.

8. PERMISSIBLE CORRELATIONS

Figure 4 shows a path diagram of a LS-COM model with three indicators per TMU, one construct, three methods and three occasions of measurement. As illustrated in the figure, the latent state factors can be correlated with each other (see **Figure 4**). Correlations between latent state factors pertaining to the same construct (e.g., empathy) and different occasions of measurement can be interpreted as indicators of construct stability. High positive correlations indicate that the construct is rather stable across time. Correlations between latent state factors pertaining to different constructs and the same occasion of measurement can be interpreted in terms of discriminant validity. High correlations indicate low discriminant validity at a given moment in time. Correlations between latent state factors pertaining to different constructs and different measurement occasions may be interpreted as coefficients of predictive validity. For example, students' self-reported level of empathy measured on the first occasion of measurement (S_{t111}) may be indicative for the self-reported level of relational aggression measured on the second occasion of measurement (S_{t212}). Moreover, correlations between the occasion-specific latent method factors pertaining to the same measurement level are permitted in the LS-COM model.

The stability of method (rater) effects can be investigated by correlations between method factors pertaining to the same construct, same method, and different occasions of measurement. For example, correlations between the unique method factors UM_{rtj3l} and $UM_{rtj3l'}$ (where $l \neq l'$) indicate to what extent the individual rater-specific effects remain stable across time. Following a similar logic, the correlations between common method factors CM_{tj3l} and $CM_{tj3l'}$ (where $l \neq l'$) indicate to what extent the common rater effects (i.e., rater effects that are not shared with the self-report, but are shared with all other raters belonging to a particular target) remain stable across time. The generalization of method effects across constructs is indicated by correlations between method factors pertaining to different constructs (e.g., empathy and relational aggression). For example, a negative correlation between the method factors pertaining to the peer reports assessing empathy and relational aggression would indicate that peers who tend to underestimate students' self-reported empathy level, tend to overrate students' self-reported aggression level and vice versa. The generalization of method effects across different methods (rater types) is indicated by the correlation between method factors

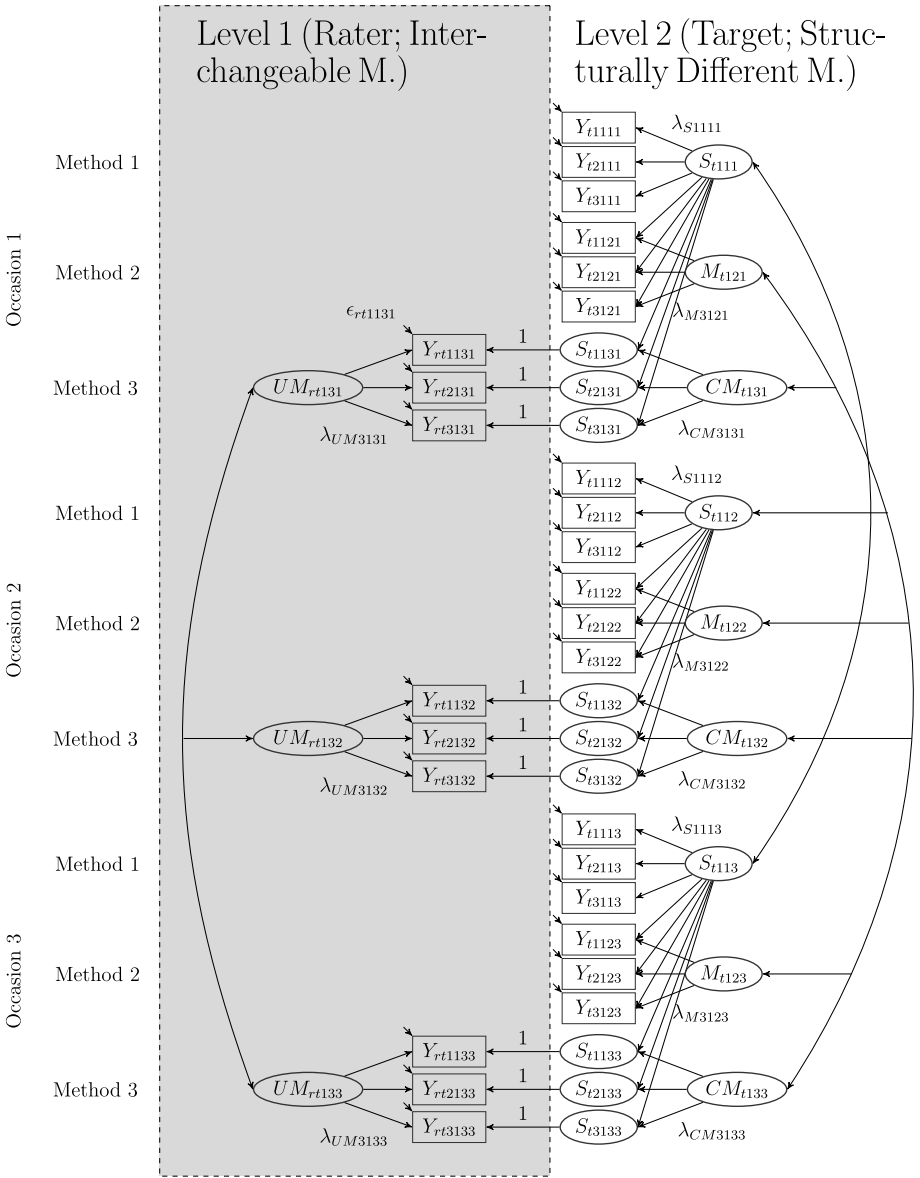

FIGURE 4 | The LS-COM model with latent state factors and three indicators ($i = 3$), one construct ($j = 1$), two structurally different and one set of interchangeable methods ($k = 3$), and three occasions of measurement ($l = 3$), where *r*: rater and *t*: target. Y_{rtijkl}: observed variables at Level 1, Y_{tijkl}: observed variables at Level 2, S_{tjkl}: latent state factors, λ_{Sijkl}: state factor loadings, M_{tjkl}: latent method factors, λ_{Mijkl}: method factor loadings, CM_{tjkl}: latent common method factors, λ_{CMijkl}: common method factor loadings, UM_{rtjkl}: latent unique method factors, λ_{UMijkl}: unique method factor loadings, ϵ_{rtijkl}: measurement errors at Level 1, and ϵ_{tijkl}: measurement errors at Level 2. Intercepts (α_{ijkl}) are not depicted in the figure. In this example, one of the structurally different methods (Method 1) serves as reference method.

pertaining to different methods (e.g., parents and peers). For example, correlations between the two method factors M_{tj2l} and CM_{tj3l} indicate whether peers and parents deviate in a similar ways (how a shared bias) from the self-reports on occasion of measurement l.

8.1. MEAN CHANGE

In addition to the investigation of the latent correlation as well as the variance components (provided in **Table 1**), many researchers seek to scrutinize the mean change of a particular construct across time. According to Equations (26–28), the expectation (mean) of the latent state factor can be identified as follows:

$$E(Y_{tij1l}) = E(\alpha_{ij1l}) + E(\lambda_{tij1l}S_{tj1l}) + E(\epsilon_{tij1l}), \tag{32}$$

$$= \alpha_{ij1l} + \lambda_{tij1l}E(S_{tj1l}). \tag{33}$$

Given that ϵ_{tij1l} is a zero-mean normally distributed residual variable, the latent mean of the latent state variables can be identified by setting one intercept for each latent state factor to zero

[e.g., $E(\alpha_{1j1l}) = 0$] and the corresponding factor loading to one [i.e., $\lambda_{S1j1l} = 1$]. Another possibility is to set the intercepts and factor loadings equal across time (i.e., assuming strong measurement invariance; see below) and to set the latent mean of the first latent state factor to zero. Then, the latent means of the remaining latent state variables reflect the true mean change of construct j from occasion of measurement 1 to occasion of measurement l.

8.2. TESTING MEASUREMENT INVARIANCE ACROSS TIME

Measurement invariance (MI) plays an important role in longitudinal analysis (Meredith, 1993; Widaman and Reise, 1997; Geiser et al., 2014). According to Widaman and Reise (1997) four different degrees of MI can be distinguished: (1) configural MI, (2) weak MI, (3) strong MI, and (4) strict MI. Configural MI implies that the number of factors as well as the factor structure as such is similar across different measurement occasions. In addition to configural MI, weak MI requires that the factor loading parameters for each indicator i are equal across different occasions of measurement. In addition to weak MI, strong MI assumes that the intercepts of indicators i are equal across different occasions of measurement. The most restrictive form of longitudinal MI (i.e., strict MI) implies that, in addition to the previous restrictions, the residual variances of indicators i are also equal across time. In the current work, we focus on one (not all possible) MI restriction that can be specified and empirically tested. In particular, we discuss the minimal set of restrictions that are necessary to meaningfully study mean change (i.e., strong MI) with respect to the reference method (e.g., self-reports). To meaningfully interpret mean change in the reference state factors, we recommend that the following MI restrictions be tested:

$$\alpha_{ij1l} = \alpha_{ij1l'} = \alpha_{ij1}, \qquad \text{where } l \neq l', \qquad (34)$$

$$\lambda_{Sij1l} = \lambda_{Sij1l'} = \lambda_{Sij1}, \qquad \text{where } l \neq l'. \qquad (35)$$

The above restrictions (34) and (35) state that the intercept and factor loading parameters of the reference state factors are time-invariant. These assumptions imply that the scale on which the reference latent state factors are measured does not change across time. Hence, researchers who are interested in studying mean change as measured by the reference method (e.g., self-reports) should at least establish strong MI as proposed above (see Equations 34–35). LS-COM models implying different degrees of MI can be compared by using a χ^2-difference test. For calculating level-specific χ^2-difference tests see Ryu and West (2009).

9. SIMULATION STUDY

To investigate the performance of the LS-COM model proposed throughout the previous sections, a Monte Carlo (MC) simulation study was performed. The main purpose of the simulation was to examine the applicability of the LS-COM model across a range of conditions and to establish a set of guidelines and recommendations concerning sample size and model complexity that ensure consistent and unbiased estimation of parameters and their standard errors and minimize potential estimation problems (so called Heywood cases).

9.1. RESULTS OF PREVIOUS SIMULATION STUDIES

Numerous simulation studies have been performed in the past focusing on the applicability and robustness of the single-level (classical) SEMs (e.g., Boomsma, 1982; Gerbing and Anderson, 1985; MacKinnon et al., 1995; Marsh et al., 1998; Fan et al., 1999; Raykov, 2000; Enders and Bandalos, 2001; Jackson, 2001; Bandalos, 2002). So far, only few simulation studies have been carried out investigating complex multilevel structural equation models (e.g., Satorra and Muthen, 1995; Hox and Maas, 2001; Julian, 2001; Stapleton, 2002; Maas and Hox, 2005) or longitudinal CFA-MTMM models (Crayen, 2008; Geiser, 2009). In this section, we briefly summarize the results of previous simulation studies that are most relevant to the present study.

With regard to single-level (classical) SEMs a ratio of 5 (sometimes 10) observations per parameter has been suggested to ensure reliable parameter estimates and standard errors (Bentler and Chou, 1987; Bollen, 1989, 2002). With regard to multilevel (two level) SEMs, simulation studies indicate that the number of Level 2 units are more important than the number of Level 1 units suggesting that at least 100 Level 2 units be sampled for accurate standard error estimates and for detecting small effects on Level 2 (Hox and Maas, 2001; Maas and Hox, 2005; Meuleman and Billiet, 2009). It has also been found that ignoring the multilevel structure completely can lead to biased parameter estimates as well as their standard errors (Julian, 2001). Recent simulation studies favor the use of Bayesian estimation techniques showing that 20 Level 2 units can be sufficient for reliable parameter estimates when using weakly informative priors (Hox et al., 2012). Nevertheless for maximum likelihood estimation, it has been generally recommended to sample at least 100 Level 2 units to ensure reliable parameter and standard error estimates (Hox and Maas, 2001; Maas and Hox, 2005; Meuleman and Billiet, 2009).

Simulation studies examining the statistical performance of longitudinal CFA-MTMM [i.e., CS-C(M-1)] models have shown that the parameter estimates and their standard errors are well-recovered in general. Nevertheless, the standard errors seem to be more sensitive to bias than the parameter estimates (Crayen, 2008; Geiser, 2009). Moreover, the statistical performance of the CS-C(M-1) model increases with larger sample sizes (i.e., more empirical informations), fewer constructs and methods (i.e., less complex models) and with low convergent validity (i.e., increasing method bias) (Crayen, 2008; Geiser, 2009).

Given that the CS-C(M-1) model by Geiser (2009) is a single-level confirmatory factor model, it is not clear to which extent the results described above apply to the LS-COM model. Similarly, the results of the simulation studies examining the performance of multilevel structural equation models (ML-SEM) may also not apply to the LS-COM model, given that the models used in those simulation studies are usually less complex (including only a few latent factors and no complex MTMM structure).

9.2. DESIGN OF THE SIMULATION STUDY

To investigate the effect of model complexity and sample size on estimation problems and precision it was necessary to vary a number of potentially influential factors. Because the LS-COM model is a longitudinal multilevel CFA-MTMM model, three main factors influence model complexity. To allow distinguishing

their influences (a) the number of constructs (1 vs. 2), (b) the number of methods (2 vs. 3), and (c) the number of occasions of measurement (2, 3, and 4) were varied independently.

In addition to these sources of model complexity, real-life applications of MTMM analysis vary greatly in the degree of convergent validity between the employed methods. To investigate whether convergent validity has an effect on the quality of the estimation this factor was also varied in two levels (high vs. low convergent validity). We used the coefficients of consistency, method specificity, and reliability to specify the true (population) model parameters. The degree of consistency and method specificity were allowed to differ across MC conditions, implying a condition of high consistency (i.e., high convergent validity) and a condition of low consistency (i.e., low convergent validity). The reliability of each indicator was obtained by the sum of the consistency and method specificity coefficients (range: 0.775–0.825). **Table 2** shows the population values for the different variance components for the different indicators.

Due to the multilevel structure of the LS-COM model sample size can be varied on the level of targets (Level 2) as well as on the level of the interchangeable raters (Level 1). As with model complexity, these two factors were varied independently of each other. The number of Level 2 units was set at 100, 250, and 500 targets (Level 2), while the number Level 1 units was set at 2, 5, 10, and 20 raters per target.

In total this simulation design resulted in $2 \times 2 \times 2 \times 3 \times 4 \times 3 = 288$ possible conditions. Of these 288 only 232 were included, because the remaining 56 conditions represented constellations in which the model is underidentified due to there being fewer targets than free model parameters. Of these 56 conditions 50 were conditions with only 100 Level 2 units and all but 8 were conditions represented models with 2 constructs.

Overall, 116,000 (232×500) data sets with a varying number or observations (200–10,000) were simulated using Mplus 6.1 (Muthén and Muthén, 2010), the free software `R` 2.14.0 (R Core Team, 2014), as well as various R packages such as `MplusAutomation` (Hallquist, 2011), `OpenMx` (Boker et al., 2011), and `corcounts` (Erhardt, 2013). All files of this simulation study can be downloaded from the following website[3]. An example Mplus syntax for the simulations study is provided in Appendix B in Supplementary Material.

Strong MI was assumed in all models (c.f. Widaman and Reise, 1997). All models were estimated using the maximum likelihood estimator implemented in M*plus* assuming multivariate normally distributed and complete data.

Table 2 | Consistency, method specificity and reliability of the LS-COM model.

	Low consistency		High consistency	
	Mean	**SD**	**Mean**	**SD**
Consistency	0.30	(±0.025)	0.60	(±0.025)
Unique method specificity	0.25	(±0.025)	0.10	(±0.025)
Common method specificity	0.25	(±0.025)	0.10	(±0.025)
Method specificity	0.50	(±0.050)	0.20	(±0.050)
Reliability	0.80	(±0.025)	0.80	(±0.025)

The variance coefficients above were standardized with regard to the observed variance of an indicator. Values in parentheses indicate the variation in standard deviations of the consistency and method specificity values across different indicators.

9.3. EVALUATION CRITERIA

The performance of the LS-COM model was examined using the following criteria: (a) rate of non-convergence after a maximum of 1000 iterations, (b) rate of improper solutions[4] (i.e., Heywood cases) due to non-positive definite covariance matrices Ψ and Θ, (c) the amount of parameter estimation as well as standard error bias, and (d) the accuracy of the χ^2-model fit statistics.

The absolute parameter bias was first calculated for each parameter p and then aggregated across all parameters of the same parameter type c for which effects were presumed to be equal (e.g., all common method factor loadings, λ_{CMij3l}; c.f. Bandalos, 2006):

$$peb(c) = \frac{1}{n_c} \sum_{c=1}^{n_c} \left(\frac{|M_{pc} - e_{pc}|}{e_{pc}} \right). \quad (36)$$

M_{pc} is the average of the MC parameter estimates across all 500 replications for parameter p of parameter type c, whereas e_{pc} is the true population value of that parameter. n_c is the number of parameters in cluster c.

In a similar way, the absolute standard error bias was calculated:

$$seb(c) = \frac{1}{n_c} \sum_{c=1}^{n_c} \left(\frac{|M_{SE_{pc}} - SD_{pc}|}{SD_{pc}} \right). \quad (37)$$

$M_{SE_{pc}}$ is the average standard error of parameter p allotted to parameter type c across all 500 MC replications, whereas SD_{pc} is the standard deviation of the parameter estimate for parameter p in cluster c across all 500 MC replications.

The aggregation of the absolute parameter estimation and standard error biases was done for two reasons. First, the LS-COM model incorporates many free parameters to be estimated (sometimes more than 100) and it would not be feasible to report bias for each single model parameter. Second, it is reasonable to assume that similar parameters (e.g., all measurement error variances) are biased in a similar way. Hence, by aggregating parameters that belong to the same parameter type, it was possible to investigate general bias in parameter estimates and their standard errors. In total 12 types of parameters were defined. Eight of these stemmed from the between part of the model: (1) the state factor loadings (λ_S), (2) the common method factor loadings (λ_{CM}), (3) the method factor loadings (λ_M), (4) the covariances of latent variables on Level 2 (cov_{L2}), (5) the latent means (μ), (6) the latent intercepts (α), (7) the variance of the latent variables at Level 2 (var_{L2}), and (8) the Level 2 residual variances (ϵ_{L2}). The

[3] http://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/psymeth/mitarbeiter/tkoch/index.html.

[4] Ψ-warning messages indicate linear dependencies in the covariance matrix of the latent variables, whereas Θ-warning messages indicate estimation problems with regard to the latent error variables (e.g., negative error variances).

remaining four parameter clusters all pertained to parameters at Level 1: (9) the unique method factor loadings (λ_{UM}), (10) the unique method factor variances (var_{L1}), (11) the covariances of the unique method factors (cov_{L1}), and (12) the Level 1 residual variances (ϵ_{L1}).

In line with previous MC simulation studies investigating MTMM-SEMs (e.g., Nussbeck et al., 2006; Geiser, 2009) 0.10 was chosen as a cut-off criterion for both parameter and SE biases, and absolute values beyond this threshold were deemed unacceptable.

10. RESULTS

10.1. RATE OF NON-CONVERGENCE

All 116,000 specified LS-COM (H0) models converged properly within 1000 iterations.

10.2. RATE OF IMPROPER SOLUTIONS

M*plus* warning messages regarding potential Ψ-problems were encountered in 65 out of 232 (28.02%) MC conditions, but in only 2,366 out of 116,000 (2.04%) total replications in the simulation. The main reason for the Ψ-warning messages were linear dependencies in the latent covariance matrix due to higher order partial correlations above |1|. Moreover, only 2 out of 232 MC conditions contained improper solutions with regard to latent residual matrix Θ. Hence, the actual amount of improper solutions with regard to this simulation study was below 5%.

Most of the conditions exhibiting general warning messages were high consistency conditions (i.e., 56 MC conditions and 2,306 out of 116,000 replications, 1.99%) and only few were low consistency conditions (i.e., 9 MC conditions and 60 out of 116,000 replications, 0.05%) Moreover, the frequency of Ψ-warning messages decreased with increasing sample size on Level 1 (number of raters per target) as well as with increasing sample size on Level 2 (number of targets). **Figure 5** shows the relationship between the average amount of Ψ-warning messages and the sample size on both levels in the low and the high consistency conditions. **Figure 5** shows that the amount of Ψ-warning messages decreased substantially with the number of targets as well as with the number of raters per target. **Figure 5** also indicates that the number of raters per target might be more important for the reduction of Ψ-warning messages than the number of Level 2 units (here: targets).

10.3. AMOUNT OF PARAMETER AND STANDARD ERROR BIAS

Across all 232 conditions the absolute parameter estimation bias (peb, see Equation 36) was below the cutoff value of 10%. However, the absolute standard error bias (seb, see Equation 37) exceeded the value of 10% in 21 out of 232 MC conditions. Higher seb values were more often found in the high consistency (14 out of 21 conditions, 66.67%) conditions than in the low consistency (7 out of 21 conditions, 33.33%) conditions. **Figure 6** shows the average peb and seb values across all parameters with respect to the sample size on Level 1 and Level 2 as well as with respect to the consistency condition (high vs low).

Figure 6 shows that the average peb and seb values decreased with increasing sample size on Level 1 and Level 2. In particular, the sample size on Level 1 (number of raters per target) seemed to be crucial for the reduction of the seb. Moreover, **Figure 6** shows that the average amount of peb and seb was lower in the low consistency condition than in the high consistency condition. Note that the average peb and seb (i.e., across all parameters) were below 10% (see **Figure 6**). Further investigations revealed that specific LS-COM model parameters were more sensitive to bias than others. Specifically, the common method factor loadings λ_{CM}, method factor loadings λ_{M}, unique method factor loadings λ_{UM}, as well as the variances of unique method factors var_{L1} showed the largest standard error biases. Additionally, the seb of the latent means on Level 2 exceeded the cutoff value of 10% in one single MC condition (i.e., one construct, two methods, two occasions of measurement, 10 Level 1 units and 100 Level 2 units). **Figure 7** shows the dependency of the seb values on the sample size at each of the measurement levels in the high and low consistency condition.

According to **Figure 7**, the standard error bias was substantially reduced with increasing sample size on both levels. In particular, the standard error bias dropped below the cutoff value of 10% when more than 2 raters per target were sampled.

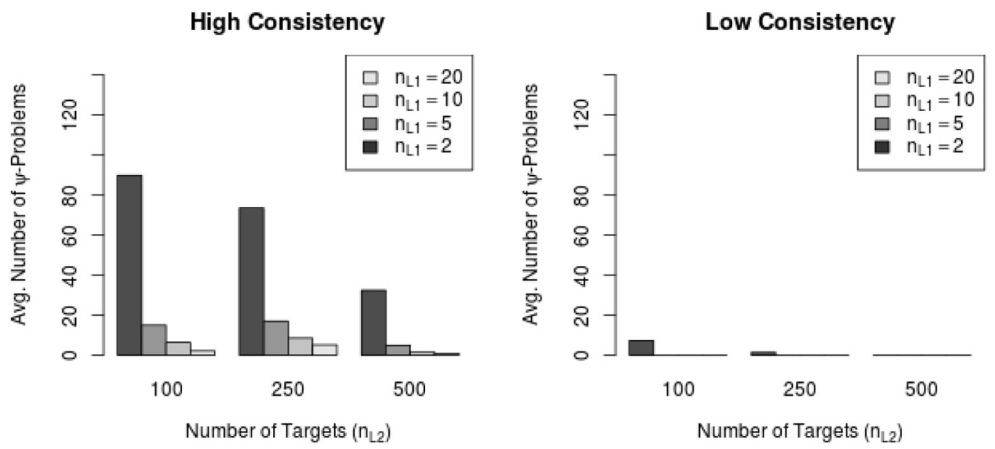

FIGURE 5 | Average number of Ψ-warning messages in high and low consistency conditions. n_{L1} = number of Level 1 units; n_{L2} = number of Level 2 units.

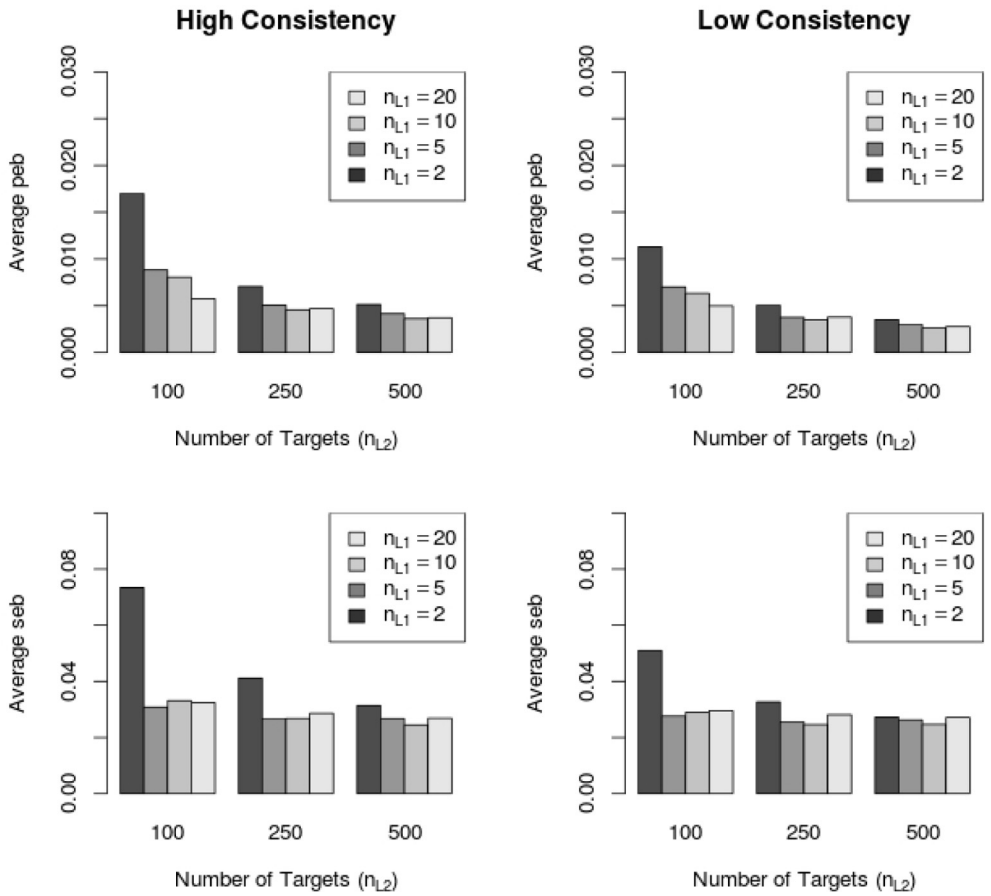

FIGURE 6 | Average peb and seb values with respect to sample size in high and low consistency conditions in the LS-COM model. n_{L1} = number of Level 1 units; n_{L2} = number of Level 2 units.

10.4. χ^2-FIT-STATISTICS

In **Figure 8A,B** the simulated and expected proportions of the χ^2 values for monoconstruct and multiconstruct LS-COM models are presented. According to these results, the simulated χ^2-values were always below the theoretically expected χ^2-values indicating a downward bias in the asymptotic type I error. These results suggest that too many specified LS-COM models would be accepted with respect to a nominal alpha level of 0.05 if researchers used the theoretical χ^2 distribution to test the model fit. Hence, the χ^2 model fit test appeared to be too liberal with respect to LS-COM models under the conditions studied here. However, the differences between the observed and the expected χ^2-distributions at a nominal alpha level of 5% were relatively small (on average 0.03 for monoconstruct condition and 0.04 for the multiconstruct condition). The results also indicate that the χ^2 model fit test was more accurate for less complex (i.e., monoconstruct) LS-COM models. We did not find a straightforward relationship between sample size and the accuracy of the χ^2 model fit test for the LS-COM model.

11. DISCUSSION

In the present work a multilevel longitudinal CFA-MTMM model for the combination of structurally different and interchangeable methods (called LS-COM model) was proposed. The LS-COM model combines the advantages of multilevel, longitudinal, and CFA-MTMM modeling approaches and is suitable for MTMM measurement designs combining different types of methods. Given that such complex MTMM measurement designs are increasingly used in psychology (e.g., 360° feedback designs, multisource, mutirater designs), the LS-COM fills a gap in the current literature on longitudinal MTMM modeling. Previous studies on longitudinal MTMM modeling have either focused exclusively on single-indicator models or on a specific type of method (e.g., structurally different methods) (e.g., Kenny and Zautra, 2001; Burns and Haynes, 2006; Courvoisier et al., 2008; Grimm et al., 2009; Geiser et al., 2010). In the present article a new CFA-MTMM model has been developed allowing the simultaneous analysis of different types of methods (i.e., structurally different and interchangeable methods) across time using a multiple indicator, multilevel latent variable approach. The LS-COM model overcomes many limitations of previous models by allowing researchers to

1. study method effects on different levels (rater and target level),
2. analyze the stability and change of construct and method effects across time,

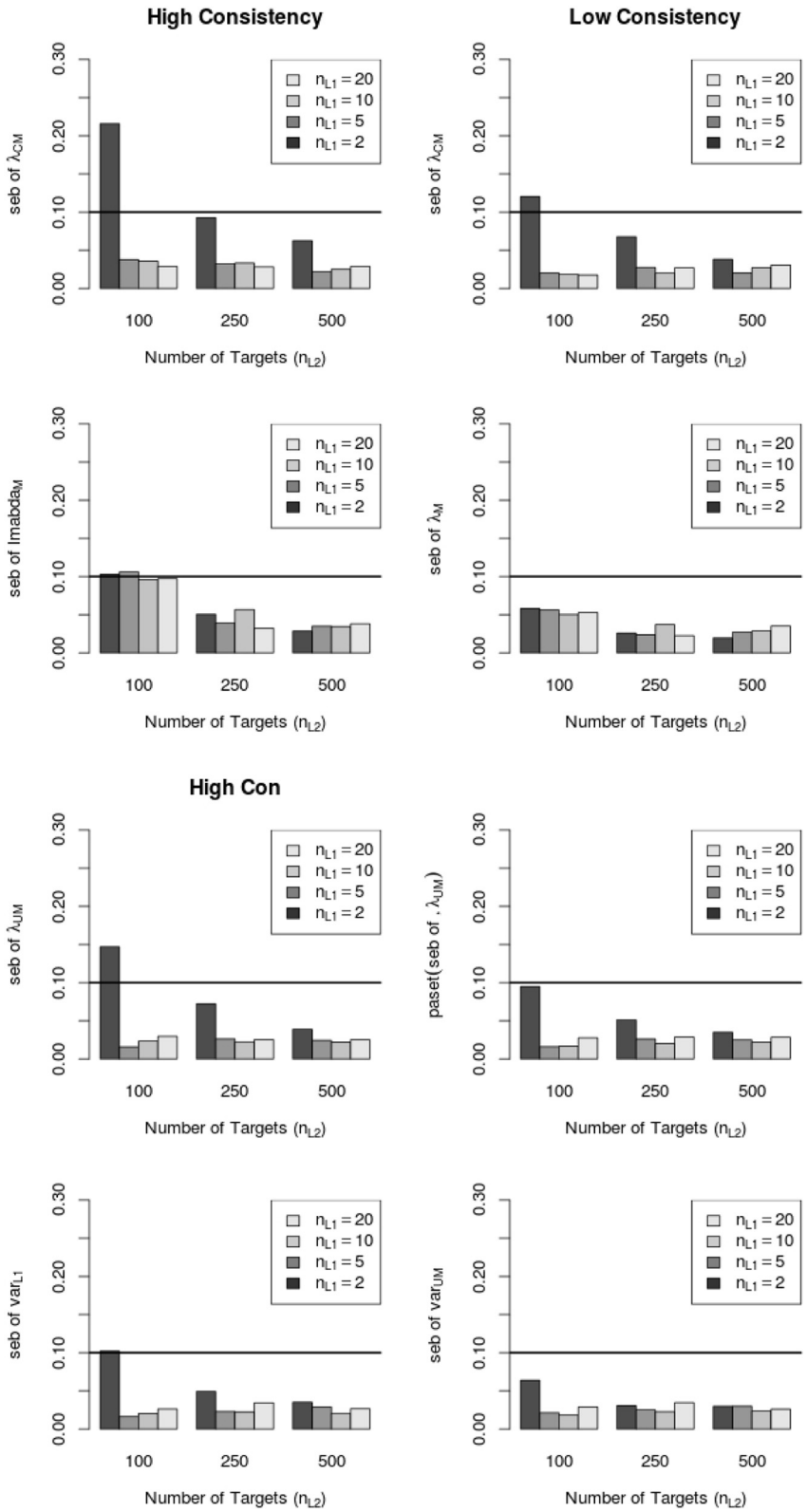

FIGURE 7 | Relationship between average standard error bias (seb) and sample size for different LS-COM model parameters in the high and low consistency condition. λ_{CM} = common method factor loading parameters; λ_M = method factor loading parameters; λ_{UM} = unique method factor loading parameters; var_{L1} variance of the unique method factors.

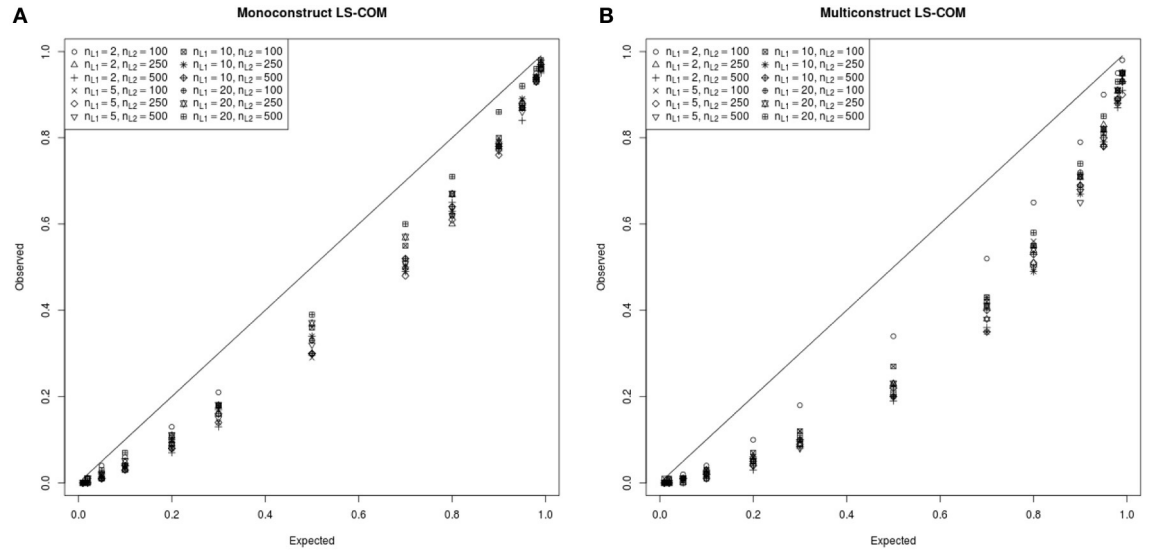

FIGURE 8 | PP-plot of the observed and theoretical proportions of the χ^2 values for the monoconstruct LS-COM model (A) and multiconstruct LS-COM model (B) for different sample sizes.

3. evaluate the convergent and discriminant validity among different methods across time,
4. investigate the stability and change of a given construct (attribute) across time,
5. examine different variance coefficients and the psychometric properties of the measures on multiple occasions of measurement,
6. test important assumptions (e.g., measurement invariance), and
7. study potential causes of method effects by including external variables in the model.

Moreover, the LS-COM model is defined based on the stochastic measurement theory (Suppes and Zinnes, 1963; Zimmerman, 1975; Steyer and Eid, 2001), which bears the advantage of defining the latent variables as random variables with a clear psychometric interpretation. That means that the latent variables in the LS-COM model are not simply assumed, but properly defined as random variables on the probability space (see Appendix A in Supplementary Material for the formal definitions). In addition, the LS-COM makes use of a latent regression modeling [CT-C(M-1)] approach, which allows contrasting different methods against a reference method. The CT-C(M-1) modeling approach bears the advantages of using "pure" method factors by defining the method variables as latent residual variables (see Geiser et al., 2008, for more details). In addition, the CT-C(M-1) modeling approach allows separating the total variance of each indicator into state, method, and measurement error components and calculating different variance coefficients (e.g., coefficients of consistency, method specificity, reliability), which is not possible in other MTMM modeling approaches [e.g., latent means (Pohl and Steyer, 2010) and latent difference modeling (Pohl et al., 2008) approaches].

Researchers who are interested in studying the mean change of an attribute across time should first test the degree of measurement invariance and then estimate the latent means of the latent state factors as described above. In addition, the stability and change of the interindividual differences in an attribute can be investigated by the correlations of the latent state factors pertaining to different occasions of measurement. The stability and change of the method factors across time can be studied with regard to the correlations between the latent method factors measured on different occasions of measurement. In total, three different types of method effects can be examined. First, the method effects of the structurally different method (e.g., parent reports) that is not shared with the reference method (e.g., self-reports). Second, the common method effect of the interchangeable methods (e.g., general peer rating) that is not shared with the reference method (e.g., self-report). Third, the unique method effect of the interchangeable methods (e.g., single peer rating) that is neither shared with the reference method (e.g., self-reports), nor with other peers. A meaningful interpretation of correlation coefficients between method factors across time (e.g., as stability of method effects), typically requires that the same raters are recruited at each time point. The generalizability of the method effects can be examined by the correlations of latent method factors pertaining to different types of methods (structurally different and interchangeable methods).

In order to examine the trustworthiness of the parameter and standard error estimates in the LS-COM model, we conducted a MC simulation study. To our knowledge, no simulation study has been performed so far scrutinizing the statistical performance of complex longitudinal, multilevel, multiple indicator CFA-MTMM models.

According to the results of our MC simulation study, the LS-COM model can produce reliable parameter estimates even

in small samples with just 100 targets and 2 raters per targets. However, for such small samples the standard errors of LS-COM model parameters will be marginally biased. Most sensitive to bias are the standard errors of the method factor loading parameters (i.e., λ_{UMijkl}, λ_{CMijkl}, λ_{Mijkl}) as well as the standard errors of the unique method factor variance [i.e., $Var(UM_{rtjkl})$]. The standard error bias can be reduced by increasing the number of Level 1 units (i.e., number of raters per target). In cases with at least 5 raters per target and 100 targets, the LS-COM produced unbiased parameters as well as standard errors in our simulation. In general, parameter estimates seemed more accurate in cases with low convergent validity. Low convergent validity is often seen in practice (e.g., Eid et al., 2003, 2008; Carretero-Dios et al., 2011; Pham et al., 2012), so that the LS-COM model should generally result in unbiased parameter and SE estimates.

The number of methods as well as the number of occasions of measurement did not seem to affect the accuracy of the parameter estimation or their standard errors. If at all, more occasions of measurement proved beneficial for the stability of the parameter estimation. This is most likely due to the fact that strong MI was assumed for the repeated measures in the simulation. Because of this, the ratio of available information to free parameters actually increased with more measurement occasions. It should be noted, however, that this condition might not be present in applications in which the assumption of strong MI does not hold or the number of occasions is very large.

In contrast to the number measurement occasions an increasing number of constructs generally does make the LS-COM model more complex, because invariance assumptions are generally not imposed across different constructs. In cases with many constructs, we recommend splitting the complete LS-COM model into multiple submodels and analyzing all combinations using two constructs simultaneously. All coefficients of interest (e.g., correlations) can still be estimated without affecting the meaning of any parameter in the model. A prerequisite for the step-by-step procedure is that the same reference method is chosen.

The results of this simulation study support previous findings of classical SEM (see Bentler and Chou, 1987; Bollen, 1989, 2002). Based on a simulation study, Bentler and Chou (1987) suggested that a ratio of 5:1 (observations per parameter) is sufficient for proper parameter estimates with regard to classical structural equation models. The results of our simulation study support this conclusion for LS-COM models. We therefore recommend sampling at least 5 raters per target and at least as many targets as there are free parameters to be estimated. Our simulation study also revealed new insights into complex multilevel SEM, namely that the sample size on Level 1 is an important factor that influences the quality of model estimation. Previous simulation studies devoted to this research area claimed that the number of Level 1 units is less important than the number of Level 2 units (Maas and Hox, 2005). Our results show that the number of Level 1 units can be crucial for the reduction of standard error bias in complex multilevel structural equation models.

So far, only few studies have investigated the accuracy of χ^2-fit-statistics in complex ML-SEMs (Ryu and West, 2009; Ryu, 2014; Schermelleh-Engel et al., 2014). The results of our simulation study are generally encouraging as they indicated that the overall χ^2-test of exact fit was only marginally biased with regard to a nominal alpha level of 5% and multivariate normal distributed and complete data. More specifically, our results indicate that the overall maximum likelihood χ^2-test of exact fit may be slightly too liberal for complex ML-SEM models. However, we recommended to use robust maximum likelihood estimation (MLR) when multivariate normality cannot be assumed.

Future studies should focus on three issues associated with complex longitudinal multilevel MTMM modeling. First, the statistical effects of attrition (i.e., missingness) of the interchangeable raters across time and the possibilities of alternative modeling approaches should be investigated. Second, the robustness of χ^2 fit statistics in complex multilevel SEM with non-normal and (un)complete data should be examined and alternative fit statistics for complex multilevel SEMs should be scrutinized. With respect to the investigation of fit statistics in multilevel SEM, researchers maybe inspired by the recent work of Schermelleh-Engel et al. (2014) and Ryu (2014). Third, future studies should focus on possible extensions of the LS-COM model to the other longitudinal modeling approaches [e.g., latent state-trait models, latent difference (change) models, latent growth curve models] with one or more sets of interchangeable methods and apply these models to real data.

12. CONCLUSION AND GENERAL RECOMMENDATION

In this work, we presented a new longitudinal multilevel CFA-MTMM model for the combination of structurally different and interchangeable methods. The model extends the spectrum of longitudinal MTMM modeling approaches by allowing the simultaneous investigation of method effects on different measurement levels across time. With respect to the results of the simulation study, we recommend that researchers should sample at least as many Level 2 units (i.e., targets) as there are free parameters to be estimated in the model and at least 5 interchangeable raters per target in order to obtain a reliable sample size for proper parameter standard error estimation. Moreover, we suggest that researchers should test the degree of MI when studying mean change of a given attribute across time.

AUTHOR NOTE

Dr. Tobias Koch: Note that the psychometric definition of the LS-COM model and the simulation study were previously presented as part of the doctoral thesis by Koch (2013). This research was funded by the German Research Foundation (Deutsche Forschungsgesellschaft, DFG, grant number: EI 379/6-1). Christian Geiser's work was funded by a grant from the National Institutes on Drug Abuse (NIH-NIDA), grant number: 1 R01 DA034770-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

SUPPLEMENTARY MATERIALS

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00311/abstract

REFERENCES

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Struct. Equ. Model.* 9, 78–102. doi: 10.1207/S15328007SEM0901_5

Bandalos, D. L. (2006). "The use of monte carlo studies in structural equation modeling research," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age Publishing), 385–426.

Bentler, P., and Chou, C.-P. (1987). Practical issues in structural modeling. *Sociol. Methods Res.* 17, 78–117. doi: 10.1177/0049124187016001004

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76, 306–317. doi: 10.1007/s11336-010-9200-6

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley series in probability and mathematical statistics. Applied probability and statistics section. New York, NY: John Wiley & Sons, pp 0271–6356.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53, 605–634. doi: 10.1146/annurev.psych.53.100901.135239

Bollen, K. A., and Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective.* New York, NY: Wiley.

Boomsma, A. (1982). *The robustness of LISREL against small sample sizes in factor analysis models.*, pages 149–173. Amsterdam: North-Holland.

Burns, G. L., and Haynes, S. N. (2006). *Clinical Psychology: Construct Validation with Multiple Sources of Information and Multiple Settings* (Washington, DC: APA Publications), 401–418.

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Carretero-Dios, H., Eid, M., and Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: an application to the validation of the State-Trait Cheerfulness Inventory. *J. Res. Pers.* 45, 153–164. doi: 10.1016/j.jrp.2010.12.007

Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C., and Cole, D. A. (2008). Analyzing the convergent and discriminant validity of states and traits: development and applications of multimethod latent state-trait models. *Psychol. Assess.* 20, 270–280. doi: 10.1037/a0012812

Crayen, C. (2008). *Eine Monte-Carlo Simulationsstudie zur Untersuchung der Anwendbarkeit von Strukturgleichungsmodellen für Multitrait-Multimethod-Multioccasion Daten. [A simulation study for investigating the applicability of structural equation models for multitrait-multimethod-multioccasion data]*, (Unpublished diploma thesis.) Freie Universität Berlin, Berlin: Germany.

Dumenci, L. (2000). *Multitrait-Multimethod Analysis.* San Diego, CA: Academic Press, 583–611.

Duncan, S. C., Duncan, T. E., and Strycker, L. A. (2006). Alcohol use from ages 9 to 16: a cohort-sequential latent growth model. *Drug Alcohol Depend.* 81, 71–81. doi: 10.1016/j.drugalcdep.2005.06.001

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika* 65, 241–261. doi: 10.1007/BF02294377

Eid, M., and Diener, E. (2006). *Introduction: The Need for Multimethod Measurement in Psychology*. Wahington, DC: American Psychological Association, 3–8.

Eid, M., and Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *J. Educ. Behav. Stat.* 23, 193–215. doi: 10.2307/1165244

Eid, M., Lischetzke, T., and Nussbeck, F. W. (2006). "Structural equation models for multitrait-multimethod data," in *Handbook of Multimethod Measurement in Psychology*, eds M. Eid and E. Diener (Washington, DC: American Psychological Association), 283–299.

Eid, M., Lischetzke, T., Nussbeck, F. W., and Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C(M-1) model. *Psychol. Methods* 8, 38–60. doi: 10.1037/1082-989X.8.1.38

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., and Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: different models for different types of methods. *Psychol. Methods* 13, 230–253. doi: 10.1037/a0013219

Enders, C. K., and Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Struct. Equa. Model.* 8, 430–457. doi: 10.1207/S15328007SEM0803_5

Erhardt, E. (2013). *Corcounts: Generate Correlated Count Random Variables*. R package version 1.4 [online]. Available online at: http://cran.r-project.org/web/packages/corcounts/index.html (Accessed March, 2014).

Fan, X., Thompson, B., and Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Struct. Equ. Model. Multidisc. J.* 6, 56–83. doi: 10.1080/10705519909540119

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

Geiser, C., Eid, M., Nussbeck, F., Courvoisier, D. S., and Cole, D. A. (2010). Multitrait-multimethod change modelling. *Adv. Stat. Anal.* 94, 185–201. doi: 10.1007/s10182-010-0127-0

Geiser, C., Eid, M., and 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). *Psychol. Methods* 13, 49–57. doi: 10.1037/1082-989X.13.1.49

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A., and Koch, T. (2014). Distinguishing state variability from trait change in longitudinal data: the role of measurement (non) invariance in latent state-trait analyses. *Behav. Res. Methods*, 1–32. doi: 10.3758/s13428-014-0457-z. Available online at: http://link.springer.com/article/10.3758/s13428-014-0457-z

Geiser, C., Koch, T., and Eid, M. (in press). Data-generating mechanisms versus constructively-defined latent variables in multitrait-multimethod analyses: a comment on Castro-Schilo, Widaman, and Grimm (2013). *Struct. Equ. Modeling*.

Gerbing, D. W., and Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behav. Res.* 20, 255–271. doi: 10.1207/s15327906mbr2003_2

Grimm, K. J., Pianta, R. C., and Konold, T. (2009). Longitudinal multitrait-multimethod models for developmental research. *Multivariate Behav. Res.* 44, 233–258. doi: 10.1080/00273170902794230

Hallquist, M. (2011). *MplusAutomation: Automating Mplus Model Estimation and Interpretation*. R package version 0.5 [online]. Available online at: http://cran.r-project.org/web/packages/MplusAutomation/index.html (Accessed March, 2014).

Hancock, G. R., Kuo, W.-L., and Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Struct. Equ. Model.* 8, 470–489. doi: 10.1207/S15328007SEM0803_7

Heck, R. H., Thomas, S. L., and Tabata, L. N. (2013). *Multilevel and Longitudinal Modeling with IBM Spss*. New York, NY: Routledge.

Hertzog, C., and Nesselroade, J. R. (1987). Beyond autoregressive models: some implications of the trait-state distinction for the structural modeling of developmental change. *Child Dev.* 58, 93–109. doi: 10.2307/1130294

Hox, J. J. (2010). *Multilevel Analysis Techniques and Applications, 2nd Edn*. New York, NY: Routledge.

Hox, J. J., and Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Equ. Model.* 8, 157–174. doi: 10.1207/S15328007SEM0802_1

Hox, J., van de Schoot, R., and Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a bayesian perspective. *Surv. Res. Methods* 6, 87–93. Available online at: http://mplus.fss.uu.nl/files/2012/07/hox-schoot-2012-how-few-countries-will-do.pdf

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: a monte carlo investigation. *Struct. Equ. Model.* 8, 205–223. doi: 10.1207/S15328007SEM0802_3

Jöreskog, K. G. (1979a). *Statistical Models and Methods for Analysis of Longitudinal Data.* Cambridge, MA: Abt Books, 129–169.

Jöreskog, K. G. (1979b). "Statistical estimation of structural models in longitudinal-developmental investigations," in *Longitudinal Research in the Study of Behavior and Development*, eds J. R. Nesselroade and P. B. Baltes (New York, NY: Academic Press), 303–351.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Struct. Equ. Model.* 8, 325–352. doi: 10.1207/S15328007SEM0803_1

Kenny, D. A., and Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychol. Bull.* 112, 165–172. doi: 10.1037/0033-2909.112.1.165

Kenny, D. A., and Zautra, A. (2001). "Trait-state models for longitudinal data," in *New Methods for the Analysis of Change.* Decade of behavior, eds L. M. Collins and A. G. Sayer (Washington, DC: American Psychological Association), 243–263. doi: 10.1037/10409-008

Khoo, S.-T., West, S. G., Wu, W., and Kwok, O.-M. (2006). "Longitudinal methods." in *Handbook of Psychological Measurement: A Multimethod Perspective*, eds M. Eid and E. Diener (Washington, DC: American Psychological Association), 301–317.

Koch, T. (2013). *Multilevel Structural Equation Modeling of Multitraitmultimethod-Multioccasion Data*, Unpublished doctoral thesis, Freie Universität Berlin, Berlin. Available online at: http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000094645. (Accessed March, 2014).

Koch, T, Eid, M. and Lochner, K. (in press). "Multitrait-Multimethod-Analysis: the psychometric foundation of CFA-MTMM models," in *The Wiley Handbook of Psychometric Testing.* (London: John Wiley & Sons).

Little, T. D., Schnabel, K. U., and Baumert, J. (2000). *Modeling Longitudinal and Multilevel Data*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-2241.1.3.86

MacKinnon, D. P., Warsi, G., and Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behav. Res.* 30, 41–62. doi: 10.1207/s15327906mbr3001_3

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Appl. Psychol. Measur.* 13, 335–361. doi: 10.1177/014662168901300402

Marsh, H. W. (1993). Multitrait-multimethod analyses: inferring each trait-method combination with multiple indicators. *Appl. Measur. Edu.* 6, 49–81. doi: 10.1207/s15324818ame0601_4

Marsh, H. W., and Grayson, D. (1994). Longitudinal confirmatory factor analysis: common, time-specific, item-specific, and residual-error components of variance. *Struct. Equ. Model.* 1, 116–145. doi: 10.1080/10705519409539968

Marsh, H. W., and Grayson, D. (1995). *Latent Variable Models of Multitrait-Multimethod Data.* Thousand Oaks: Sage Publications, Inc, 177–198.

Marsh, H. W., Hau, K.-T., Balla, J. R., and Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behav. Res.* 33, 181–220. doi: 10.1207/s15327906mbr3302_1

Marsh, H. W., and Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: application of second-order confirmatory factor analysis. *J. Appl. Psychol.* 73, 107–117. doi: 10.1037/0021-9010.73.1.107

McArdle, J. J. (1988). "Dynamic but structural equation modeling of repeated measures data," in *Handbook of Multivariate Experimental Psychology*, eds R. B. Cattell and J. Nesselroade (New York, NY: Plenum Press), 561–614.

McArdle, J. J., and Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Dev.* 58, 110–133. doi: 10.2307/1130295

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/BF02294746

Meuleman, B., and Billiet, J. (2009). A monte carlo sample size study: how many countries are needed for accurate multilevel sem? *Surv. Res. Methods* 3, 45–58. Available online at: http://hdl.handle.net/1854/LU-1041001.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Muthén, L. K., and Muthén, B. O. (1998–2010). *Mplus User's Guide, 6th Edn* Los Angeles, CA: Muthén & Muthén.

Nussbeck, F. W., Eid, M., and Lischetzke, T. (2006). Analysing multitrait-multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: what sample size is needed for valid results? *Br. J. Math. Stat. Psychol.* 59, 195–213. doi: 10.1348/000711005X67490

Pham, G., Koch, T., Helmke, A., Schrader, F. W., Helmke, T., and Eid, M. (2012). Do teachers know how their teaching is perceived by their pupils? *Procedia Soc. Behav.* 46, 3368–3374. doi: 10.1016/j.sbspro.2012.06.068

Pohl, S., and Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behav. Res.* 45, 45–72. doi: 10.1080/00273170903504729

Pohl, S., Steyer, R., and Kraus, K. (2008). Modelling method effects as individual causal effects. *J. R. Stat. Soc. Ser. A*, 171, 41–63. doi: 10.1111/j.1467-985X.2007.00517.x

Raykov, T. (2000). On sensitivity of structural equation modeling to latent relation misspecifications. *Struct. Equ. Model.* 7, 596–607. doi: 10.1207/S15328007SEM0704_4

Rabe-Hesketh, S., and Skrondal, A. (2004). Generalized multilevel structural equation modeling. *Psychometrica* 69, 167–190. doi: 10.1007/BF02295939

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/. (Accessed March, 2014).

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Front. Psychol.* 5:81. doi: 10.3389/fpsyg.2014.00081

Ryu, E., and West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Struct. Equ. Model.* 16, 583–601. doi: 10.1080/10705510903203466

Satorra, A., and Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociol. Methodol.* 25, 267–316. doi: 10.2307/271070

Schermelleh-Engel, K., Kerwer, M., and Klein, A. G. (2014). Evaluation of model fit in nonlinear multilevel structural equation modeling. *Front. Psychol.* 5:181. doi: 10.3389/fpsyg.2014.00181

Schultze, M. (2012). *Evaluating What the Crowd Says. A Longitudinal Structural Equation Model for Exchangeable and Structurally Different Methods for Evaluating Interventions*. Diploma thesis, Freie Universität Berlin.

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford university press. doi: 10.1093/acprof:oso/9780195152968.001.0001

Snijders, T. A. B., and Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Struct. Equ. Model.* 9, 475–502. doi: 10.1207/S15328007SEM0904_2

Steele, F. (2008). Multilevel models for longitudinal data. *J. R. Stat. Soc. Ser. A* 171, 5–19. doi: 10.1111/j.1467-985X.2007.00509.x

Steyer, R. (1988). *Experiment, Regression und Kausalität: Die Logische Struktur Kausaler Regressionsmodelle [Experiment, Regression and Causality: The Logical Structure of Causal Regression Models]*. Unpublished thesis, Universität Trier.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika* 3, 25–60.

Steyer, R. (1995). Das MTMM-Modell ist nicht identifiziert [The MTMM model is not identified]. *Newsletter der Fachgruppe Methoden*, 3, 5.

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology* 1:39. doi: 10.1027/1614-1881.1.1.39

Steyer, R., and Eid, M. (2001). *Messen und Testen. [Measurement and testing.], 2nd Edn*. Heidelberg: Springer. doi: 10.1007/978-3-642-56924-1

Steyer, R., Eid, M., and Schwenkmezger, P. (1997). Modeling true intraindividual change: true change as a latent variable. *Methods Psychol. Res.* 2, 21–33.

Steyer, R., Ferring, D., and Schmitt, M. J. (1992). States and traits in psychological assessment. *Eur. J. Psychol. Assess.* 8, 79–98.

Steyer, R., Partchev, I., and Shanahan, M. J. (2000). *Modeling True Intraindividual Change in Structural Equation Models: The Case of Poverty and Children's Psychosocial Adjustment.* Mahwah, NJ: Lawrence Erlbaum, 109–126.

Steyer, R., Schmitt, M., and Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *Eur. J. Pers.* 13, 389–408.

Suppes, P., and Zinnes, J. L. (1963). *Basic Measurement Theory*, Vol. 1. Wiley, New York, 1–76.

Tisak, J., and Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychol. Methods* 5, 175–198. doi: 10.1037/1082-989X.5.2.175

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. J., and Muthén, B. (2013). Facing off with scylla and charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Measur.* 9, 1–26. doi: 10.1177/014662168500900101

Widaman, K. F., and Reise, S. P. (1997). *Exploring the Measurement Invariance of Psychological Instruments: Applications in the Substance Use Domain.* Washington, DC: American Psychological Association, 281–324.

Wothke, W. (1995). "Covariance components analysis of the multitrait-multimethod matrix," in *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, ed P. E. Shrout (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.), 125–144.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika* 40, 395–412. doi: 10.1007/BF02291765

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 November 2013; accepted: 25 March 2014; published online: 17 April 2014.

Citation: Koch T, Schultze M, Eid M and Geiser C (2014) A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. Front. Psychol. ***5****:311. doi: 10.3389/fpsyg.2014.00311*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.