Analyzing True Change in Longitudinal Multitrait-Multimethod Studies: Application of a Multimethod Change Model to Depression and Anxiety in Children

Christian Geiser and Michael Eid
Freie Universität Berlin

Fridtjof W. Nussbeck
University of Zürich

Delphine S. Courvoisier
University of Geneva

David A. Cole
Vanderbilt University

The authors show how structural equation modeling can be applied to analyze change in longitudinal multitrait-multimethod (MTMM) studies. For this purpose, an extension of latent difference models (McArdle, 1988; Steyer, Eid, & Schwenkmezger, 1997) to multiple constructs and multiple methods is presented. The model allows investigators to separate true change from measurement error and to analyze change simultaneously for different methods. The authors also show how Campbell and Fiske's (1959) guidelines for analyzing convergent and discriminant validity can be applied to the measurement of latent change. The practical application of the multimethod change model is illustrated in a reanalysis of child depression and anxiety scores ($N = 906$ American children) that were assessed by self- and parent reports on three measurement occasions. The analyses revealed that (a) the convergent validity of change was low for both constructs and (b) sex was a significant predictor of self-reported, but not of parent reported, anxiety states. Finally, the authors discuss advantages and limitations and compare the model with other approaches for analyzing longitudinal MTMM data.

Keywords: multitrait-multimethod analysis, longitudinal confirmatory factor analysis, convergent and discriminant validity, latent change analysis, latent difference models

Supplemental materials: http://dx.doi.org/10.1037/a0017888.supp

Eid and Diener (2006) observed that multimethod research designs are nowadays preferred to single method designs in almost all areas of psychological research. Developmental psychology is no exception, as is proven by an increasing number of longitudinal studies in the field that use multitrait-multimethod (MTMM) designs. Recent examples of such multitrait-multimethod-multioccasion (MTMM-MO) studies include multimethod investigations of adolescent popularity, social adaptation, and deviant behavior (Allen, Porter, McFarland, Marsh, & McElhaney, 2005); childhood depression and anxiety (Cole, Martin, Peeke, Henderson, & Harwell, 1998; Tram & Cole, 2006); parenting effects on the mental health of bereaved children (Kwok et al., 2005); temperament in early childhood (Majdandzic & van den Boom, 2007); the development of aggressive behavior in children (Ostrov &

Crick, 2007); and the relation between parental warmth/positive expressiveness and children's empathy-related responding and social functioning (Zhou et al., 2002).

In this article, our goal is to demonstrate how structural equation modeling (SEM) can be used to analyze change in MTMM-MO studies in a comprehensive and sophisticated way. SEM allows researchers to test an overall theoretical model and to account explicitly for measurement error in the observed (change) scores. Furthermore, SEM has been applied successfully to the analysis of cross-sectional MTMM data (e.g., Dumenci, 2000; Eid, Lischetzke, & Nussbeck, 2006; Marsh, 1989; Widaman, 1985; Wothke, 1995).

MTMM-MO designs have become increasingly popular for several reasons. First, panel studies that rely on only a single method (e.g., a single item or scale) or a single source (e.g., self-rating) are often less informative than are studies that combine multiple sources of information. Second, the generalizability of multimethod investigations tends to be broader than that of single-method studies (Morris, Robinson, & Eisenberg, 2006). Results based on a single method may be specific to that particular measurement instrument (e.g., test, rater, or observation) and thus limited in scope. Third, in contrast to monomethod studies, MTMM-MO designs make it possible to update the convergent and discriminant validity information about the measures each time measurement takes place. This is important given that the

Christian Geiser and Michael Eid, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany; Fridtjof W. Nussbeck, Department of Psychology, University of Zürich, Zürich, Germany; Delphine S. Courvoisier, Faculty of Medicine, University of Geneva, Geneva, Switzerland; David A. Cole, Department of Psychology, Vanderbilt University.

Correspondence concerning this article should be addressed to Christian Geiser, who is now at the Department of Psychology, Arizona State University, 950 South McAllister, Tempe, AZ 85287-1104. E-mail: geiserch@gmail.com

degree of validity of a particular method (e.g., rater, questionnaire, test, observation) might change in the course of a longitudinal study.

Changes in the validity and other psychometric properties of the measures over time are a serious problem in many longitudinal studies. Imagine, for example, a study that investigates the development of aggressive behavior from childhood to adolescence using self-report questionnaires. The question arises, are the same questionnaires appropriate for both children and adolescents? By using MTMM-MO designs, researchers can test whether the convergent and discriminant validity of their measures remains constant throughout their study.

Finally (and perhaps most importantly), changes in the underlying construct might not be correctly identified by a single method. Furthermore, different methods might disagree as to the kind or amount of change that occurs. Collectively, these constitute problems pertaining to the "convergent validity of change." For example, parents might underestimate the amount of change in depression of their children compared with the children's self-reports. Compared with mono-method designs, MTMM-MO designs provide more information about such changes in the underlying construct (see Hertzog & Nesselroade, 2003, for an overview about measurement of change).

MTMM-MO data sets tend to be large and complex. The analysis of such data requires careful consideration. Some investigators simply examine correlations among manifest variables. Others conduct separate MTMM analyses for each wave. A drawback of such approaches is that no overall theoretical model for the entire data structure is ever tested. Analyses that are solely based on manifest correlations have the additional disadvantage that measurement error is not taken into account. Consequently, estimates of the relations among the underlying constructs may be biased (e.g., Bollen, 1989). One might, for instance, underestimate the stability of interindividual differences in anxiety or depression over time. One might also obtain biased estimates of cross-sectional associations among constructs, potentially affecting estimates of the convergent and discriminant validity at a given point in time.

Measurement error also affects the assessment of change. Change scores based on observed variables may be especially affected by measurement error, in part, because the unreliability of both the pretest and the posttest score affects an observed difference score. Clearly, we need adequate methods for taking measurement error into account as we model change in MTMM-MO designs.

In this article, we show how these problems can be handled with an SEM approach. Using SEM, an overall theoretical model can be formulated and tested against the data. Furthermore, change can be analyzed in terms of latent variables (McArdle, 1988; McArdle & Hamagami, 2001; Steyer et al., 1997). We show how latent change can be analyzed simultaneously for different methods. In particular, we present a confirmatory factor analysis (CFA) approach to analyzing change in MTMM-MO studies. This approach is an extension of McArdle's (1988) latent difference (LD) model (see also McArdle & Nesselroade, 1994; Steyer et al., 1997; Steyer, Partchev, & Shanahan, 2000) to accommodate multiple constructs and multiple methods.

Our article is organized as follows. We first provide a brief review of existing SEM approaches for modeling MTMM-MO

data and motivate the need for a multimethod change model. Second, the conventional (single-method) CFA model of change (McArdle, 1988; McArdle & Hamagami, 2001; Steyer et al., 1997, 2000) is introduced. Third, we illustrate the usefulness of extending this model to multiple methods. Fourth, an application of the multimethod change model to the development of childhood depression and anxiety is presented as an illustrative example. Finally, we discuss the advantages and limitations of this approach and show links to other recently developed methods for analyzing MTMM-MO data.

SEM Approaches to Modeling MTMM-MO Data

To date, relatively few comprehensive statistical models have been developed for the analysis of MTMM-MO data in developmental psychology. Exceptions include Cudeck's (1988) multiplicative approach; Cole and Maxwell's (2003) multi-occasion correlated uniqueness (CU) approach; Burns, Walsh, and Gomez's (2003) extension of the correlated trait-correlated method (CT-CM) model (Marsh, 1989; Marsh & Bailey, 1991; Marsh & Grayson, 1995) to a multioccasion CT-CM model (see also Burns & Haynes, 2006); and the development of multimethod latent state trait (LST) models (Courvoisier, Nussbeck, Eid, Geiser, & Cole, 2008; see also Vautier, 2004). The primary focus of these models is on analyzing latent states (LSs). These models do not focus on change; they do not even include latent change variables. Nevertheless, change is often the primary focus in developmental psychology. In the present article, we present a model for analyzing change.

Vautier, Steyer, and Boomsma (2008) have developed a latent change model that takes method effects (because of indicator-specific [IS] effects) into account. In contrast to the approach presented here, the model discussed by Vautier et al. is not specifically designed for MTMM-MO investigations, does not provide coefficients for quantifying the convergent and discriminant validity of change, and does not allow for a comparison of change scores across different methods. Furthermore, Vautier et al.'s model assumes that method effects are perfectly stable over time, whereas the model presented here provides for all of these possibilities.

The Single-Method LS and LD Models

Let us assume that a researcher is interested in the development of child depression and wants to find out whether (a) children differ in their level of depression at each point in time, (b) the mean level of depression changes over time, (c) some children show a stronger increase or decrease of their depression scores over time than do other children (i.e., whether there is interindividual variability in the amount of change), and (d) other variables (such as, e.g., gender, parenting, social support, change of personal competences) can predict why some children's depression scores change more than others'. These important research questions can be answered by LD models. McArdle (1988) and Steyer et al. (1997, 2000) have shown how the basic longitudinal CFA measurement model can be reformulated as a LD model to study interindividual differences in intraindividual change with respect to a single construct. In contrast to these authors, we assume the more general case in which one has to deal with multiple con-

structs throughout this article. Below, we discuss the extension to multiple methods.

In line with Steyer and colleagues (Steyer et al., 1997, 2000), we present models for multiple indicators per construct-method-occasion (CMO) combination. Multiple indicator models have several advantages compared with single indicator models. For example, multiple indicator models allow researchers to investigate the degree of homogeneity versus indicator-specificity of different indicators of a construct. Most important, these models make it possible to test the assumption of measurement invariance over time. Both advantages are illustrated in detail below.

We first introduce the basic longitudinal CFA measurement model that serves as the base for analyzing latent change. To be in line with Steyer, Ferring, and Schmitt's (1992) classical article on LS and LST theory, we refer to this model as the *LS model*. In contrast to the LD model, the LS does not contain LDs as latent factors but just a time-specific latent factor for each construct at each time point. As these time-specific factors capture the true state of an individual with respect to a particular construct at a particular time point, the variables are called LS factors and hence the name LS model (see Steyer et al., 1992).

The LS Model

Figure 1A shows an LS model for three manifest variables or indicators Y_{ict} (i = indicator; i = 1, 2, 3; e.g., items or test scores) measuring two constructs (c = 1, 2; e.g., depression and anxiety) on two time points (t = 1, 2). In the LS model, there is a time-specific latent factor S_{ct} (c: construct; t: occasion of measurement) for each construct at each measurement occasion. Each LS factor S_{ct} is measured by the same indicators Y_{ict} at each measurement occasion. For example, depression might always be measured by the items of the Child Depression Inventory (CDI; Kovacs, 1985). Each factor S_{ct} represents the latent scores of individuals on a particular construct at a given point in time (e.g., the depression scores on a specific measurement occasion corrected for measurement error). Hence, these factors characterize the LS of individuals (i.e., the degree of depression) at a given time point and are therefore called LS factors.[1] (The meanings of the IS_{ic} latent variables are explained in detail below.)

Note that the covariance structure and the mean structure of the variables are part of the model. That is, means are also analyzed, especially because changes in latent factor means are also of interest in most longitudinal studies. (In contrast, most cross-sectional structural equation analyses focus only on the covariance structure.) Further, note that for reasons of clarity, we do not present the mean structure in the figures to avoid cluttering and to reduce the complexity of the figures; however, we emphasize that the mean structure is included in all models presented in this article. Figure 1A implies the following general measurement model for the indicators Y_{ict}:

$$Y_{ict} = \alpha_{ict} + \lambda_{ict}S_{ct} + \varepsilon_{ict}. \quad (1)$$

Equation 1 shows us that each indicator Y_{ict} has a factor loading λ_{ict} on the corresponding factor. The factor loading can be interpreted as a *discrimination parameter*. For example, the higher the loading of a depression scale, the greater the scale discriminates between individuals with different latent depression scores. It is

common practice to fix the loading of one indicator per factor to 1.00 to identify the scale of each LS factor.[2]

In addition, each indicator has an additive constant (intercept) α_{ict}. The intercept α_{ict} can be interpreted as a *difficulty parameter*—the greater α_{ict}, the easier the scale. For example, the statement "sometimes, I feel sad" is probably easier to endorse than the statement "I feel sad all the time." To identify the means of the LS factors S_{ct}, it is convenient to fix the intercept of those indicators to 0.00 whose state factor loadings are fixed to 1.00. Again, this choice is arbitrary; other possibilities exist to identify the latent means (see, e.g., Steyer et al., 2000).

In the general LS model, the loading and intercept parameters both carry three indices, i, c, and t. This implies that the loadings and intercepts can be different for each indicator and can also be different for the same indicator at different time points. However, as we see below, when LD scores are considered, one has to impose *measurement invariance constraints* on these parameters to make the LD scores meaningful. More specifically, one usually holds these parameters equal for the same indicator over time.

The variable ε_{ict} is a residual variable with a mean of zero. The residual variable captures measurement error and variable-specific effects (indicator-specificity). Indicator-specificity refers to reliable effects that are unique for an indicator and not shared with other indicators. These effects might be due to a specific response format, item wording, or content. For example, one depression scale might focus more strongly on sadness, whereas another scale might capture aspects of depression that are more related to lethargy. The variance of the residual variable, $Var(\varepsilon_{ict})$, tells us something about the degree of unreliability of the indicators and about the degree of indicator-specificity: the greater $Var(\varepsilon_{ict})$, the more unreliable and/or specific variance is present in a scale. (As we see below, with longitudinal data, we are able to separate the IS variance from random error variance; this is not possible with cross-sectional data, in which measurement error and indicator-specificity are confounded.)

The model in Figure 1A assumes that the common state factors S_{ct} are uncorrelated with all residual variables $\varepsilon_{ic't'}$, and that all residual variables ε_{ict} are uncorrelated with each other.

Measurement invariance. An important assumption that should generally be tested in longitudinal investigations is the assumption of measurement invariance over time (Meredith & Horn, 2001; Tisak & Tisak, 2000; Widaman & Reise, 1997). In the longitudinal context, measurement invariance means that the rela-

---

[1] Note that in LST theory (Steyer et al., 1992), the time-specific LS factors S_{ct} are further decomposed into a latent trait (stable component) and a latent occasion-specific residual factor (situation-specific component). However, in the present case, our goal is not to separate trait and state components but to model change over time as a latent variable. Note also that in the present context, a *LS variable* does not refer to a *categorical* latent variable (i.e., latent class variable) but to a *continuous* latent variable (i.e., a factor).

[2] Identification generally refers to the question of whether enough information is available so that a unique solution can be determined for each unknown model parameter, for example, factor loadings, factor variances, and so forth. The present restrictions for identifying the scale of the state factor are arbitrary; the metric of this factor could also be identified by other restrictions, for example, by fixing its variance to 1.00 (see Bollen, 1989, for a detailed discussion of this issue).
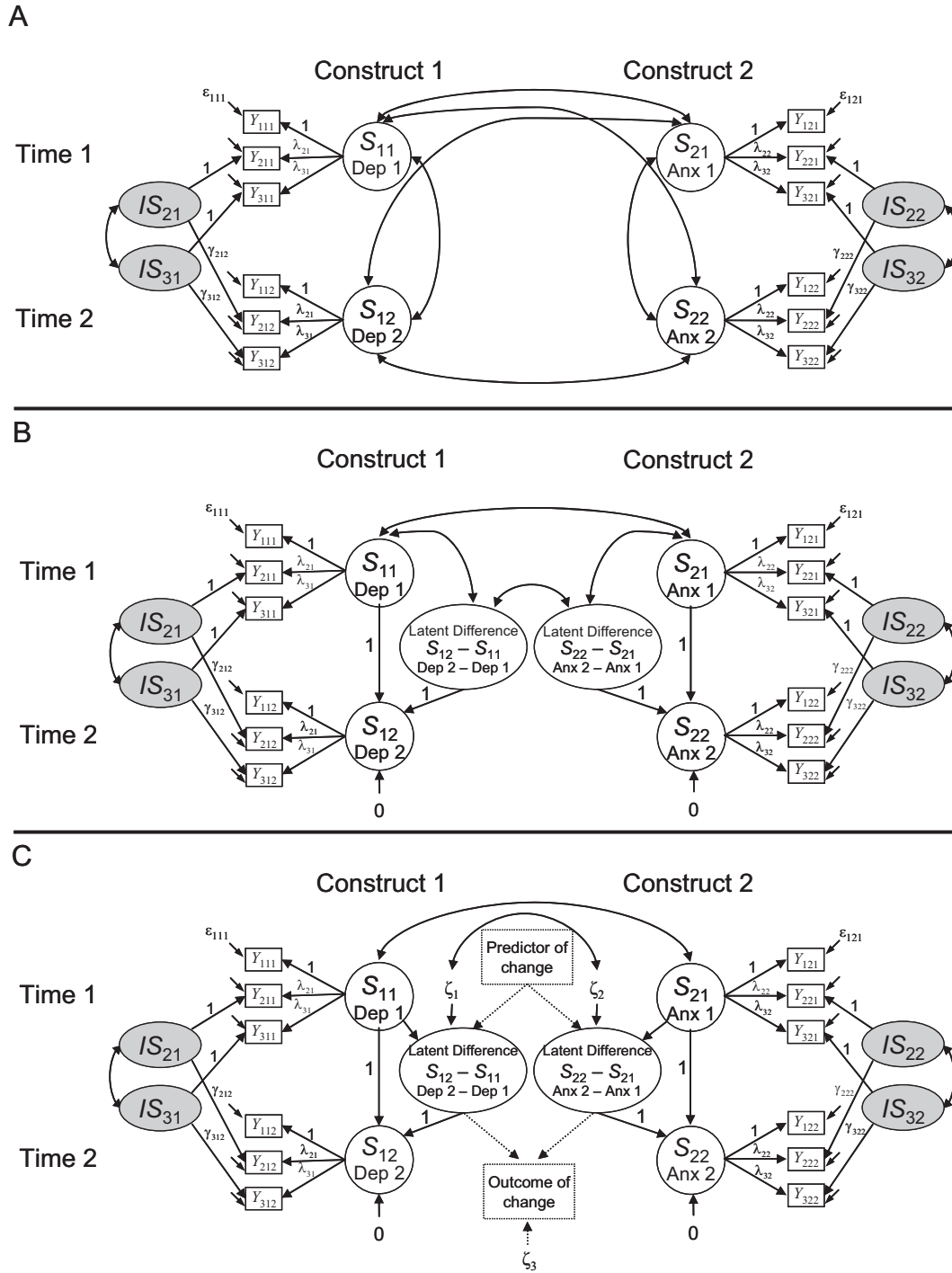
Figure 1. Longitudinal confirmatory factor analysis models for two constructs (Dep = Depression; Anx = Anxiety) measured by one single method on two time points. Y_{ict} denotes the ith observed variable measuring construct c on occasion t; λ_{icm} = time-invariant state factor loading; S_{ct} = latent state factor; γ_{icm} = indicator-specific factor loading; IS_{ic} = indicator-specific factor; ε_{ict} = residual (error) variable (not shown for all indicators). (A) Latent state model with correlated state factors. (B) Reformulation of the latent state model as latent difference model. (C) Extended latent difference model including a predictor variable and an outcome variable of change. ζ_1–ζ_3 denote latent residual variables. The variances of the latent variables S_{c1}, $(S_{c2} - S_{c1})$, IS_{ic}, and ε_{ict} are estimated in the model. Intercepts (α_{ict}) and latent means are also included in the model but are not shown in the figure for simplification. The models are identified by fixing the intercept of the first indicator to zero and its loading to one for each factor.

tion between the observed variables and their underlying latent variables do not change over time. In other words, the observed (item, scale, test, etc.) scores should have the same meaning at each point in time. Measurement invariance is particularly important in the LD model, as this model is concerned with differences between latent variables. To make sure that these differences are meaningful (that one does not subtract "apples from oranges"), measurement invariance has to be established.

How can measurement invariance be represented in the parameters of the LS model? An important aspect of measurement invariance is that the links between the indicators and the latent factors should remain the same over time. In the LS model, these links are represented by the factor loadings λ_{ict} and intercepts α_{ict}. If both the factor loadings λ_{ict} and the intercepts α_{ict} are time-invariant (i.e., if, for the same indicator, these parameters take on the same values on each measurement occasion), the condition of *strong measurement invariance* is satisfied (Widaman & Reise, 1997).[3]

Time-invariant intercepts and loadings imply that the measurement structure of the construct has not changed over time. In contrast, variation in the intercepts over time would indicate changes in the difficulty of a scale, and variation in the factor loadings would indicate that the discrimination of a scale has changed over time. Under the condition of measurement noninvariance, one would be faced with changes in the metric and meaning of the underlying latent variables. This would lead to complications when researchers want to compare, for instance, latent means over time, given that in cases in which measurement invariance is severely compromised, apples would be compared with oranges.

Does measurement invariance always hold? No, but fortunately, we are able to test this assumption empirically in the LS model. This is possible because the LS model makes use of multiple indicators for each LS factor at each time point. For this purpose, we can specify a restricted version of the LS model, in which all loadings and intercepts are constrained to be equal over time. When such invariance constraints are imposed on all intercepts and loadings, the measurement model in Equation 1 simplifies to

$$Y_{ict} = \alpha_{ic} + \lambda_{ic}S_{ct} + \varepsilon_{ict}. \tag{2}$$

We can see that the index t for the measurement occasion has now been dropped from the intercepts and loadings. This indicates that these parameters can no longer vary over time. In practical applications, the fit of this restricted model can be statistically tested against the fit of the more general model in Equation 1 to investigate whether the assumption of measurement invariance is tenable. When the more constrained model with time-invariant parameters fits worse than the model with freely estimated parameters, measurement invariance is not tenable for at least some indicators.

If invariance of intercepts and loadings is only tenable for some, but not all indicators, one speaks of *partial measurement invariance* (Byrne, Shavelson, & Muthén, 1989). Partial invariance can be sufficient to warrant proper interpretation of differences in the structural model (e.g., differences in latent means over time). According to Byrne et al. (1989, p. 458), a sufficient degree of partial invariance is present if, in addition to the indicator whose loading and intercept are fixed to set the scale of the latent factor, at least one additional indicator has invariant parameters over time.

However, this is a rather arbitrary rule. In practice, researchers need to think carefully of the possible causes and consequences of violations of measurement invariance even if just one indicator is non-invariant.

What does the LS model tell us about development and change? If we go back to our depression and anxiety example, the variance of a factor S_{ct} tells us something about interindividual differences in the true (i.e., error free) scores at a given time point. For example, the question "how large are true interindividual differences in depression at the first measurement occasion?" can be answered by estimating the variance of S_{11}. Factor means for the same construct can be compared over time to discover whether the average level of depression and anxiety has changed over time.

The correlations among factors pertaining to the same construct over time [i.e., $Cor(S_{ct}, S_{ct'})$] inform us about the stability of interindividual differences over time. If these correlations equal 1.00, then interindividual differences on the same construct are perfectly predictable from the preceding measurement occasion (note that this would not preclude mean changes over time).

A limitation of the LS model is that it does not allow direct analysis of interindividual differences in change (in terms of LD scores). The following reformulation of the LS model to an LD model allows us to analyze latent change variables that can be used to relate interindividual differences in change to other variables (McArdle, 1988; Steyer et al., 1997, 2000).

The LD Model

The basic idea of LD modeling is that a latent factor S_{ct} can be decomposed into the initial status (Time 1 = T1) factor S_{c1} plus a LD factor $(S_{ct} - S_{c1})$, where $(S_{ct} - S_{c1})$ represents latent change (growth or decline) from T1 to time t:

$$S_{ct} = (S_{ct} + S_{c1} - S_{c1}) = 1S_{c1} + 1(S_{ct} - S_{c1}). \tag{3}$$

Equation 3 is a simple restatement and does not contain any restrictive assumptions. Basically, we just added a variable (S_{c1}) and subtracted the same variable at the same time. This might seem weird at first sight, but it allows us to consider true change as a latent variable in the SEM. By implementing Equation 3 in the structural model (as shown in Figure 1B), the LS model can be transformed into a LD model (the measurement model remains unchanged). The LD factors $(S_{ct} - S_{c1})$ represent true interindividual differences in intraindividual change from T1 to time t on construct c. The term *true* here means that the difference scores are corrected for measurement error. Therefore, these models have also been referred to as *true change* models (Steyer et al., 1997).

To illustrate, we can imagine that each individual has a score on the LD factor $(S_{ct} - S_{c1})$. A negative score indicates a true decline in depression for an individual, whereas a positive score indicates a true increase, and a score of zero indicates no true change at all. The LD factor mean and variance can also be estimated. The LD factor mean reflects the average growth or decline between the two time points for this sample. With regard to the depression example,

---

[3] Widaman and Reise (1997) distinguished between four different degrees of measurement invariance. However, for our purpose, it is sufficient to consider the case of *strong measurement invariance*, in which both the factor loadings and intercepts are time-invariant.

a positive mean would indicate that on average, individuals show an increase in depression over time. The LD factor variance reflects the degree of true variability (interindividual differences) in change. In our example, a high variance would indicate that individuals differ a lot in the degree to which their true depression level changes over time.

LD modeling offers a direct approach to investigating change. The LD model is statistically equivalent to the LS model, as it is just a reformulation. That is, one does not specify a new model but just makes the information about change more accessible (Steyer et al., 1997, 2000). Furthermore, LD models are quite flexible. LD variables can be included for investigating change between a given occasion of measurement and the first time point or for change between any specific pair of state factors that is of interest (e.g., to study change between adjacent time points; see Steyer et al., 1997, 2000). As illustrated in Figure 1C, LD models can be used as endogenous variables (outcomes) or predictor variables in slightly expanded models. For example, one might be interested in finding predictor variables that explain why depression scores increase more for some people than others or why intervention programs cause greater changes in cognitive abilities in some children than in others. Alternatively, change scores may themselves be used as predictors of other variables (e.g., change in cognitive abilities may cause or at least predict change in school grades). Given that LD models focus on interindividual differences in intraindividual change, they can also be very useful for analyzing treatment effects in intervention studies.

IS Effects

As we noted before, the residual variables ε_{ict} in an LS or LD model capture both indicator-specificity and random measurement error. Consequently, in longitudinal studies in which the same indicators are repeatedly measured, these residuals are often correlated over time (Sörbom, 1975). Such auto-correlated errors indicate that the error variables do not only contain random measurement error but also reliable IS effects that generalize across time (Jöreskog, 1979; Marsh & Grayson, 1994; Raffalovich & Bohrnstedt, 1987; Tisak & Tisak, 2000). Even though such effects are not usually of substantive interest, failure to account for them in the model can lead to specification error, misfit, and biased parameter estimates (e.g., the overestimation of construct stability).

One possible way to model IS effects in the LS and LD model is to allow for correlations between residual variables pertaining to the same indicator (so-called CU approach; Cole & Maxwell, 2003; Kenny, 1976; Lance, Noble, & Scullen, 2002; see discussion above). High correlations between residual variables indicate a high degree of indicator-specificity. The CU approach is a straightforward method to represent indicator-specificity in longitudinal studies. However, it is limited in that it does not explain systematic sources of variance by additional latent factors. As a consequence, IS variance and random error variance remain confounded in this approach. The inability to separate these variance components leads to an underestimation of the reliable variance of an indicator. Furthermore, when the number of indicators is large, the CU approach lacks parsimony, as many residual correlations need to be estimated.

Therefore, we recommend an alternative approach in which indicator-specificity is represented by IS factors (Eid, Schneider, & Schwenkmezger, 1999; Raffalovich & Bohrnstedt, 1987; Tisak & Tisak, 2000) rather than correlated error variables. Figures 1A and 1B show the LS and LD models with additional IS factors. According to Eid et al. (1999), it is not necessary to include an IS factor for each repeatedly measured indicator. These authors have shown that it is sufficient to use $i - 1$ indicator specific factors per construct (i.e., one IS factor less than there are indicators per factor). Specifying all i IS factors per construct can cause underidentification and other estimation problems. The extended LS model with $i - 1$ IS factors is given by the following:

$$Y_{ict} = \begin{cases} \alpha_{1ct} + \lambda_{1ct}S_{ct} + \varepsilon_{1ct}, & \text{for } i = 1 \\ \alpha_{ict} + \lambda_{ict}S_{ct} + \gamma_{ict}IS_{ic} + \varepsilon_{ict}, & \text{for } i \neq 1 \end{cases}, \qquad (4)$$

where the first indicators ($i = 1$) have been selected as "reference indicators" for which no IS factors are specified. The constant α_{ict} again denotes the intercept, λ_{ict} denotes the factor loading on the state factor S_{ct}, γ_{ict} is the factor loading on the IS factor IS_{ic}, and ε_{ict} is a residual variable with a mean of zero. It is assumed that IS_{ic} also has a mean of zero and is uncorrelated with all state factors S_{ct} that belong to the same construct.[4] Furthermore, it is assumed that the residual variables ε_{ict} are uncorrelated with all other latent variables and with all other residual variables. Note that in contrast to S_{ct}, the factor IS_{ic} is time-unspecific and therefore carries no occasion index t, because IS_{ic} represents IS effects that generalize across time.

We recommend that IS factors be included in longitudinal models whenever indicators show IS effects over time. If such effects are not present (i.e., because of a high degree of homogeneity of the indicators or long time intervals between the measurements), these factors can be dropped, leading to the simpler model in Equation 1. In the final section of this article, we discuss some specific problems that may occur with IS factors in practical applications and appropriate solutions for these problems.

In the extended LS model with IS factors, measurement invariance can also be tested. For this purpose, one simply constrains the intercepts and factor loadings to be time-invariant and then compares the fit of such a restricted model with the fit of an unconstrained model in which the intercepts and loadings are free to take on any value.

The Multimethod Change Model

The LS/LD approach can easily be generalized for use in an MTMM-MO design. For example, imagine that a researcher wants to assess change in depression and anxiety as measured by two

---

[4] According to their psychometric definition (see Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003), indicator-specific factors may be correlated with all other indicator-specific factors and with all state factors that belong to different constructs. Correlations between indicator-specific factors are often meaningful and interpretable (e.g., if two items share common method variance because of a similar wording, one would expect their indicator-specific factors to be correlated). In contrast, correlations between indicator-specific factors and other state factors are often difficult to interpret and also close to zero in many cases. In cases in which it is not plausible to assume these correlations to be different from zero, one may consider fixing them to zero a priori for reasons of parsimony.

different methods (e.g., self- and parent report). As shown in Figure 2, the only difference is that the variables and constants in the new model receive an additional index m to represent the method. For example, the indicators Y_{icmt} now carry four, and the state factors S_{cmt} carry three indices (i = indicator, c = construct, m = method, t = time point). An LD model can then be specified for each construct-method unit. To our knowledge, an application of such an extended LD model to MTMM-MO data has not yet been presented in psychology.

The multimethod change model is especially relevant to researchers analyzing MTMM-MO data. According to Campbell and Fiske (1959), different methods should converge in measuring the same attribute (*convergent validity*). In addition, a new measure or construct should be distinguishable from an already established measure or construct (*discriminant validity*). For example, a new test should not correlate too highly with an existing test supposed

to measure a different construct. If we transfer Campbell and Fiske's guidelines to the measurement of change, we can assess the *convergent validity of change* by correlating the LD variables pertaining to the same construct but different methods [e.g., $(S_{112} - S_{111})$ and $(S_{122} - S_{121})$ in Figure 2]. High correlations indicate that different methods are in agreement as to the assessment of interindividual differences in intraindividual change. Individuals with high change scores according to one method would also have high change scores according to the other method.

In addition, the means of the LD factors can be compared across different methods measuring the same construct. Convergent validity with respect to *mean change* suggests that different methods agree with regard to the direction of mean change over time: mean increase, mean decrease, or mean stability. For example, do self- and parent reports converge with regard to the assessment of mean change in depression and anxiety over time? Note that mean
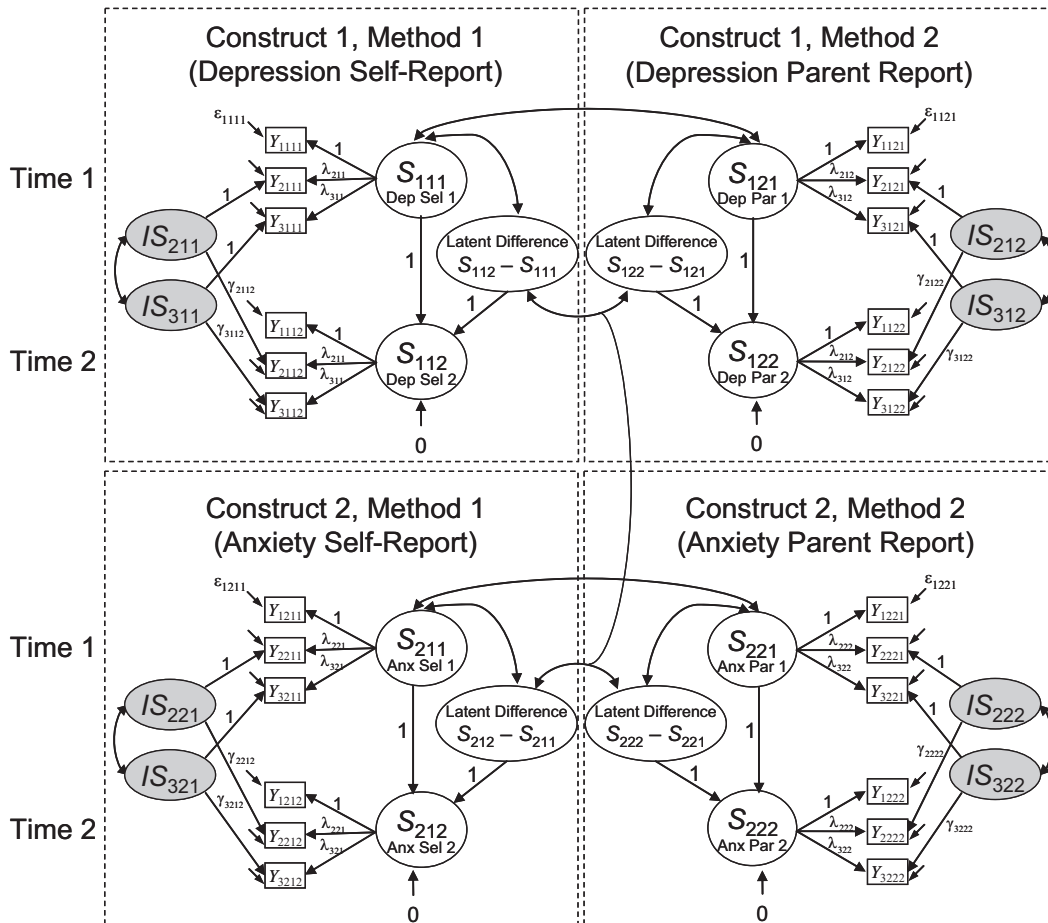

Figure 2. Multimethod change model for two constructs (Dep = Depression; Anx = Anxiety) measured by two methods (Sel = Self-Report; Par = Parent Report) on two time points. Each square highlights a construct method unit. Y_{icmt} denotes the ith observed variable measuring construct c by method m on occasion t. Latent difference variables are included for each construct-method unit. λ_{icm} = time invariant state factor loading; γ_{icmt} = indicator-specific factor loading. The variances of the latent variables—that is, $Var(S_{cm1})$, $Var(S_{cm2} - S_{cm1})$, $Var(IS_{icm})$, and $Var(\varepsilon_{icmt})$—are estimated in the model. Intercepts (α_{icm}) and latent means are also included in the model but are not shown in the figure for simplification. The models are identified by fixing the intercept of the first indicator to zero and its loading to one for each factor. To simplify, not all possible factor correlations are shown (but see the discussion in the text).

comparisons across methods are only meaningful if the different methods (e.g., raters) used equivalent scales (e.g., a self- and a parent rating form of the same questionnaire, where both forms contain equivalent items).

We can also apply Campbell and Fiske's (1959) concept of discriminant validity to variables that measure change. Correlations among change variables that belong to the same method, but different constructs [e.g., $(S_{112} - S_{111})$ and $(S_{212} - S_{211})$ in Figure 2] can be interpreted in terms of *discriminant validity of change*. High correlations might indicate that a method is not sufficiently sensitive to differences in change processes for the two different constructs. For example, parents rating the depression and anxiety of their children might be unable to distinguish change in depression from change in anxiety. More broadly, one can ascertain whether the pattern of correlations among change scores pertaining to different constructs is consistent from one method to another. An application of Campbell and Fiske's criteria to the measurement of change is shown in the illustrative example discussed below.

The present model can also be used to determine the reliability of observed change scores ($Y_{icmt} - Y_{icmt'}$). If we assume time-invariant intercepts α_{icm}, time-invariant state loadings λ_{icm}, and time-invariant IS loadings γ_{icm}, we can express the observed difference scores as follows:

$$(Y_{icmt} - Y_{icmt'}) = \lambda_{icm}(S_{cmt} - S_{cmt'}) + \varepsilon_{icmt} - \varepsilon_{icmt'}. \quad (5)$$

Note that there is no intercept in Equation 5 because we assume $\alpha_{icmt} = \alpha_{icmt'} = \alpha_{icm}$, so that the intercepts cancel out. Given the assumption of equal loadings γ_{icm} over time, the IS factors also cancel out—IS effects do not affect the difference score under this condition. Assuming uncorrelated error variables, the variance of the difference score, $Var(Y_{icmt} - Y_{icmt'})$, can be decomposed as follows:

$$Var(Y_{icmt} - Y_{icmt'}) = \lambda_{icm}^2 Var(S_{cmt} - S_{cmt'}) + Var(\varepsilon_{icmt})$$
$$+ Var(\varepsilon_{icmt'}). \quad (6)$$

Notice that Equation 6 shows us why observed difference scores often are so unreliable: The error variances of both time t and time t' affect the variance of the difference score. We may calculate the reliability for ($Y_{icmt} - Y_{icmt'}$) as follows:

$$Rel(Y_{icmt} - Y_{icmt'}) = \frac{\lambda_{icm}^2 Var(S_{cmt} - S_{cmt'})}{Var(Y_{icmt} - Y_{icmt'})}, \quad (7)$$

where λ_{icm}^2 and $Var(S_{cmt} - S_{cmt'})$ are parameters to be estimated on the basis of the model, and $Var(Y_{icmt} - Y_{icmt'})$ is the model-implied difference score variance.

Identification and Estimation of the Models

Identification

In SEM with latent variables, each latent factor has to be assigned a scale (Bollen, 1989). For the present models, a convenient way to identify the scale and mean of each LS factor is to fix the loading of one indicator to 1.00 and the intercept of the same indicator to 0.00. Furthermore, if IS factors are included, one can identify the scale of these factors by also fixing the loading of one

indicator to 1.00. As noted above, these restrictions are arbitrary. Other, equivalent possibilities to identify the model exist (see, e.g., Steyer et al., 2000).

It is important to note that identification problems may arise in measurement designs with only two time points. In this case, the model part related to the IS factors may not be identified. The following guidelines can be provided for these cases:

1. If there is only one construct measured by only two indicators pertaining to only one method on only two time points, the IS factor is underidentified when the second loading is left to be freely estimated, unless the IS factor is substantially correlated with an external covariate in the model. If there is no external covariate or if the IS factor is not substantially correlated with the covariate, one has to fix both IS factor loadings (e.g., to 1.00) to achieve identification.

2. If there are just two time points, just one construct, and just one method, but three or more indicators per state factor, an identification problem only occurs if one or more of the IS factors are not substantially correlated with any other IS factor or with any external covariate in the model. If one or more of the IS factors are not substantially correlated with any other variable in the model, both loadings of this IS factor have to be fixed (e.g., to 1.00) to achieve identification.

3. If there are just two time points and just two indicators per state factor, but two or more methods or constructs, an identification problem only occurs if one or more of the IS factors are not substantially correlated with any other factor or with any external covariate in the model. If one or more of the IS factors are not substantially correlated with any other variable in the model, both loadings of this IS factor have to be fixed (e.g., to 1.00) to achieve identification.

Estimation

All models presented in this article are special models of CFA. The parameters of the models can therefore be estimated using estimation methods (e.g., maximum likelihood [ML]) available in conventional SEM software (e.g., AMOS, EQS, LISREL, Mplus, Mx). The models should be fit to raw data or to the covariance matrix and mean vector of the observed variables. ML estimation is based on the assumption of multi-variate normality. If this assumption is violated, standard errors and fit statistics may be biased. The use of robust ML estimation (e.g., Satorra & Bentler, 1994) or Bootstrap methods (Bollen & Stine, 1992) is recommended in this case (e.g., Finney & DiStefano, 2006).

Annotated Mplus (L. K. Muthén & Muthén, 1998–2007) input files to estimate the models presented in this article are available from the online supplementary materials. The following is an illustrative application of these methods to an actual MTMM-MO data set.

Method

Sample

The three-wave data set used for the present illustration was drawn from a larger data set collected by Cole and colleagues (Cole et al., 1998; Cole, Martin, & Powers, 1997; Cole, Martin, Powers, & Truglio, 1996; Cole, Truglio, & Peeke, 1997). Participants were U.S. children ($N = 906$) who were in the third and sixth grades at Time 1. Demographic characteristics were not available for data analysis but are described at length in previous publications (Cole et al., 1998, Cole, Martin, & Powers, 1997; Cole et al., 1996; Cole, Truglio, & Peeke, 1997).

For simplicity, we analyzed the combined data of third and sixth graders for the present illustration. We emphasize, however, that in an actual substantive application, a *multigroup approach* (Jöreskog, 1971) should be preferred, in which the age groups are treated separately. The time interval between each of the three measurement occasions was 6 months. For the present illustration, we selected two constructs (depression and anxiety) and two methods (self-report and parent report).

Measures

Depression was measured by the child and parent form of the CDI (and CDI–Parent Form; 26 items measured on a 3-point scale ranging from 0 to 2; Kovacs, 1985). Anxiety was assessed by the child and parent form of the Revised Children's Manifest Anxiety Scale (RCMAS–Child Form and RCMAS–Parent Form; 28 items measured on a 3-point scale ranging from 0 to 2; Reynolds & Richmond, 1978). The number and the wording of the items was equivalent across the child and parent forms for both the CDI and the RCMAS. A more detailed description of the scales can be found in Cole, Martin, and Powers (1997), as well as Cole, Truglio, and Peeke (1997).

Details on the Statistical Analysis

Model structure. The LS/LD models estimated for the present data had the same basic structure as the model shown in Figure 2, with the exception that we considered three (instead of two) occasions of measurement and used two (instead of three) indicators per state factor. Hence, there were a total of 16 latent factors in the LS version of the model: 12 state factors (three state factors for self-reported depression, three state factors for self-reported anxiety, three state factors for parent-reported depression, and three state factors for parent-reported anxiety) and four IS factors (one for the second indicator of self-reported depression, one for the second indicator of self-reported anxiety, one for the second indicator of parent-reported depression, and one for the second indicator of parent-reported anxiety). The LD version additionally contained eight LD factors (one difference factor for self-reported depression T2 – T1, one difference factor for self-reported depression T3 – T1, one difference factor for self-reported anxiety T2 – T1, one difference factor for self-reported anxiety T3 – T1, one difference factor for parent-reported depression T2 – T1, one difference factor for parent-reported depression T3 – T1, one difference factor for parent-reported anxiety T2 – T1, and one difference factor for parent-reported anxiety T3 – T1).

We estimated two versions of the LS/LD model (referred to as Model 1 and Model 2 in the following) to test whether measurement invariance across time was tenable. In Model 1, we imposed no measurement invariance constraints over time (i.e., all loadings and intercepts were free to take on any value except those fixed to ensure identification). In Model 2, all factor loadings and intercepts were held equal over time (we also constrained the IS factor loadings to be equal over time). A statistical comparison of the fit of Model 1 and Model 2 allowed us to test the assumption of strong factorial invariance (i.e., time-invariant intercepts and loadings). If Model 2 fit worse than Model 1, the assumption of strong factorial invariance for all indicators would need to be rejected.

Factor indicators. In principle, we could have used the single items of the CDI and RCMAS as manifest indicators in the models; however, two considerations led us to use parcels instead. First, the item responses were measured on a 3-point scale and thus represented discrete (ordered categorical = ordinal) rather than continuous variables. Ordered categorical outcomes require the use of specific models and estimators (e.g., Finney & DiStefano, 2006). The treatment of these issues is beyond the scope of this article.

Second, the large number of items per construct would have led to extremely large models with an excessive number of estimated parameters. Therefore, we created two parcels per CMO unit (CMOU; i.e., there were two parcels for depression measured by child report on each occasion of measurement, two parcels for depression measured by parent report on each occasion, and so on). We created the parcels by calculating the mean of half of the items for each CMOU. Parcels consisted of identical items on each occasion of measurement. Each depression parcel thus consisted of 13 items, and each anxiety parcel consisted of 14 items. We wish to emphasize that in an actual substantive application, researchers should conduct detailed preliminary item analyses to make sure that the use of item parcels instead of single items is warranted (T. D. Little, Cunningham, Shahar, & Widaman, 2002).

Clustering of observations. A complication that often arises in psychological studies is that observations are not independent but clustered within a hierarchical structure. For example, in the present data set, the children were nested within 49 school classrooms. Such complex sample structure is often referred to as multilevel structure and requires a special treatment. When the nonindependence of observations is ignored, standard errors and test statistics in conventional covariance structure analyses can be biased (Julian, 2001). In the present analysis, we handled this problem by using robust ML estimation in which a so-called Sandwich estimator is used to obtain adjusted standard errors and test statistics that are robust to clustering and nonnormality of the data (B. O. Muthén & Muthén, 1998–2004; B. O. Muthén & Satorra, 1995; these robust statistics were obtained by using the TYPE = COMPLEX option in Mplus).

Handling of missing data. We used full information ML (FIML) estimation to handle missing data (Arbuckle, 1996; R. J. A. Little & Rubin, 2003; Wothke, 2000). The FIML method relies on the assumption that missing values are missing at random. Missing at random requires that missing data are either missing completely at random or that the missing information depends "on variables that are included in the statistical model (which in the longitudinal setting usually includes previous measurements of the same variable)" (Ghisletta & Lindenberger, 2005, p. 566). The FIML method is generally preferred to listwise deletion and other ad hoc approaches to handling missing data (Schafer & Graham, 2002; Wothke, 2000).

Goodness-of-fit assessment.   Goodness of fit was assessed using the chi-square test, the root-mean-square error of approximation (RMSEA; Steiger, 1990), and the comparative fit index (CFI; Bentler, 1990). A nonsignificant chi-square value indicates that the assumption of exact fit in the population is not rejected, suggesting a good fit of the model to the data. The RMSEA coefficient is a measure of approximate fit. RMSEA values smaller than .05 point to an acceptable fit (Hu & Bentler, 1998, 1999). The CFI compares the fit of the target model with the fit of a baseline model. The baseline model is a null model that assumes zero covariation among the observed variables. For a good model, the CFI should be greater than 0.95 (Hu & Bentler, 1998, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). To test for measurement invariance, we performed chi-square difference tests to statistically compare nested models.[5] As an additional descriptive index for model comparisons, we use the Bayesian information criterion (BIC; Schwarz, 1978). According to this criterion, the model with the smallest BIC value fits the data best.

IS effects.   Preliminary analyses showed that IS effects were present for all constructs and raters. (Models without IS factors did not fit the data well.) Therefore, as explained above, we focused on models that had the same structure as the one shown in Figure 2.

Results

Goodness of Fit

We first tested measurement invariance over time. For this purpose, we compared the fit of Model 1 (without invariance restrictions on loadings and intercepts) against Model 2 (with loadings and intercepts held equal over time for all indicators). Both models fit the data well. Model 1 had a nonsignificant chi-square value: $\chi^2(132) = 142.86$, $p = .24$. Approximate fit indices were also good ($RMSEA = .010$; $CFI = .999$).

The chi-square value of the more constrained Model 2 was also nonsignificant, $\chi^2(156) = 161.02$, $p = .38$, and the model showed a very good fit according to the other indices as well ($RMSEA = .006$; $CFI = 1.000$). A chi-square difference test revealed that Model 2 did not fit significantly worse than Model 1, $\Delta\chi^2(24) = 19.03$, $p = .75$. Model 2 also had a lower BIC value ($BIC_{\text{Model 2}} = -4,704$) than Model 1 ($BIC_{\text{Model 1}} = -4,566$). The results thus indicate that the assumption of measurement invariance was tenable. Hence, we present detailed outcomes for Model 2. We first discuss important findings from the measurement model.

Measurement Model: Indicator-Specificity and Reliability

The measurement model provides us with information on the degree of homogeneity and reliability of the indicators. Table 1 shows that the unstandardized intercepts, factor loadings, and residual variances are very similar for the indicators of the same LS factor (e.g., the intercepts, state factor loadings, and residual variances for DS11 and DS21 are very similar). This shows that the item parcels are essentially homogeneous.

Standardized factor loadings (see the fifth and eighth columns of Table 1) can be interpreted as correlations between the indicator and the corresponding LS and IS factors, respectively. Squared standardized loadings represent variance components (see the last three columns in Table 1) that give the percentage of variance explained by the

LS and IS factors, respectively. The sum of the squared standardized loadings on both types of factors equals the reliability of an indicator.

In the present example, all indicators show high standardized factor loadings on the state factors (range = .85–.96), whereas the standardized loadings on the IS factors are substantial but much smaller (range = .22–.37). This shows us that although IS effects are present and nonnegligible, they play a minor role in the present application.

Overall, the variance components indicate that 72%–92% of the observed variance is due to the corresponding state factor, whereas only 5%–14% of the observed variance is due to reliable IS effects. That is, the different indicators generally measure the same latent dimension. Further, the reliability coefficients are high for all indicators (range = .83–.92).

Structural Model: Convergent and Discriminant Validity of States and Change

The structural model provides us with information on convergent and discriminant validity and change. The latent MTMM-MO correlation matrices estimated in the state and change versions of the multimethod change model are shown in Tables 2 and 3, respectively. These matrices contain coefficients of convergent validity, discriminant validity, and stability among latent factors that are corrected for measurement error. In both matrices, convergent validity coefficients are printed in bold face, whereas discriminant validity coefficients are italicized. Stability coefficients are framed in boxes in Table 2. The MTMM-MO matrix obtained from the change version allows us to study the convergent and discriminant validity of change.

From the moderate to small bold-face correlations we can see that convergent validity is low for both the state scores (range = .26–.36; see Table 2) and the LD scores (range = .14–.28; see Table 3). The coefficients are higher for depression (range state factors = .32–.36; range difference scores = .22–.28) than for anxiety (range state factors = .26–.28; range difference scores = .14–.18). Although the differences are not large, we speculate that parents may find it easier to judge their children's depression than their anxiety. Further, the convergent validities are higher for the state than for the change scores.

The convergent validity of change is very low. That is, interindividual differences in intraindividual change estimated from the self-ratings do not correspond well with interindividual differences in intraindividual change estimated from the parent ratings. Consequently, an important substantive finding of the present modeling is that conclusions about change in depression and anxiety drawn from the parent ratings would be very different from conclusions drawn on the basis of the self-ratings. This finding could indicate three different things: Change scores based on the self-report may not be valid; change scores based on the parent report may not be valid; or both types of change scores may not be valid. One important consequence of these findings is that researchers interested in studying change in depression and anxiety of children would be ill-advised if they based their assessment on a single method only. Another important consequence is that in the next step, researchers should study the

---

[5] We calculated the correctly scaled chi-square difference value under robust ML estimation using the procedure described in Satorra and Bentler (2001) and on the Mplus homepage at http://www.statmodel.com/chidiff.shtml

Table 1

Estimated Intercepts, Factor Loadings, Residual Variances, and Variance Components in the Multimethod Change Model

Report indicator	Intercept α_{icm}	Time-invariant state factor loading λ_{icm}			Time-invariant indicator-specific factor loading γ_{icm}			Residual variance $Var(\varepsilon_{icmt})$	Variance components		
		Estimate	SE	Standardized estimate	Estimate	SE	Standardized estimate		State	Indicator-specific	Reliability
DS11	0.00*	1.00*	—	0.93	—	—	—	0.01	0.86	—	0.86
DS21	0.00[a]	0.92[e]	0.02	0.91	1.00*	—	0.22	0.01	0.83	0.05	0.88
DS12	0.00*	1.00*	—	0.92	—	—	—	0.02	0.85	—	0.85
DS22	0.00[a]	0.92[e]	0.02	0.90	1.00*	—	0.22	0.01	0.81	0.05	0.86
DS13	0.00*	1.00*	—	0.94	—	—	—	0.01	0.88	—	0.88
DS23	0.00[a]	0.92[e]	0.02	0.90	1.00*	—	0.23	0.01	0.81	0.05	0.86
AS11	0.00*	1.00*	—	0.94	—	—	—	0.02	0.88	—	0.88
AS21	0.06[b]	0.95[f]	0.02	0.91	1.00*	—	0.24	0.02	0.83	0.06	0.89
AS12	0.00*	1.00*	—	0.96	—	—	—	0.02	0.92	—	0.92
AS22	0.06[b]	0.95[f]	0.02	0.92	1.00*	—	0.23	0.02	0.85	0.05	0.90
AS13	0.00*	1.00*	—	0.96	—	—	—	0.02	0.92	—	0.92
AS23	0.06[b]	0.95[f]	0.02	0.91	1.00*	—	0.25	0.02	0.83	0.06	0.89
DP11	0.00*	1.00*	—	0.92	—	—	—	0.01	0.85	—	0.85
DP21	0.03[c]	1.01[g]	0.03	0.86	1.00*	—	0.32	0.01	0.74	0.10	0.84
DP12	0.00*	1.00*	—	0.91	—	—	—	0.01	0.83	—	0.83
DP22	0.03[c]	1.01[g]	0.03	0.85	1.00*	—	0.35	0.01	0.72	0.12	0.84
DP13	0.00*	1.00*	—	0.93	—	—	—	0.01	0.86	—	0.86
DP23	0.03[c]	1.01[g]	0.03	0.88	1.00*	—	0.32	0.01	0.77	0.10	0.87
AP11	0.00*	1.00*	—	0.94	—	—	—	0.01	0.88	—	0.88
AP21	−0.06[d]	1.07[h]	0.03	0.87	1.00*	—	0.35	0.02	0.76	0.12	0.88
AP12	0.00*	1.00*	—	0.95	—	—	—	0.01	0.90	—	0.90
AP22	−0.06[d]	1.07[h]	0.03	0.88	1.00*	—	0.37	0.01	0.77	0.14	0.91
AP13	0.00*	1.00*	—	0.94	—	—	—	0.01	0.88	—	0.88
AP23	−0.06[d]	1.07[h]	0.03	0.86	1.00*	—	0.36	0.02	0.74	0.13	0.87

Note. The first number refers to the indicator, whereas the second number refers to the occasion of measurement. For example, DS12 stands for the first indicator of depression self-report measured on the second occasion of measurement. Fixed parameters are marked with an asterisk. Standard errors are not available for fixed parameters. Parameters set equal to each other (time invariant parameters) have the same superscript. i = indicator; c = construct; m = method; t = occasion of measurement; DS = depression self-report indicator; AS = anxiety self-report indicator; DP = depression parent report indicator; AP = anxiety parent report indicator.

causes of these large discrepancies between the two rater types in more detail to find out which method should be trusted more. One way to investigate method effects in more detail would be to add covariates that might explain why parents judge their children's depression and anxiety so differently than do the children themselves (e.g., Eid et al., 2006) or to add additional (possibly more objective) measures or methods and study their association with the child and parent reports.

With regard to discriminant validity, we can see that the depression and anxiety state factors are highly positively correlated. This is the case for both the self-report (range = .74–.76) and the parent report (range = .79–.83), indicating that the discriminant validity of the two constructs is rather low for both types of ratings. Depression and anxiety represent two closely related concepts. Not surprisingly, their change scores are also highly positively correlated (range = .56–.83). According to both methods, change in depression is strongly related to change in anxiety. Interindividual differences in both depression and anxiety remained relatively stable over time as shown by the high stability coefficients in Table 2 (range = .61–.85).

The latent means and variances are shown at the bottom of Tables 2 and 3. All variances are different from zero, indicating that there is significant variability in the LS scores and in intraindividual change over time (i.e., children differ in their status and in the degree of change over time). If we compare the state means for depression and anxiety across self- and parent ratings (see the bottom of Table 2), we find that the parents reported consistently

lower levels of both depression (self-rating = 0.29–0.32; parent rating = 0.18–0.20) and anxiety (self-rating = 0.56–0.72; parent rating = 0.44–0.50) compared with the self-report.

The means of the latent change factors (see the bottom of Table 3) are all negative. This indicates that both constructs show a mean decline over time. Although this mean decline is rather small, it is consistent across both methods. However, the amount of mean decline differs between self- and parent ratings. That is, parent ratings show lower mean change over time (especially with regard to anxiety) than do self-ratings (self-rating depression = −0.03; parent rating depression = −0.02; self-rating anxiety = −0.10 to −0.16; parent-rating anxiety = −0.04 to −0.06).[6] In sum, the latent

---

[6] As a reviewer of this article noted, raw mean differences can sometimes be difficult to interpret, especially when raters used different scales. In this case, one may consider using standardized mean differences. In the present study, we did not use standardized mean differences because the self- and parent reporters used identical scales so that the raw means were directly comparable across raters. It should be noted that a standardization of mean differences using the standard deviations can also be misleading in some cases. For example, if groups differ in variability, a standardization magnifies the mean difference in the group with smaller variability and attenuates the mean difference in the group with larger variability. We recommend that standardized effect sizes be used only in cases in which measurement instruments are not comparable in terms of their metric.

Table 2

Estimated Latent MTMM-MO Correlation Matrix Estimated in the State Version of the Multimethod Change Model

Report indicator	1	2	3	4	5	6	7	8	9	10	11	12
1. S_{111} (Depression SR T1)	—											
2. S_{112} (Depression SR T2)	.73	—										
3. S_{113} (Depression SR T3)	.61	.74	—									
4. S_{211} (Anxiety SR T1)	.75	.60	.50	—								
5. S_{212} (Anxiety SR T2)	.61	.74	.59	.76	—							
6. S_{213} (Anxiety SR T3)	.55	.64	.76	.68	.80	—						
7. S_{121} (Depression PR T1)	**.36**	.29	.27	.25	.21	.17	—					
8. S_{122} (Depression PR T2)	.31	**.33**	.31	.24	.26	.26	.85	—				
9. S_{123} (Depression PR T3)	.21	.30	**.32**	.18	.15	.26	.68	.72	—			
10. S_{221} (Anxiety PR T1)	.30	.24	.24	**.27**	.22	.22	.79	.72	.52	—		
11. S_{222} (Anxiety PR T2)	.27	.28	.26	.27	**.28**	.29	.67	.83	.58	.80	—	
12. S_{223} (Anxiety PR T3)	.20	.26	.25	.21	.21	**.26**	.60	.64	.81	.74	.76	—
Ms	0.32 (0.01)	0.29 (0.01)	0.29 (0.01)	0.72 (0.02)	0.62 (0.02)	0.56 (0.02)	0.20 (0.01)	0.18 (0.01)	0.18 (0.01)	0.50 (0.01)	0.46 (0.01)	0.44 (0.02)
Variances	0.10 (0.01)	0.09 (0.01)	0.09 (0.01)	0.18 (0.01)	0.19 (0.01)	0.17 (0.01)	0.04 (0.01)	0.03 (0.00)	0.04 (0.01)	0.09 (0.01)	0.08 (0.01)	0.08 (0.01)

Note. Values in parentheses are standard errors. Bolded entries are monotrait-heteromethod correlations that indicate convergent validity on the same measurement occasion. Italicized entries are heterotrait-monomethod correlations that indicate discriminant validity on the same measurement occasion. Stability coefficients are framed in boxes. MTMM-MO = multitrait-multimethod-multioccasion; T = time; S_{cmt} = latent state factor; c = construct; m = method; t = occasion of measurement; SR = self-report; PR = parent report.

Table 3

Estimated Latent MTMM-MO Correlation Matrix Estimated in the Change Version of the Multimethod Change Model

Report indicator	1	2	3	4	5	6	7	8	9	10	11	12
1. S_{111} (DSR T1)	—											
2. $(S_{112} - S_{111})$ (DSR T2 – DSR T1)	-.41	—										
3. $(S_{113} - S_{111})$ (DSR T3 – DSR T1)	-.49	.63	—									
4. S_{211} (ASR T1)	.75	-.24	-.32	—								
5. $(S_{212} - S_{211})$ (ASR T2 – ASR T1)	-.18	.56	.38	-.32	—							
6. $(S_{213} - S_{211})$ (ASR T3 – ASR T1)	-.28	.43	.65	-.44	.65	—						
7. S_{121} (DPR T1)	**.36**	-.11	-.12	.25	-.06	-.11	—					
8. $(S_{122} - S_{121})$ (DPR T2 – DPR T1)	-.14	**.22**	.19	-.06	.18	.24	-.43	—				
9. $(S_{123} - S_{121})$ (DPR T3 – DPR T1)	-.18	.26	**.28**	-.09	.03	.25	-.39	.47	—			
10. S_{221} (APR T1)	.30	-.10	-.09	**.27**	-.07	-.08	.79	-.25	-.33	—		
11. $(S_{222} - S_{221})$ (APR T2 – APR T1)	-.08	.16	.11	-.02	**.14**	.14	-.27	.66	.38	-.40	—	
12. $(S_{223} - S_{221})$ (APR T3 – APR T1)	-.15	.23	.18	-.10	.10	**.18**	-.30	.32	.83	-.41	.50	—
Ms	0.32 (0.01)	-0.03 (0.01)	-0.03 (0.01)	0.72 (0.02)	-0.10 (0.01)	-0.16 (0.02)	0.20 (0.01)	-0.02 (0.01)	-0.02 (0.01)	0.50 (0.01)	-0.04 (0.01)	-0.06 (0.01)
Variances	0.10 (0.01)	0.05 (0.01)	0.07 (0.01)	0.18 (0.01)	0.09 (0.01)	0.11 (0.01)	0.04 (0.01)	0.01 (0.00)	0.03 (0.01)	0.09 (0.01)	0.03 (0.01)	0.05 (0.01)

Note. Values in parentheses are standard errors. Bolded entries are monotrait/change-heteromethod correlations that indicate convergent validity. Italicized entries are heterotrait/change-heteromethod correlations that indicate discriminant validity. MTMM-MO = multitrait-multimethod-multioccasion; T = time; S_{cmt} = latent state factor; c = construct; m = method; t = occasion of measurement; DSR = depression self-report; DPR = depression parent report; ASR = anxiety self-report; APR = anxiety parent report.

means point to a lack of convergent validity of mean level and mean change. Parent ratings appear to be less sensitive measures for assessing change in their children's depression and anxiety than are the children's self-ratings.

In summary, we notice that Campbell and Fiske's (1959) criteria for convergent and discriminant validity are not met in the present example. The convergent validity of self- and parent reports as well as the discriminant validity between depression and anxiety are very low. As one reviewer of this article noted, researchers may view these findings as rather discouraging. It should be noted, however, that the present results are rather typical for multimethod studies of personality traits (e.g., Fiske & Campbell, 1992), especially when traits with a rather low level of "visibility"—such as depression and anxiety—are considered (e.g., Funder, 1995). In such cases, low convergent validity is not unusual. With regard to the low level of discriminant validity, one must keep in mind that depression and anxiety represent two rather closely related concepts for which high correlations must be expected.

Structural Model: Prediction by Sex

To illustrate the use of covariates, we estimated an extended model in which the variable *sex* (dummy coded with 0 = boy, 1 = girl) was used as a predictor variable for all latent factors. The extended model also fit the data well, $\chi^2(164) = 162.66$, $p = .52$; $RMSEA = .000$; $CFI = 1.000$. Because we had no specific hypotheses and because there were 12 regression paths of interest (we did not attempt to interpret paths between sex and the IS factors), we used two-tailed z tests in conjunction with a Bonferroni-adjusted alpha level of $.05/12 = .004$ to identify significant relations with sex.

In the state version, sex was a significant predictor of all three anxiety self-report state factors (T1–T3; standardized path coefficients $\beta = .22, .26, .27$; SEs $= .03, .03, .04$; z scores $= 6.63, 8.34, 7.67$; ps $< .001$, respectively), indicating a sex difference in self-reported anxiety. Girls tended to report more anxiety than did boys. Interestingly, this sex difference is not evident on the basis of the parent reports. All paths from sex to the parent report state factors are close to zero (range $= -.10$ to $.003$) and not significantly different from zero according to the Bonferroni-adjusted alpha level.

For self-reported depression, a similar trend emerged (girls were somewhat more likely to be depressed), but the coefficients were smaller and significant only at T3 ($\beta = .16$, $SE = .04$; $z = 3.69$, $p < .001$). Latent change in depression and anxiety was not related to sex.

Discussion

Multimethod measurement is routinely used by many developmental psychology researchers. In this article, we have discussed an SEM approach that is useful to study change in longitudinal MTMM studies. In the final section of this article, we summarize advantages and limitations, provide guidelines for applied researchers, and show links to other models for MTMM-MO data.

Advantages

Why should a researcher prefer the present approach to conventional data analytic strategies for analyzing MTMM-MO data? The main advantage is that using SEM allows us to formulate and test a model for the entire data structure. At the same time, SEM allows us to take measurement error into account. As a consequence, we obtain more reliable estimates of convergent validity, discriminant validity, and stability because the relevant correlation coefficients are corrected for measurement error.

In addition, we can use the model to study latent change for different methods simultaneously and to test important assumptions such as, for instance, the assumption of measurement invariance over time. Furthermore, we can assess important properties of the measures such as convergent validity, discriminant validity, indicator-specificity, and reliability. Studying the convergent validity of change is very important as different methods may lead to different conclusions about change. For example, in the present application, we found that self- and parent reports of depression and anxiety showed poor convergent validity of change—with respect to both, interindividual differences in change and mean change over time.

A very useful feature of the multimethod change model is that it provides a latent MTMM-MO correlation matrix. This matrix is in line with Campbell and Fiske's (1959) MTMM matrix, but it has the important advantage that the correlations are estimated among latent factors and thus are corrected for measurement error. The latent MTMM-MO matrix available in the LS version of the model allows quantifying the convergent and discriminant validity as well as the stability of constructs. The MTMM-MO matrix obtained from the LD version allows researchers to apply Campbell and Fiske's criteria for convergent and discriminant validity to the measurement of change.

Another advantage of the SEM approach is its flexibility. For example, the multimethod change model can be used to study change with respect to a baseline measurement occasion or between adjacent measurement occasions (Steyer et al., 1997, 2000). In addition, the model can easily be expanded to include predictors, outcomes, or covariates of the latent factors already in the model. In this way, we can, for example, relate latent change in one construct to change in other constructs. We can also include predictor variables to explain why some individuals change more than others or to study the causes of low convergent validity of different methods.

To investigate why methods show low convergent validity, we can test whether outcome variables are differentially related to change variables on the basis of different methods (convergent validity of prediction). For instance, in the present reanalysis, we identified sex as a significant predictor of anxiety states as measured by self-report (there was a tendency for girls to have higher latent anxiety scores) but not of any of the parent report states or any change scores. As noted by an anonymous reviewer of this article, this might either indicate that the gender differences represent a method artifact of the self-report method or that parents are insensitive to gender differences in anxiety.

Limitations

Sample size. The model presented here requires relatively large samples. We encourage researchers who are in doubt as to whether their sample size is sufficient to conduct an application-oriented simulation study (Bandalos, 2006; Paxton, Curran, Bollen, Kirby, & Chen, 2001). B. O. Muthén and Muthén (2002)

discussed how such a simulation can easily be implemented in the Mplus program.

Number of time points. Although the model presented here can in principle be applied to any number of time points, readers should be aware of a potential limitation that emerges when only two time points are used. This limitation has to do with the specification of IS factors to capture IS effects. When only two occasions are used, identification and estimation problems (e.g., improper solutions with out-of-range parameter estimates) occur more frequently (see Ciesla, Cole, & Steiger, 2007). The problems are often related to the IS factors, whose identification status can be deficient when there are only two time points (see our discussion of remedies in this case above).

Also, models with IS factors may not fit well if IS effects are not unidimensional. For example, residual covariances among indicators that are due to IS effects might follow an autoregressive structure rather than the unidimensional (one-factor) structure implied by the model with IS factors. With more than three time points, deviations from unidimensionality would be indicated by a lack of fit.

Another problem related to IS effects that does not depend on the number of time points occurs if these effects account for a large portion (say more than 20–30%) of the variance of the indicators. This would indicate that the indicators are rather heterogeneous and mirror distinct facets of a construct. The results obtained from the structural model (e.g., the MTMM-MO matrix or correlations with covariates) might then strongly depend on the choice of the reference indicators for which no IS factors are specified.

The best way to avoid all these problems is to identify and select homogeneous indicators for each construct. For homogenous indicators, no IS factors are needed, and a less complicated version of the model without these factors can be used. If this is not possible and if IS effects do not follow a unidimensional structure, then the use of a CU approach (e.g., Cole & Maxwell, 2003), in which the residual variables for the same indicators are allowed to correlate over time, is an appropriate alternative modeling strategy.

If IS factors account for a large portion of the indicator variance, researchers should conduct sensitive analyses using different reference indicators. If the parameter estimates for the structural model differ significantly for different reference indicators, then it is important to carefully select the reference indicator on the basis of substantive considerations: Which indicator can be seen as the "gold standard" that best represents the construct?

Measurement invariance. A requirement for the proper interpretation of the LD factors is that the measurement model for each construct is invariant over time. At least strong factorial invariance (time-invariant intercepts and factor loadings) is required to ensure that model parameters related to the LD factors are interpretable. As we showed above, this assumption is testable, and in the present application, it was not rejected by the goodness-of-fit criteria. If measurement invariance were not tenable, the interpretation of the LD factors would be compromised. In some cases, however, researchers will be able to establish partial measurement invariance (Byrne et al., 1989). As mentioned above, partial measurement invariance means that invariance is tenable at least for some indicators of a construct, a condition that can re-enable judicious interpretation.

Links to Other Models

Finally, we think that it is useful to point out some links to other models for analyzing MTMM-MO data. In Cole and Maxwell's (2003) CU approach, three types of shared method variance can be represented by correlations among residual variables ("uniquenesses"): (a) within-wave, cross-construct CUs for indicators pertaining to the same method mirror method effects on a given time point; (b) cross-wave, within-construct CUs for the same indicator capture method variance caused by stable IS effects; and (c) cross-wave, cross-construct CUs may be admitted to account for additional effects due to the same method.

Cole and Maxwell's (2003) approach of modeling different types of shared method effects through residual correlations parallels Kenny's (1976) correlated trait-CU model for cross-sectional MTMM data (see also Lance et al., 2002). The advantage of Cole and Maxwell's model is that method and error effects are taken into account so that structural (e.g., mediational) hypotheses can be more accurately tested.

On the other hand, the CU approach has a number of limitations: (a) many uniqueness correlations need to be estimated in designs with many indicators, constructs, methods, and time points; (b) random measurement error cannot be separated from IS variance and shared method variance leading to an underestimation of the reliabilities of the measures; and (c) variance components due to construct, method, indicator-specificity, and error are not available.

A model that overcomes some of the limitations of the CU approach is Burns et al.'s (2003; see also Burns & Haynes, 2006) multi-occasion extension of the CT-CM model (Marsh, 1989; Marsh & Bailey, 1991; Marsh & Grayson, 1995). Instead of including various kinds of CUs, Burns et al.'s model uses occasion-specific method factors to capture cross-sectional method effects. The model represents an extension of the LS model with one state factor for each construct and one method factor for each method on each occasion of measurement.

Burns et al.'s (2003) approach allows separating occasion-specific variance due to a construct from occasion-specific method variance and error influences. Furthermore, different methods can be correlated and external variables can be included to explain method effects. A potential shortcoming is that this approach is based on the cross-sectional CT-CM model and thus might be prone to similar identification, estimation, and interpretation problems. Like Cole and Maxwell's (2003) model, the model presented by Burns et al. is a single indicator model. Hence, it cannot be used to study construct-specific method effects.

Vautier et al. (2008) recently presented a true change model with method effects. Their model is useful to deal with heterogeneous indicators in longitudinal studies. However, it is not specifically designed to analyze MTMM-MO data. Vautier et al.'s model includes only one indicator per CMOU, whereas the models discussed here are multiple indicator models. As a consequence, Vautier et al.'s model assumes perfect temporal stability of method effects and thus is not suitable for analyzing the convergent and discriminant validity of change. In contrast, the models presented here are specifically designed to deal with changes in all methods. Note, however, that Vautier et al.'s model could be extended to a multiple indicator model in which the assumption of perfect stability of method effects could be tested.

Some research questions in developmental psychology are less concerned with measuring change and are more concerned with the assessment of situation-specific fluctuations around a stable trait value. For example, a researcher might be primarily interested in the degree to which situation-specific variables affect the measurement of these constructs (e.g., Vautier, 2004). If the goal were to separate stable from situation-specific components of variance in a MTMM-MO study, the model proposed by Courvoisier et al. (2008) can be applied. Courvoisier et al.'s model represents an extension of the conventional LST model (Steyer, Schmitt, & Eid, 1999) to a multimethod LST model. A related approach has been presented by Vautier (2004), who showed how method effects caused by bipolar items can be studied in an extended LST model.

Readers might wonder how the LD approach discussed here compares with latent growth curve modeling (LGCM; Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006). The difference between LD models and LGCM is that LD models consider change as such, whereas LGCM are concerned with a special component of change (e.g., linear or curvilinear growth; Steyer et al., 2000). In that sense, LD models are less restrictive than LCGM, as the parameters of a LD model do not imply any specific functional form of change but just consider "pure" change. As a practical consequence, LD models may be used as alternative models if LGCM do not fit because change does not follow a simple pattern, such as, for example, straight line growth.

Conclusion

Using more than one method or source to assess the concepts of interest in a study is highly recommended in all areas of psychological research (Eid & Diener, 2006). Multimethod assessment is of particular importance to developmental psychology research, given that the measurement of change is the focus of many studies. With only a single method or source, it may not be possible to draw valid conclusions about change. MTMM-MO studies enable researchers to study the convergent and discriminant validity of change. Therefore, we encourage developmental researchers to use multiple constructs and multiple methods as well as sophisticated methods of data analysis, such as SEM, not only in cross-sectional investigations but also in longitudinal investigations.

References

Allen, J. P., Porter, M. R., McFarland, C. F., Marsh, P. A., & McElhaney, K. B. (2005). The two faces of adolescents' success with peers: Adolescent popularity, social adaptation, and deviant behavior. *Child Development, 76,* 747–760.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Erlbaum.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385–426). Greenwich, CT: Information Age Publishing.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

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

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21,* 205–229.

Burns, G. L., & Haynes, S. N. (2006). Clinical psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 401–418). Washington, DC: American Psychological Association.

Burns, G. L., Walsh, J. A., & Gomez, R. (2003). Convergent and discriminant validity of trait and source effects in ADHD-inattention and hyperactivity/impulsivity measures across a 3-month interval. *Journal of Abnormal Child Psychology, 15,* 529–541.

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

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

Ciesla, J. A., Cole, D. A., & Steiger, J. H. (2007). Extending the trait-state-occasion model: How important is within-wave measurement equivalence? *Structural Equation Modeling, 14,* 77–97.

Cole, D. A., Martin, J. M., Peeke, L., Henderson, A., & Harwell, J. (1998). Validation of depression and anxiety measures in White and Black youths: Multitrait-multimethod analyses. *Psychological Assessment, 10,* 261–276.

Cole, D. A., Martin, J. M., & Powers, B. (1997). A competency-based model of child depression: A longitudinal study of peer, parent, teacher, and self-evaluations. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 38,* 505–514.

Cole, D. A., Martin, J. M., Powers, B., & Truglio, R. (1996). Modeling causal relations between academic and social competence and depression: A multitrait-multimethod longitudinal study of children. *Journal of Abnormal Psychology, 105,* 258–270.

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

Cole, D. A., Truglio, R., & Peeke, L. (1997). Relation between symptoms of anxiety and depression in children: A multitrait-multimethod-multigroup assessment. *Journal of Consulting and Clinical Psychology, 65,* 110–119.

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

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics, 13,* 131–147.

Dumenci, L. (2000). Multitrait-multimethod analysis. In S. D. Brown & H. E. A. Tinsley (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 583–611). San Diego, CA: Academic Press.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-

multimethod models: A multiple indicator CT-C(*M*–1) model. *Psychological Methods, 8,* 38–60.

Eid, M., Schneider, C., & Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality, 13,* 283–306.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670.

Ghisletta, P., & Lindenberger, U. (2005). Exploring the structural dynamics of the link between sensory and cognitive functioning in old age: Longitudinal evidence from the Berlin Aging Study. *Intelligence, 33,* 555–587.

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging, 18,* 639–657.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 129–169). Cambridge, MA: Abt.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8,* 325–352.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Kovacs, M. (1985). The Children's Depression Inventory (CDI). *Psychopharmacology Bulletin, 21,* 995–998.

Kwok, O., Haine, R. A., Sandler, I. N., Ayers, T. S., Wolchik, S. A., & Tein, J.-Y. (2005). Positive parenting as a mediator of the relations between parental psychological distress and mental health problems of parentally bereaved children. *Journal of Clinical Child and Adolescent Psychology, 34,* 260–271.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait–correlated method and correlated uniqueness models for multitrait–multimethod data. *Psychological Methods, 7,* 228–244.

Little, R. J. A., & Rubin, D. B. (2003). *Statistical analysis with missing data.* New York, NY: Wiley.

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

Majdandzic, M., & van den Boom, D. C. (2007). Multimethod longitudinal assessment of temperament in early childhood. *Journal of Personality, 75,* 121–167.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15,* 47–70.

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

Marsh, H. W., & Grayson, D. A. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousands Oaks, CA: Sage.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In R. B. Cattell & J. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York, NY: Plenum Press.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analysis with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 137–175). Washington, DC: American Psychological Association.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–267). Hillsdale, NJ: Erlbaum.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association.

Morris, A. S., Robinson, L. R., & Eisenberg, N. (2006). Applying a multimethod perspective to the study of developmental psychology. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 371–384). Washington, DC: American Psychological Association.

Muthén, B. O. (1998–2004). *Mplus technical appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Muthén, L. K. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

Ostrov, J. M., & Crick, N. R. (2007). Forms and functions of aggression during early childhood: A short-term longitudinal study. *School Psychology Review, 36,* 22–43.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287–312.

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

Reynolds, C. R., & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6,* 271–280.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 28,* 138–151.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

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

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

Steyer, R., Partchev, I., & Shanahan, M. (2000). Modeling true intraindividual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 109–126). Hillsdale, NJ: Erlbaum.

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

Tisak, J., & Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods, 5,* 175–198.

Tram, J. M., & Cole, D. A. (2006). A multimethod examination of the stability of depressive symptoms in childhood and adolescence. *Journal of Abnormal Psychology, 115,* 674–686.

Vautier, S. (2004). A longitudinal SEM approach to STAI data: Two comprehensive multitrait-multistate models. *Journal of Personality Assessment, 83,* 167–179.

Vautier, S., Steyer, R., & Boomsma, A. (2008). The true change model with individual method effects: Reliability issues. *British Journal of Mathematical and Statistical Psychology, 61,* 369–399.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1–26.

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

Wothke, W. (1995). Covariance components analysis of the multitrait-multimethod matrix. In P. E. Shrout (Ed.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Hillsdale, NJ: Erlbaum.

Zhou, Q., Eisenberg, N., Losoya, S. H., Fabes, R. A., Reiser, M., Guthrie, I. K., . . . Shepard, S. A. (2002). The relations of parental warmth and positive expressiveness to children's empathy-related responding and social functioning: A longitudinal study. *Child Development, 73,* 893–915.