Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Modeling Common Traits and Method Effects in Multitrait-Multimethod Analysis

Steffi Pohl[a]; Rolf Steyer[a]
[a] Otto-Friedrich-University Bamberg,

Online publication date: 25 February 2010

To cite this Article Pohl, Steffi and Steyer, Rolf(2010) 'Modeling Common Traits and Method Effects in Multitrait-Multimethod Analysis', Multivariate Behavioral Research, 45: 1, 45 — 72
To link to this Article: DOI: 10.1080/00273170903504729
URL: http://dx.doi.org/10.1080/00273170903504729

Multivariate Behavioral Research, 45:45–72, 2010

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170903504729

Psychology Press
Taylor & Francis Group

Modeling Common Traits and Method Effects in Multitrait-Multimethod Analysis

Steffi Pohl and Rolf Steyer
Otto-Friedrich-University Bamberg

Method effects often occur when constructs are measured by different methods. In traditional multitrait-multimethod (MTMM) models method effects are regarded as residuals, which implies a mean method effect of zero and no correlation between trait and method effects. Furthermore, in some recent MTMM models, traits are modeled to be specific to a certain method. However, often we are not interested in a method-specific trait but in a trait that is *common to all methods*. Here we present the Method Effect model with common trait factors, which allows modeling "common" trait factors and method factors that represent method "effects" rather than residuals. The common trait factors are defined as the mean of the true-score variables of all variables measuring the same trait and the method factors are defined as differences between true-score variables and means of true-score variables. Because the model allows estimating mean method effects, correlations between method factors, and correlations between trait and method factors, new research questions may be investigated. The application of the model is demonstrated by 2 examples studying the effect of negative, as compared with positive, item wording for the measurement of mood states.

MULTITRAIT-MULTIMETHOD RESEARCH

Constructs regarded in the social sciences may usually be measured by different methods; for example, self, parent, and teacher ratings are used for the measure-

Correspondence concerning this article should be addressed to Steffi Pohl, Otto-Friedrich-University Bamberg, NEPS, 96045 Bamberg, Germany. E-mail: steffi.pohl@uni-bamberg.de

ment of children's competencies, questionnaire and behavior observation for the measurement of anxiety, or positively and negatively worded items for the measurement of mood states. For the validation of a measurement instrument Campbell and Fiske (1959) proposed to use multitrait-multimethod (MTMM) designs, where each of several *traits* (constructs) is measured by each of several *methods*. According to Campbell and Fiske, discriminant validity is supported when the trait under investigation may well be distinguished from other traits and convergent validity is achieved when different measurement methods yield similar results in measuring the same trait.

Today, MTMM designs are not only used for the validation of a measurement instrument but also for the measurement of constructs in general. Because there is usually no method with undoubted validity, researchers frequently use several methods to measure a construct. Examples are Mount (1984), who used superior, self, and peer ratings for the measurement of managerial performance; Villar, Luengo, Gómez-Fraguela, and Romero (2006), who measured parenting constructs using adolescent, mother, and father reports; and McConnell and Leibold (2001), who used implicit association tests, explicit measures, and external ratings for the measurement of racial attitudes. Usually different methods do not yield the same result for the measurement of a construct, but instead systematic differences between the different measurements exist that are person-specific. If systematic person-specific method effects exist, these measures are not unidimensional. Thus, in data analysis, ignoring systematic individual method effects often results in an unsatisfactory model fit. In order to get a satisfactory model fit, method effects need to be accounted for in the model.

Although in some studies method effects are regarded as nuisance effects (e.g., Cole, Martin, Powers, & Truglio, 1996; Gignac, 2006; Motl & DiStefano, 2002), they are the focus of research in other studies. Typical examples are the investigation of the effect of different raters on the measurement of constructs (see Clausen, 2002; Conway & Huffcuff, 1997, for examples of rater effects on teaching quality and job performance, respectively) or the investigation of the effect of item wording (positive or negative) on ratings measuring different psychological constructs (see, e.g., Horan, DiStefano, & Motl, 2003; Marsh, 1996; Russell & Carroll, 1999; Steyer & Riedl, 2004; Vautier & Pohl, 2008; Watson & Tellegen, 1999).

In order to analyze data measured with multiple methods different models have been proposed, most of them based on confirmatory factor analysis (see, e.g., Marsh, 1989; Widaman, 1985). The most frequently applied models are the correlated trait-correlated uniqueness model (CTCU; Kenny, 1976; Marsh, 1989; Marsh & Craven, 1991), the correlated trait-correlated methods model (CTCM; Jöreskog, 1974; Widaman, 1985), and the correlated trait-correlated method minus one model (CTC[M-1]; Eid, 2000). A more recent model is the Method Effect model with a reference method (Pohl, Steyer, & Kraus, 2008).

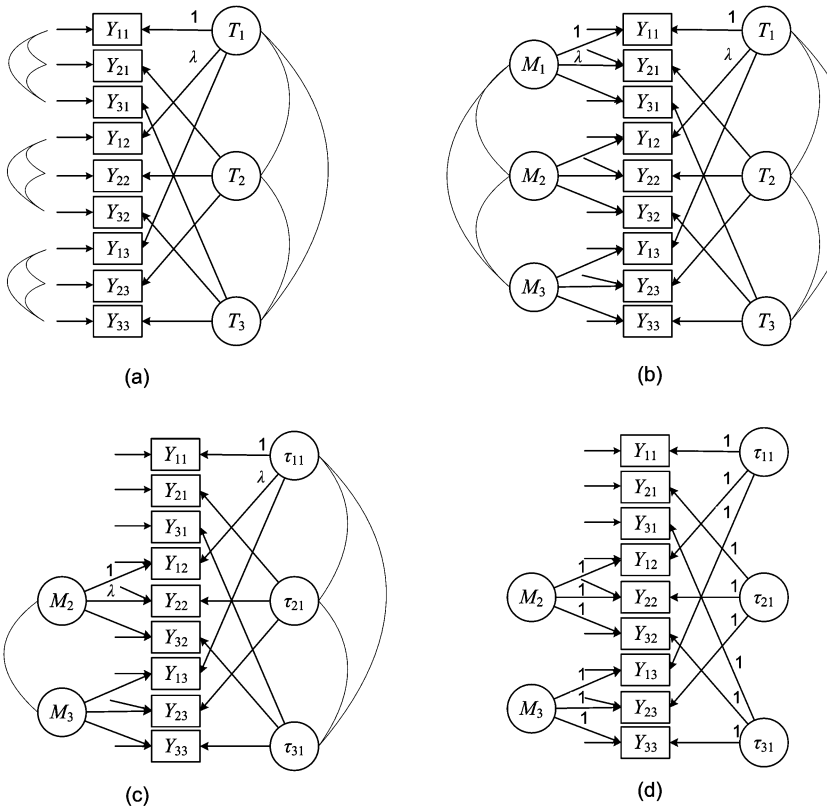

FIGURE 1 The (a) correlated trait-correlated uniqueness (CTCU), (b) correlated trait-correlated methods (CTCM), (c) correlaetd trait-correlated (methods minus 1) (CTC(M-1)), and (d) Method Effect model with a reference method (MEref) model for nine variables measuring three traits by three methods. Arcs denote correlations.

These models are depicted in Figure 1 for nine manifest variables Y_{tj} measuring three traits (t) by three methods (j). The latent variables represent trait factors (T_t and τ_{tj}) and method factors (M_j).

The Correlated Trait-Correlated Uniqueness Model

In the CTCU model (see Figure 1(a)), method effects are accounted for by error covariances. Thus the error terms in this model not only represent unsystematic measurement error but also systematic method effects. Because the method effects are part of the error term, the reliability of the measurement is underestimated. Correlations between different method effects are not allowed.

If, for example, the effect of different raters on the measurement of managerial performance of an employee is investigated, the model does not allow that different raters show similar response tendencies. Instead, it is assumed that the effects of different raters are independent of each other. This is not always a plausible assumption and Conway, Lievens, Scullen, and Lance (2004) showed that biased estimates for the trait variances and covariances may result if the assumption of uncorrelated method effects is not met. Furthermore, because the method effects are only accounted for by allowing for covariances between error variables, no explanatory variables can be included in the model that may explain the interindividual differences in the method effects.

The Correlated Trait-Correlated Methods Model

As in the CTCU model, in the CTCM model (see Figure 1(b)) the trait factors represent "trait variance" (Marsh, 1989, p. 357), that is, variance that is *common* to all variables measuring this trait by different methods. Method effects are represented by a method factor for each method. The variance of the manifest variables may be additively decomposed into trait, method, and error variance. Correlations between different method factors may be estimated and the method factors may be explained by external variables. However, in contrast to the CTCU model, the CTCM model suffers from identification and estimation problems, especially when the method factors are correlated (Marsh, 1989; Marsh & Grayson, 1995).

The Correlated Trait-Correlated (Methods Minus One) Model

The more recent correlated trait-correlated method minus one (CTC[M-1]) model (see Figure 1(c)) overcomes some of the limitations of the other MTMM models. In this model a reference method needs to be chosen for which no method factor is modeled, whereas all other methods have their own method factor. This model avoids overparameterization and the resulting identification and estimation problems. The trait factors are defined as true-score variables of the manifest variables measuring the trait by the reference method (τ_{t1}) and the method factors are defined as residual variables of the regression of the true-score variable of a manifest variable measuring a trait with a specific method on the true-score variable of a manifest variable measuring the same trait with the reference method (i.e., on the trait variable). Note that here, in contrast to the CTCU and CTCM model, where *common* trait factors are modeled, *method-specific* trait factors are defined. As in the CTCM model the method factors may correlate with each other and the variance of the manifest variables may be additively decomposed into trait, method, and error variance. A limitation of the model is that model fit and variance of the method factors are not invariant to the choice

of the reference method, thus leading to different substantive interpretations for different reference methods.

How to Model Method Effects

In the MTMM models presented earlier, method effects are implicitly (CTCM, CTCU) or explicitly (CTC[M-1]) regarded as residuals. Jöreskog (1971) interpreted method factors as "what is left over after all trait factors have been eliminated" (p. 128) and stated that they are "independent of the particular traits that the method is used to measure" (p. 128). In the CTCU model method effects are represented in the error terms (which are uncorrelated with the traits) and in the CTC(M-1) model the method factors are explicitly defined as regression residuals. Defining method effects as residuals implies that they have the typical properties of a residual: their mean is zero and they do not correlate with the predictor, that is, with the trait factor in MTMM models.

But why should we assume that the mean of the method factor, that is, the mean over the individual method effects, is zero? In fact, it is quite plausible that participants rate their mood state on average higher on negatively worded items than on positively worded ones or that parents rate the competencies of their children on average higher than the children themselves. Furthermore, why should the method effect be uncorrelated with the trait? Marsh and Grayson (1995) already stated that "the lack of correlation between trait and method factors is an assumption that may be unrealistic in some situations. The constraint seems to be routinely applied to avoid technical estimation problems and to facilitate decomposition of variance into trait and method effects, not because of the substantive likelihood or empirical reasonableness" (p. 181). In applications the amount of overestimation of competencies by parents compared with the self rating of their children may be larger for low competencies than for high ones, or the participants may differentiate more between their answers on negatively as compared with positively worded items when they are tired than when they are alert.

Furthermore, in models defining method effects as residuals a method effect of zero does not necessarily indicate that there is no difference between the competency ratings of parents and children. Instead a method effect of zero indicates that the observed rating does not differ from the expected rating given the rating on another method. Consider an artificial example where all parents overestimate the competencies of their child by one unit. Then the regression equation of the true score of the parent rating (τ_{CP}) on the true score of the child rating (τ_{CC}) would be $\tau_{CP} = 1 + \tau_{CP}$. Although the parent and the child rating differ, the error term, and thus the method effect, would be zero for all persons. This, however, is not always what we want. Usually we want the method effect to represent the difference in the ratings with a method effect

of zero indicating no difference. Residuals are, thus, not necessarily the most appropriate representation of method effects.

The Method Effect Model With a Reference Method

A new model in which the method factors are not modeled as residuals but as "effects" is the Method Effect model with a reference method (MEref; Pohl et al., 2008). In this model, a reference method needs to be chosen and the trait factors (τ_{t1}) are, as in the CTC(M-1) model, defined as the true-score variable of the manifest variable measuring the respective trait with the reference method. The method effect variables are defined by

$$M_{tj} \equiv \tau_{tj} - \tau_{t1}. \tag{1}$$

Hence, the method effect variable M_{tj} is the difference between the true-score variable of the measurement of trait t obtained under method j and the true-score variable of the measurement of the same trait t measured by method $k = 1$, the so-called reference method.[1] In Equation 1, the difference $\tau_{tj} - \tau_{t1}$ between the two true-score variables represents the systematic effects of using method j, instead of reference method 1, for measuring trait t. Note that each method may be chosen to be the reference method and that the method effects represent the effect of a certain method compared with the reference method. In order to specify an identified model, the measurement of different traits is regarded. In Figure 1(d), the MEref model for measuring three traits by three methods is depicted. It is assumed that the method effect variables M_{tj} of method j are the same for all traits, that is,

$$M_{1j} = M_{2j} = M_{3j} \equiv M_j. \tag{2}$$

The *trait factors* τ_{11}, τ_{21}, and τ_{31} in Figure 1(d) represent the *true-score variables* of the manifest variables, measuring the traits *with the reference method* 1. Note that the trait factors will change (to some degree) if the reference method is changed. The values of the *method factor* M_2 and M_3 in the MEref model represent the method effects of methods 2 and 3, respectively, compared with the reference method 1, *for all three traits*. Note that there are no restrictions on the means or on the covariances of the latent variables. The method factors

[1]Note that this definition of a method effect may be regarded as a latent difference according to the latent difference approach by McArdle (2001) and Steyer, Eid, and Schwenkmezger (1997). It is also compatible with the definition of individual causal effects presented in Steyer, Partchev, Kröhne, Nagengast, and Fiege (in press).

may have a mean different from zero; they may correlate with each other as well as with the trait factors.

The MEref model has many advantages. Because the method effects are defined as effects, they have the usual properties of an effect, that is, (a) an effect of zero indicates no difference and (b) the size of the effect does not change with the direction of the comparison. In the example of using self-report and subordinate ratings for the measurement of managerial performance, a method effect of zero indicates that the self and the subordinate rating do not differ in their true scores. Furthermore, the size of the method effect is rather the same when comparing the self rating with the subordinate rating than when comparing the subordinate rating with the self rating. These properties do not hold for the CTC(M-1) model (see Pohl et al., 2008, for a detailed explanation). The model not only allows estimating the covariances of the method effects among each other but—because method effects are defined as effects rather than residuals—also the covariances of the method effects with the traits as well as the average method effect.

ADVANTAGES AND LIMITATIONS OF THE MODELS FOR MULTITRAIT-MULTIMETHOD DESIGNS

In traditional MTMM models, like the CTCM and the CTCU model, the factors represent trait or method variance (Marsh, 1989). Although the trait and method variances may be estimated in the models it is not clear what the "trait" or "method" variable is, the variance of which we are looking at, that is, the factors are not clearly defined in these models. Also in the CTC(M-1) and the MEref model trait variance and method variance is estimated. There, however, the factors are clearly defined. In the CTC(M-1) model the method effects are residuals of a regression, whereas in the MEref model they are differences in true scores. Due to the different definition of the method factors the method variance is different in the two models. Specifying the factors as just representing trait or method variance is thus not enough to clearly define the factors. Instead, the variables the factors represent need to be clearly defined. Only a clear definition of the factors helps us to distinguish the factors in different models (e.g., the method factors in the CTC[M-1] and in the MEref model) and thus to decide which model to choose for our research question. A substantive interpretation of the factors is only possible when we know what variables these factors represent. Furthermore, a clear definition of the factors also helps to derive restrictions (e.g., on the mean and correlation of method factors).

Whereas in the CTCM and CTCU model the trait factors represent a common trait of all variables measuring this trait, in the CTC(M-1) and the MEref model method-specific traits are modeled. In both models, the CTC(M-1) and

the MEref model, a reference method has to be chosen and the trait factor represents the true-score variable of the the reference method measuring this trait. In some applications a reference-method specific trait may be appropriate, for example, when measuring extraversion with self and peer ratings and when we are interested in extraversion measured by the self rating. However, we are often not interested in a *reference-method specific trait* but rather in *a trait common to all variables* measuring that construct. For example, we are not always interested in well-being measured by positively (or by negatively) worded items. Instead, we might want to aggregate these different measures in order to get a measure of well-being *common to both methods*. The same applies to the measurement of school performance by oral and by written exams, where we are usually interested in an aggregate of performance measures over both types of exams. Already Epstein (1983, 1986) argued that aggregation over modes of measurement cancels out method-specificity and may increase the reliability and validity of the measurement.

Whereas in the CTCU, CTCM, and CTC(M-1) model method effects are defined as residuals, the method effects in the MEref model are defined as effects. Mean method effects as well as the correlation between method factors and traits may be estimated in this model. In contrast to the CTC(M-1) and the MEref model, where the traits are reference-method specific, in the CTCU and the CTCM model common trait factors are modeled. However, only in the CTC(M-1) and the MEref model are the factors clearly defined. In the method effect model with common trait factors (MEcom model) we combine the different advantages of the previous models. In the MEcom model the latent variables are clearly defined. A common trait is modeled and the method effects are defined as effects rather than residuals, allowing the estimation of the mean method effect as well as the correlation between method factors and traits.

THE METHOD EFFECT MODEL WITH COMMON TRAIT FACTORS

The Method Effect model with common trait factors is a new parametrization of the MEref model. Here we start with the definition of the factors and then introduce the assumptions defining the model.

Definition of the Factors

Let Y_{11} and Y_{12} be two manifest variables measuring the same trait 1 by methods 1 and 2, respectively. The *common trait* is defined as

$$\tau_1 \equiv \frac{\tau_{11} + \tau_{12}}{2}, \tag{3}$$

the mean of the true-score variables τ_{11} and τ_{12} of the two manifest variables Y_{11} and Y_{12}. Hence, the trait factor is no longer specific to one of the two methods but *common to both methods*. The *method-effect variables* may be defined as

$$M_{11} \equiv \tau_{11} - \tau_1 \tag{4}$$

$$M_{12} \equiv \tau_{12} - \tau_1, \tag{5}$$

representing the difference between the true-score variables of the manifest variables measuring a trait by a certain method and the common trait.[2] If there are systematic person-specific method effects, the true-score variables of Y_{11} and Y_{12} will differ from each other and also from the common trait factor τ_1. Note that the method-effect variables are still *trait-specific*, that is, for each trait there may be different person-specific method effects. Using the definition of the method-effect variables M_{11}, the manifest variable Y_{11} may be decomposed as follows:

$$\begin{aligned} Y_{11} &= \tau_{11} + \varepsilon_{11} \\ &= \tau_1 + \tau_{11} - \tau_1 + \varepsilon_{11} \\ &= \tau_1 + M_{11} + \varepsilon_{11}, \end{aligned} \tag{6}$$

which is always true. Similarly, for the manifest variable Y_{12},

$$\begin{aligned} Y_{12} &= \tau_{12} + \varepsilon_{12} \\ &= \tau_1 + \tau_{12} - \tau_1 + \varepsilon_{12} \\ &= \tau_1 + M_{12} + \varepsilon_{12}, \end{aligned} \tag{7}$$

which is also always true.

The definition of the method-effect variables has some implications. Because a method-effect variable is defined as the difference between the true-score variable of a manifest variable and the mean of the true-score variables of all manifest variables measuring the same trait (see Equations 4 and 5), the following equation always holds:

$$M_{11} + M_{12} = 0, \tag{8}$$

[2]This definition of a method effect corresponds to the definition of individual causal effect variables for comparing a treatment with all treatments in the theory of individual and average causal effects by Steyer et al. (in press).

implying,

$$M_{11} = -M_{12}. \tag{9}$$

This leads to the following measurement equations:

$$Y_{11} = \tau_1 - M_{12} + \varepsilon_{11} \tag{10}$$

$$Y_{12} = \tau_1 + M_{12} + \varepsilon_{12}. \tag{11}$$

Considering m manifest variables Y_{1j}, measuring trait 1 by m methods, Equations 8 and 9 generalize as follows:

$$M_{11} + M_{12} + \ldots + M_{1m} = 0, \tag{12}$$

and thus

$$M_{11} = -M_{12} - \ldots - M_{1m}. \tag{13}$$

Hence, one method-effect variable—here method 1—pertaining to trait 1 is a deterministic function of the other method-effect variables pertaining to that trait.

Assumptions Defining the Model

In the preceding paragraph we defined the method-effect variable for a *single trait*. In order to define the Method Effect model with common trait factors, let us now consider two more manifest variables measuring a second trait by the same two methods. Consistent with the definitions of the first trait, the common trait and the method-effect variables may be defined for the manifest variables Y_{21} and Y_{22} measuring trait 2 by methods 1 and 2, respectively:

$$\tau_2 = \frac{\tau_{21} + \tau_{22}}{2} \tag{14}$$

$$M_{21} = \tau_{21} - \tau_2 \tag{15}$$

$$M_{22} = \tau_{22} - \tau_2, \tag{16}$$

where τ_{21} and τ_{22} denote the true-score variables of the manifest variables Y_{21} and Y_{22}, respectively. The manifest variables may now be written as

$$Y_{21} = \tau_2 - M_{22} + \varepsilon_{21} \tag{17}$$

$$Y_{22} = \tau_2 + M_{22} + \varepsilon_{22}, \tag{18}$$

which is always true.

So far, we have provided definitions; however, no assumptions have been specified. In order to identify the theoretical parameters such as the expected values, the variances, and the covariances of the latent variables, we have to introduce appropriate assumptions. The *first assumption* is

$$M_{11} = M_{21} \equiv M_1, \tag{19}$$

saying that the method-effect variables M_{11} and M_{21} are the same for each of the two traits. If we investigate the effect of using a scale with positively worded items (method 1), as compared with a scale with negatively worded items (method 2), on the measurement of alertness (trait 1) and calmness (trait 2), this assumption implies that the method effect due to item wording for the measurement of alertness is the same as for the measurement of calmness. This assumption may not be realistic in all applications. In longitudinal designs, however, where the same construct is measured with the same methods at different occasions, this assumption is quite plausible (see Application 1).

Because $M_{11} = -M_{12}$ and $M_{21} = -M_{22}$ (see Equation 9), the assumption in Equation 19 implies

$$M_{12} = M_{22} \equiv M_2 \tag{20}$$

and therefore

$$M_1 = -M_2. \tag{21}$$

Hence, the equations of the measurement model for all manifest variables simplify as follows:

$$Y_{11} = \tau_1 - M_2 + \varepsilon_{11} \tag{22}$$

$$Y_{12} = \tau_1 + M_2 + \varepsilon_{12} \tag{23}$$

$$Y_{21} = \tau_2 - M_2 + \varepsilon_{21} \tag{24}$$

$$Y_{22} = \tau_2 + M_2 + \varepsilon_{22}. \tag{25}$$

A path diagram representing these equations is depicted in Figure 2(a). The trait factors in Figure 2(a) represent the common trait variables τ_1 and τ_2. The *method factor* M_2 in Figure 2(a) represents the *method effects* of method 2 compared with the common trait. Note that these method effects are assumed to be the same *for both traits*. The *method effects* of method 1 compared with the common trait are modeled by factor loadings of -1 on the method factor M_2. There are no restrictions on the means of the latent variables nor on the covariances between the latent variables.

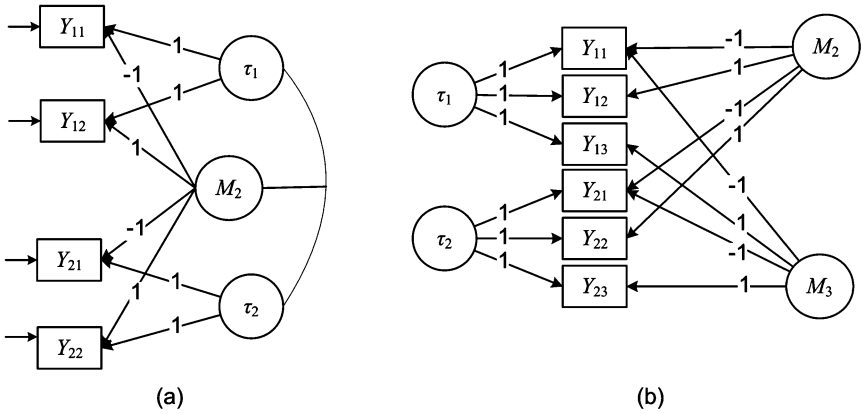

FIGURE 2 (a) The MEcom model for four variables measuring two traits by two methods. Arcs denote correlations. (b) The MEcom model for six variables measuring two traits by three methods. The variance of the error terms and the correlations between the latent variables are not restricted. However, for simplicity they are not depicted in this diagram.

A *second assumption* is that the errors do not correlate with each other:

$$Cov(\varepsilon_{tj}, \varepsilon_{sk}) = 0 \quad (t, j) \neq (s, k); \quad t, s = 1, 2; \quad j, k = 1, 2, \tag{26}$$

where t and s denote traits and j and k methods. Following from the definitions of true-score and error variables (see, e.g., Novick, 1966; Steyer, 1989, 2001; Zimmerman, 1975), the true-score variables τ_t and the method factor M_j are uncorrelated with the error variables:

$$Cov(\tau_t, \varepsilon_{sj}) = Cov(M_j, \varepsilon_{sk}) = 0 \quad t, s = 1, 2; \quad j, k = 1, 2. \tag{27}$$

The MEcom model may also be applied to data measuring more than two traits by more than two methods. The path diagram for a MEcom model for six variables measuring *two traits* by *three methods* is depicted in Figure 2(b). Note that for simplicity the error terms and the covariances between the latent variables are not represented. The number of method factors modeled in the MEcom model is one less than the number of methods used. Hence, there are just two method factors in Figure 2(b). For each manifest variable, except for the first one, there is a method factor on which its loading is 1. This method factor represents the person-specific effects of the corresponding method as compared with all methods considered. All manifest variables measured by method 1 have a loading of -1 on all method factors. The method-effect variable pertaining to method 1 is the negative sum of the method-effect variables of all method factors, that is, $M_1 = -(M_2 + \ldots + M_m)$ (see Equation 13). The general measurement

equations for a MEcom model for variables measuring r traits by m methods are

$$Y_{tj} = \begin{cases} \tau_t - M_2 - \ldots - M_m + \varepsilon_{t1} & \text{for } j = 1 \\ \tau_t + M_j + \varepsilon_{tj} & \text{for } j \neq 1, \end{cases} \tag{28}$$

with $t = 1, \ldots, r; j = 1, \ldots, m$.

In the MEcom model presented so far it is assumed that the method effects are the same for each trait. This assumption may not be feasible in all applications. In some applications the effect of a method may be larger for some traits than for others. Parents may overestimate the intelligence of their children more than they overestimate their children's talent for music. In order to allow the magnitude of the method effects to differ between the different traits, the model may be specified as follows:

$$Y_{tj} = \begin{cases} \tau_t - M_2 - \ldots - M_m + \varepsilon_{tj} & \text{for } t = 1, j = 1 \\ \tau_t + M_j + \varepsilon_{tj} & \text{for } t = 1, j \neq 1 \\ \tau_t - \lambda_{t2} \cdot M_2 - \ldots - \lambda_{tm} \cdot M_m + \varepsilon_{tj} & \text{for } t \neq 1, j = 1 \\ \mu_{tj} + \tau_t + \lambda_{tj} \cdot M_j + \varepsilon_{tj} & \text{for } t \neq 1, j \neq 1, \end{cases} \tag{29}$$

with $t = 1, \ldots, r; j = 1, \ldots, m$. A path diagram of such a model for the example in Application 1 measuring well-being with positively and negatively worded items at four measurement occasions (with well-being at the four occasions representing the four "traits") is depicted in Figure 3(b). The values on the method factors (M_j) represent the difference between the true score of the manifest variable measured with the respective method and *the first trait*. The method effects for the other traits (M_{tj}) are represented by $M_{tj} = \lambda_{tj} \cdot M_j$. The factor loadings λ_{tj} capture the magnitude of the method effects for each trait. Factor loadings greater than one indicate a larger method effect for the respective trait as compared with the magnitude of the method effects for the first trait. Factor loadings smaller than one indicate smaller method effects than the method effects for the first trait. Note that although the magnitude of the method effects may be different for each trait, it is still assumed that the method effects are unidimensional for all traits. Thus, it is assumed that participants having high method effects on one trait also have high method effects on the other traits and vice versa.

Extensions of the Model

So far we have assumed that the method effects are unidimensional for each trait. This is a very strong assumption that does not hold in all applications. The method effect for parent report (vs. self report) is probably different for

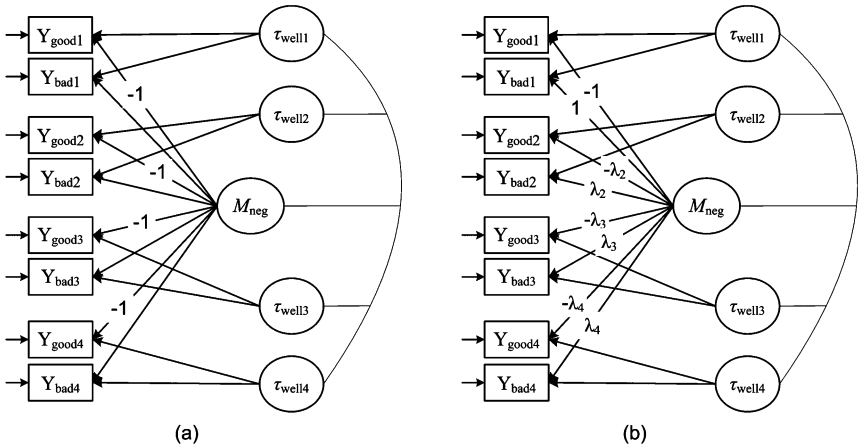

FIGURE 3 MEcom model for eight variables measuring well-being with positively and negatively worded items at four measurement occasions (a) assuming equal method effects for each occasion and (b) allowing for a different magnitude of the method effect for each occasion. All factor loadings that are not displayed are set to one. There are no restrictions on the means and correlations of the latent variables.

the measurement of intelligence than for the measurement of tidiness. Parents who overestimate the intelligence of their children do not necessarily also overestimate their children's tidiness. In order to allow for method effects that are not unidimensional or even the same for each trait, but *trait-specific*, the MEcom model may be extended using multiple indicators for each trait-method unit. For the measurement of calmness and alertness by positively and negatively worded items, for example, two test halves per trait and method (i.e., per trait-method unit) may be used as indicators i (see Application 2). Different measurement models may be assumed for the indicators of the same trait-method unit. For tau-equivalent indicators within each trait-method unit the resulting model may be formalized as

$$Y_{tji} = \begin{cases} \tau_t - M_{t2} - \ldots - M_{tm} + \varepsilon_{tji} & \text{for } j = 1 \\ \tau_t + M_{tj} + \varepsilon_{tji} & \text{for } j \neq 1, \end{cases} \tag{30}$$

with $t = 1, \ldots, r; j = 1, \ldots, m; i = 1, \ldots, p$. Figure 4(a) shows this model for the example presented in Application 2, measuring alertness and calmness with positively and negatively worded items. Note that—in contrast to the model presented in Equations 28 and 29—there is a separate method factor for each trait. This method factor does not represent the method effects *for all traits* (as in the model in Equation 28) but the method effects *for the respective trait*. Thus,

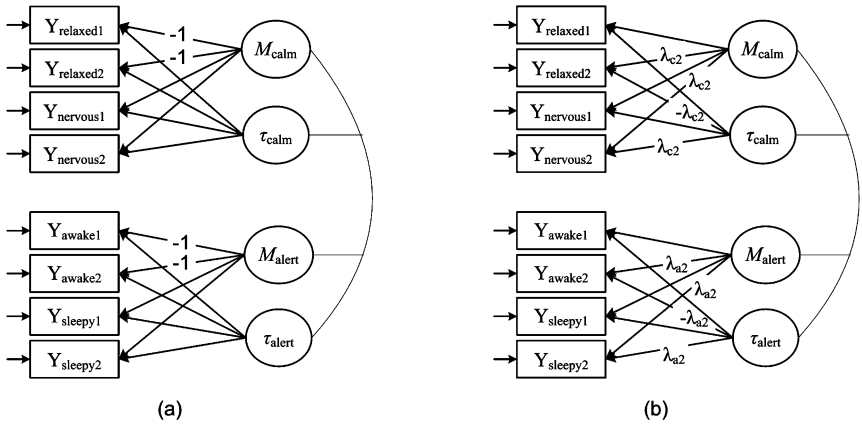

FIGURE 4 MEcom model for eight variables measuring calmness and alertness with positively and negatively worded items (a) assuming tau-equivalent indicators and (b) assuming tau-congeneric indicators for each trait-method unit. All factor loadings that are not displayed in Figure (a) are set to one. All factors that are not displayed in Figure (b) are as in the model displayed in Figure (a). There are no restrictions on the means and correlations of the latent variables.

method effects are not assumed to be unidimensional but may be *trait-specific*. For tau-congeneric indicators within each trait-method unit the model equations are

$$Y_{tji} = \begin{cases} \tau_t - M_{t2} - \ldots - M_{tm} + \varepsilon_{tji} & \text{for } j = 1, i = 1 \\ \mu_{ti} + \tau_t - \lambda_{ti} \cdot M_{t2} - \ldots - \lambda_{ti} \cdot M_{tm} + \varepsilon_{tji} & \text{for } j = 1, i \neq 1 \\ \tau_t + M_{tj} + \varepsilon_{tji} & \text{for } j \neq 1, i = 1 \\ \mu_{ti} + \lambda_{ti} \cdot \tau_t + \lambda_{ti} \cdot M_{tj} + \varepsilon_{tji} & \text{for } j \neq 1, i \neq 1, \end{cases} \quad (31)$$

with $t = 1, \ldots, r; j = 1, \ldots, m; i = 1, \ldots, p$. The respective model for Application 2 is depicted in Figure 4(b). Note that we propose measurement invariance (see, e.g., Borsboom, 2006) across the different methods. Within the same trait the factor loadings and the intercept for the indicators are set equal across the different methods. Thus, the different methods need to be measured on the same scale. If the same items for the self rating and the parent rating are used, this assumption is quite plausible. If, however, nervousness is measured using a questionnaire and by measuring electrodermal activity, the different methods are not on the same scale and measurement invariance will not hold. In this case a comparison of the different methods using the ME models is not meaningful because the difference between the measurements is not only due to method effects but also due to differences in scaling. Also note that the

scores on the trait and the method factors are on the scale metric of the first indicator. There is a method factor for each trait, thus allowing for different method effects for different traits. Again, there are no restrictions on the means and covariance between the latent variables and it is assumed that the error terms do not correlate with each other.

APPLICATION

In this section, the application of the MEcom model is illustrated by two examples examining the effect of item wording on the measurement of mood states. In the first application the simple MEcom model is illustrated on data measuring well-being by positively and negatively worded items at different measurement occasions. Note that this is a longitudinal study where well-being at each measurement occasion represents a separate "trait." In the second example the extended MEcom model is demonstrated on cross-sectional data measuring the two traits alertness and calmness with positively and negatively worded items, allowing for trait-specific method effects. The research questions investigated in both applications are (a) Is there a difference in the average mood state rated by different item wordings? (b) Do the participants differ considerably in the size of the method effect due to item wording, that is, are the different item wordings unidimensional measures of the same construct? (c) Is the size of the method effect correlated with the level of the mood state?

Data

The data (Steyer, Schwenkmezger, Notz, & Eid, 1997) used for this analysis stem from a validation study of a questionnaire measuring mood states. The sample consists of 291 women and 212 men with a mean age of 31.2 years (range from 17 to 77 years). For a more detailed description of the sample see Steyer, Schwenkmezger, Eid, and Notz (1991). Due to missing values six cases were deleted listwise for the first analysis and one case for the second analysis. The participants were asked to complete some questionnaires on each of four occasions with an interval of 3 weeks between measurements. Among others, the participants rated their well-being, alertness, and calmness on items from the multidimensional mood state questionnaire (MDBF; Steyer, Schwenkmezger, Notz, & Eid, 1997). The items comprising these scales are statements containing adjectives describing the mood states, which are either positively (e.g., "I feel good") or negatively (e.g., "I feel bad") worded. The participants could agree or disagree with these items on a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much*). For further analyses, the scores of the negatively worded items are reversed.

For the first application scale scores on the well-being variables are yielded for both sets of items (positive-item scale and negative-item scale) by calculating the mean of all positively worded items and of all negatively worded items within each occasion. Altogether eight manifest variables (two scales at four occasions) are analyzed; for each measurement occasion one scale score of positively ($Y_{good1}, Y_{good2}, Y_{good3}$, and Y_{good4}) and one scale score of negatively (Y_{bad1}, Y_{bad2}, Y_{bad3}, and Y_{bad4}) worded items. For the second application only the data of the third measurement occasion were used and items were aggregated to two test halves per trait-method unit by splitting the items for each trait and method combination in two sets and calculating the mean. Thus, there are eight manifest variables: two indicators measuring alertness (alert) with positively worded items (Y_{awake1}, Y_{awake2}), two with negatively worded items (Y_{tired1}, Y_{tired2}), two indicators measuring calmness (calm) with positively worded items ($Y_{relaxed1}, Y_{relaxed2}$), and two with negatively worded items ($Y_{nervous1}, Y_{nervous2}$). In both examples the scores on the manifest variables vary between 1 and 5, with high scores indicating high well-being, alertness, or calmness.

All analyses were done using M*plus* 5 (Muthén & Muthén, 1998–2007). The distributions of all variables differ significantly from a normal distribution with the negatively worded items being especially skewed. Therefore, the MLM estimator, using the Satorra-Bentler chi-square, was used for estimation.[3]

Analyses

Application 1: The effect of item wording on the measurement of well-being in longitudinal data. The MEcom model was fitted twice to the data, first assuming that the method effects are the same across the four well-being measures (see Equation 28 and Figure 3(a)) and second allowing for a different magnitude of the method effects across the four well-being measures (see Equation 29 and Figure 3(b)). The M*plus* syntax for the model in Figure 3(a) can be found in Appendix A. Both models fit the data well (first model: $\chi^2(16) = 21.301$, $p = .1672$, $RMSEA = 0.026$; second model: $\chi^2(13) = 18.919$, $p = .1256$, $RMSEA = 0.030$). Because the model fit of the second model was not significantly better than the model fit of the first model ($\chi^2_{diff}(3) = 2.22$, $p > .05$), the first, more restrictive model was used for the interpretation of the parameters. The magnitude of the method effects is the same across the four occasions. The estimated variances, covariances, correlations, and means of the MEcom model in Figure 3(a) are shown in Table 1.

The trait-factors τ_{well1} to τ_{well4} in the model in Figure 3(a) represent the mean of the true well-being ratings over both item wordings, whereas the method

[3]See recommendations of Olsson, Foss, Troye, and Howell (2000).

TABLE 1
Estimates of the MEcom Model in Application 1: Variances (Diagonal), Covariances (Lower Triangle), Correlations (Upper Triangle), and Means (Last Column) of the Latent Variables (*indicates $p < 0.05$)

Variables	τ_{well1}	τ_{well2}	τ_{well3}	τ_{well4}	M_{neg}	M
τ_{well1}	0.638*	0.294*	0.303*	0.291*	0.155*	3.696*
τ_{well2}	0.204*	0.757*	0.347*	0.410*	0.207*	3.747*
τ_{well3}	0.207*	0.258*	0.732*	0.447*	0.188*	3.711*
τ_{well4}	0.183*	0.282*	0.302*	0.623*	0.161*	3.809*
M_{neg}	0.024*	0.035*	0.031*	0.025*	0.038*	0.366*

factors represent the effects of the negatively worded items with respect to the common trait factor. On average the participants feel quite well (estimated means on the trait factors range from 3.696 to 3.809). When asked with negatively worded items the participants rate their well-being on average 0.366 scale scores higher and when asked with positively worded items 0.366 scale scores lower than the average trait score (see estimated mean of the method factor). The difference between the answers to positively and negatively worded items is on average $0.366 + 0.366 = 0.732$ scale points. Thus, there is a considerable mean effect of item wording. The estimated variances of the trait factors range from $\widehat{\sigma}^2(\tau_{well4}) = 0.623$ to $\widehat{\sigma}^2(\tau_{well2}) = 0.757$. The participants differ in their well-being ratings on the common well-being scores. The variance of the method factor, $\widehat{\sigma}^2(ME_{neg}) = 0.038$, $p < .001$, is substantive compared with the variance of the trait factors. Note that the method factor "only" represents the difference between the true-score variable of the negative scale and the *average of the true-score variables of both methods*. The difference in the true-score variables *between both methods* is twice as large and, thus, the respective variance is four times as large (i.e., 0.152). The standard deviation of the method effects in the model in Figure 3(a) is $\widehat{\sigma}(ME_{neg}) = 0.195$, indicating—when related to the 5-point Likert scale—substantive interindividual differences in the size of the method effects. On average the individual method effects are about 0.195 scale points higher or lower than the average method effect.[4] Related to the 5-point Likert scale and the standard deviation of the trait factors, the variation of the method effects is substantive. Some participants differentiate stronger between

[4]As a comparison, when defining the method effects as in the MEref model as the difference between the true scores of the positively and the negatively worded items, the standard deviation of the method effects would be 0.390. Thus on average the individual method effects are about 0.390 scale points higher or lower than the average method effect. Note that the MEcom is equivalent to the MEref model and the difference in parameter estimates for the method factors is just due to the different definitions of the method effects, that is, due to the different comparison value.

the different item wordings and others not at all. The different item wordings are, thus, not unidimensional but measure different latent variables. The method factors correlate significantly with the trait factors (between $\widehat{\rho}(\tau_{well1}, ME_{neg}) = 0.155$ and $\widehat{\rho}(\tau_{well2}, ME_{neg}) = 0.207$). The higher the participants rate their well-being, the more they differ in the answer to positively and negatively worded items. These correlations are, however, small.

Application 2: The effect of item wording on the measurement of alertness and calmness in cross-sectional data. In the second application the effect of item wording is investigated on the measurement of the two constructs alertness and calmness in cross-sectional data. Because it is not assumed that the effect of the item wording is the same for calmness and alertness, the extended MEcom model, which allows for trait-specific method effects, is applied. First, a model assuming tau-equivalent indicator variables within each trait-method-unit (according to Equation 30) was fitted to the data. The respective model is depicted in Figure 4(a). Two indicators are used to measure each trait-method unit. The trait factors τ_{calm} and τ_{alert} represent the means of the true-score variables of all four manifest variables measuring calmness and alertness, respectively. The method factors represent the effect of the negative item wording with respect to the common trait factor. Note that the effect of the positive item wording is just the negation of the effect of the negative item wording. The model fits the data well ($\chi^2(22) = 39.073, p = .0138, RMSEA = 0.039$). Second, a model assuming tau-congeneric indicator variables within each trait-method unit (according to Equation 31) was applied. The respective model is depicted in Figure 4(b) and the M*plus*-Syntax is given in Appendix B.[5] The model fits the data well ($\chi^2(16) = 23.067, p = .1120, RMSEA = 0.030$) and significantly better than the model assuming tau-equivalent indicators ($\chi^2(6) = 16.222, p < .05$). Thus, for the interpretation of the parameters the model in Figure 4(b) is used. The factor loadings of the second indicator do not differ much from one ($\lambda = 0.1031$ and $\lambda = 0.973$ for calmness and altertness, respectively) and the intercepts of the second indicator differed slightly for alertness ($\mu = -0.124$) but not much for calmness ($\mu = 0.066$). Thus, the rejection of the tau-equivalent model seems to be mainly due to the additive constant present for the measurement of alertness. The estimated variances, covariances, correlations, and means of the model are shown in Table 2.

On average, the participants rate themselves as being slightly calm, $\widehat{E}(\tau_{calm}) = 3.442$, and moderate alert, $\widehat{E}(\tau_{alert}) = 3.263$. These ratings vary between participants with even larger interindividual differences for alertness,

[5]It might be necessary to set starting values for the free factor loadings in order to get proper solutions. This is also shown in the M*plus*-Syntax.

TABLE 2
Estimates of the MEcom Model in Application 2: Variances (Diagonal), Covariances (Lower Triangle), Correlations (Upper Triangle), and Means (Last Column) of the Latent Variables (*indicates $p < 0.05$)

Variables	τ_{calm}	τ_{alert}	M_{calm}	M_{alert}	M
τ_{calm}	0.671*	0.457*	0.113	0.007	3.442*
τ_{alert}	0.375*	0.984*	−0.100	0.499*	3.263*
M_{calm}	0.025	−0.026	0.069*	0.348*	0.266*
M_{alert}	0.001	0.117*	0.022*	0.056*	0.226*

$\widehat{\sigma}^2(\tau_{alert}) = 0.984$, than for calmness, $\widehat{\sigma}^2(\tau_{calm}) = 0.671$. Calmness and alertness correlate moderate, $\widehat{\rho}(\tau_{calm}, \tau_{tired}) = 0.457$, with each other. There is a larger average method effect due to item wording for the measurement of calmness, $\widehat{E}(M_{calm}) = 0.266$, than for alertness, $\widehat{E}(M_{alert}) = 0.226$. Participants rate their calmness and alertness on average higher on negatively worded items than on the average over both item wordings. The participants differ in the size of the method effect, $\widehat{\sigma}^2(M_{calm}) = 0.069$ and $\widehat{\sigma}^2(M_{alert}) = 0.056$, with some participants differentiating more between the answers to different item wordings than others.[6] The respective standard deviations are $\widehat{\sigma}(M_{calm}) = 0.263$ and $\widehat{\sigma}(M_{alert}) = 0.237$. On average the individual method effects are about 0.263 and 0.237 for calmness and alertness, respectively, higher or lower than the average method effect.[7] Relating these values to the 5-point-Likert scale the constructs are measured on and to the standard deviation of the trait factors, the interindividual differences in the method effects are substantive. There is no significant correlation between the trait factor calmness and its method factor. There is, however, a significant correlation between the trait factor alertness and its method factor, $\widehat{\rho}(\tau_{alert}, M_{alert}) = 0.499$. Alert participants differentiate more between the item wordings for measuring alertness than less alert participants. The trait-specific method factors correlate slightly with each other, $\widehat{\rho}(M_{calm}, M_{alert}) = 0.3481$, indicating that the method effects are not identical for both traits but trait-specific. However, participants who differentiate more between item wordings for the measurement of calmness also tend to differentiate more between item wordings for the measurement of alertness.

[6]Note again that the method factors only represent the difference between the answers on the negative scale and the trait scores. The difference between the answers on the negative and on the positive scale is twice as large and thus the variance of this difference four times as large as the method variance in the MEcom model (i.e., 0.267 and 0.224, respectively).

[7]The standard deviations of the respective method effects in the MEref model would be 0.526 and 0.474 for calmness and alertness, respectively.

Summarizing the results of both applications, there is a considerable effect of item wording for the measurement of all three mood states. Participants rate their mood states on average higher when asked with negatively as compared with positively worded items. The participants differ considerably in the size of the method effect, with participants with high well-being and high alertness showing larger method effects, that is, differences between their answers to positively and negatively worded items. Hence, the use of different item wordings does not result in a unidimensional measure of mood states but questions the construct validity of the questionnaire. Researchers should thus be careful with the dimensionality of a questionnaire when using different item wordings.

DISCUSSION

Advantages and Limitations of the MEcom Model

We here introduced the MEcom model, which combines many advantages of previous MTMM models, thus being an excellent model for the investigation of usual MTMM research questions. The MEcom model does not suffer from estimation or identification problems. The method effects are not assumed to be independent of each other but may correlate and may be explained by other variables. The trait factors are defined as common to all methods, which is what we are interested in in many applications. The method effects are clearly defined. They are modeled as effects and have all properties of an effect. Furthermore, the means of the method factors as well as the correlation of the method factors with the traits may be estimated, thus, also allowing the investigation of new research questions, as for example: Do parents on average overestimate the competencies of their children compared with the self and teacher ratings? Is the amount of the overestimation by the parents related to the competency level of the child, that is, do parents overestimate the competencies of their children more for low competencies of the children than for higher ones? Or, do participants rate their mood on average higher on negatively than on positively worded items? Is the size of the effect of item wording related to the level of the mood state? Because for these research questions the estimation of mean method effects and the correlation of method effects with the trait is necessary, these questions can only be investigated using the MEref or MEcom model.

Another strength of the MEref and MEcom model is the clear definition of the latent variables. They are either true-score variables, true-score variable differences, or means of true-score variables. Such a clear definition of the latent variables is necessary for a substantive interpretation of the results and enables one to distinguish trait and method factors in different MTMM models. In the MEref model a method effect is defined by comparing one method with

a reference method whereas in the MEcom model a method effect is defined by comparing one method with an aggregate of all methods. As a result of these different definitions the parameter estimates of the method factor differ between the two models. Although the two ME models are equivalent, when using only two methods the variance of the method factor in the MEref model, for example, is four times as large as the variance of the method factor in the respective MEcom model. Without knowing the definition of the method factor a meaningful interpretation of the size of the method effect is not possible. Such a clear definition of the latent variables is also present in the CTC(M-1) model but not in traditional models. According to the authors (e.g., Marsh, 1989) trait factors in traditional models represent trait variance and method factors represent method variance. However, in these models it is not clear what the variables are, the variance of which we are looking at. Thus, we cannot decide whether these trait and method factors are what we aim to model.

There are also some restrictions to the application of the MEref and MEcom (ME) model. When comparing different methods the scaling of the measurements is important. In order to get interpretable estimates of the method effects, the different methods should be measured on the same scale. In the example of positively and negatively worded items the participants respond to both methods on a 5-point Likert scale. It is thus quite reasonable to assume that both methods are measured on the same scale and that differences between the two measurements are due to method effects. The same applies when regarding different raters who answer the same set of questions on the same response scale. If, however, a self rating of prejudices on a 5-point Likert scale and the implicit association test measuring prejudices by comparing reaction times are used, the measurements are on different scales. A direct comparison of the different measures, for example, the difference between a prejudice rating of 3 and a reaction time difference of 150 ms, is then not meaningful. Of course, even if the values, the mean, and the variances of the method effects for measures on different scales may not be meaningfully interpreted, correlations between the measures using different methods or correlation of method effects with other variables may. This problem of noncomparability of measures using different scales is not prevalent in the CTC(M-1) model because there is not a difference between the different measures but the difference between true score and expected true score is regarded.

Also the choice of the items in each scale should carefully be regarded. The items of the well-being scale in Application 1 are antonyms, thus assuring that a difference between the measurements is due to the direction of item wording and not due to the choice of items in each scale. In contrast the items used in the calmness and alertness scale in Application 2 are not direct antonyms. This may question the validity of the analysis. If the items of the negative scale are more extremely worded as compared with the items of the positive scale (independent

of the direction of item wording) the method effects estimated in the MEcom model may not only be due to effects of item wording but also due to effects of item phrasing. Thus, a researcher should aim at using scales that are parallel in order to draw valid conclusions about method effects.

One limitation of the ME models is that the variance of the manifest variables cannot be decomposed into additive components representing trait-, method-, and error-specific variance because the trait and the method factors may correlate with each other. However, convergent validity can still be evaluated by interpreting the size of the standard deviation of the method effects relative to the scale of measurement and relative to the standard deviation of the trait factors. When the standard deviation of the method factor is small, convergent validity is assured. It is of note that for the evaluation of the size of the method effect, the type of ME model and, thus, the way the method effects are defined needs to be considered. Although the MEref and MEcom model are equivalent, that is, they have the same model fit, analyzing data with both models results in different parameter estimates for the method effects. Using only two different methods the mean of the method effects in the MEref model is twice as large as the mean of the method effects in the MEcom model and the variance of the method effects is four times as large in the MEref as compared with the MEcom model. This difference in parameter estimates is due to the fact that in the MEref model method effects are defined by comparing the true score of one method with *the true score of the other method* whereas in the MEcom model the same true score is compared with *the mean of the true scores of both methods*. Thus, even if the parameters of the method effects are much smaller in the MEcom than in the MEref model, the method effect itself is the same. For the interpretation of the method effects, the definition of the method effects should, thus, be taken into account.

Choice of a Model

In both models, the MEref and the MEcom model, a method effect of a person is defined as an *effect*. In the MEref model we regard the effect of using method j as compared with method 1, whereas in the MEcom model a method effect is defined by comparing method j to all methods. Both definitions of a method effect are legitimate and lead to equivalent models. Defining method effects as it is done in the MEref model leads to method-specific traits, whereas the definition of a method effect in the MEcom model leads to traits common to all methods. The MEref model may thus be convenient when there is a clear reference method, for example, when subordinate and superior ratings are compared with the self rating and we are interested in the self rating as the trait factor. There are, however, applications where there is no clear reference method. When regarding oral and written exams for the measurement of school

performance or positively and negatively worded items for the measurement of well-being, we are not concerned with a reference-method specific trait, for example, performance measured by oral (or by written) exams but with a trait that is an aggregate over all measures. In these cases, the MEcom model is a good alternative to the MEref model. Of course, in order to get more information a researcher may also apply both the MEref model (using different reference methods) and the MEcom model.

Further Research

The ME models, presented here, are specified for scale scores. Of course, the model may also be extended to model on item level. This may be especially important because method effects may already occur on item level. Therefore, one future research task is to extend the ME models to item level.

Although method effects are often modeled in applications, not many studies deal with their explanation. Steyer and Riedl (2004) and Vautier and Pohl (2008) investigated whether the multidimensionality found when using positively and negatively worded items is due to effects of item wording or due to the existence of two substantially different constructs. They developed different models for the investigation of this research question and the results of their data analyses supported the hypothesis of the bipolarity of the construct and the existence of temporally stable method effects due to item wording. Horan et al. (2003), for example, investigated whether the method effects due to positively and negatively worded items in the Rosenberg self-esteem scale are a methodological artifact or due to response style by estimating the correlation between the method effects for different traits and the stability of method effects across time. Bernhardson (1970) studied social desirability as a confounding variable in the reversed item approach in the Minnesota Multiphasic Personality Inventory. Further research should focus not only on accounting for method effects but also on explaining them.

In order to investigate the validity of different methods and to decide which method to choose for the measurement of a construct, researchers need to investigate the relation of the constructs measured by different methods with external criteria. Until now inconclusive results have been obtained for the validity of the scales. Lai (1994), for example, investigated the validity of the positive and the negative scale of the Life Orientation Test by looking at the predictive power of both scales for symptom levels. Only the complete test and the positive scale predicted symptom levels. It was concluded that positive items validly measure optimism. Different results have been obtained by Burke (1999) and Greenberger et al. (2003). Burke investigated the validity of the Job Diagnostic Survey. She estimated the correlation of the questionnaire consisting of both positively and negatively worded items and of the questionnaire consisting only of the

positively worded items with work outcomes. From the results she concluded that both questionnaires have the same validity. Greenberger et al. came to similar results for the measurement of self-esteem. They found that there is hardly any difference in the validity, that is, in the relation of the positive, the negative, and the mixed scale of self-esteem items with depression and self-deception. Further research is needed concerning the validity of other methods for the measurement of different constructs. The Method Effect models may well be used for such investigations because they allow different definitions of the trait factors.

REFERENCES

Bernhardson, C. S. (1970). Social desirability as a confounding variable in the reversed item approach to studying acquiescence in the MMPI. *Canadian Journal of Behavioural Science, 2*, 148–156.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*, 176–181.

Burke, B. G. (1999). Item reversals and response validity in the Job Diagnostic Survey. *Psychological Reports, 85,* 213–219.

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

Clausen, M. (2002). *Unterrichtsqualitä: Eine Frage der Perspektive?* [Teaching quality: A question of the perspective?]. Münster, Germany: Waxmann.

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.

Conway, J. M., & Huffcuff, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer and self-ratings. *Human Performance, 10,* 331–360.

Conway, J. M., Lievens, F., Scullen, S. E., & Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling, 11,* 535–559.

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

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability? *Journal of Personality and Social Psychology, 50,* 1199–1210.

Gignac, G. E. (2006). Evaluating subtest 'g' saturation levels via the single trait-correlated uniqueness (STCU) SEM approach: Evidence in favor of crystallized subtests as the best indicators of 'g'. *Intelligence, 34,* 29–46.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences, 35,* 1241–1254.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10,* 435–455.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 409–426.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. Atkinson, D. Krantz, R. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: W. H. Freeman.

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

Lai, J. C. L. (1994). Differential predictive power of the positively versus the negatively worded items of the Life Orientation Test. *Psychological Reports, 75,* 1507–1515.

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

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810–819.

Marsh, H. W., & Craven, R. G. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teacher, mothers, and fathers. *Journal of Educational Psychology, 83,* 393–404.

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

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Soerbom (Eds.), *Structural equation modeling: Present and future*. Lincolnwood, IL: Scientific Software International, Inc.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

Motl, R. W., & DiStefano, C. (2002). Longitudinal Invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling, 9,* 562–578.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37,* 687–702.

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

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

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7,* 557–595.

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A, 171,* 1–23.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3–30.

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

Steyer, R. (2001). Classical test theory. In C. Ragin & T. Cook (Eds.), *International encyclopedia of the social and behavioural sciences: Logic of inquiry and research design* (pp. 481–520). Oxford, UK: Elsevier Science.

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., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (in press). *Probability and causality.* New York: Springer. (Manuscript available from http://www.causal-effects.de/)

Steyer, R., & Riedl, K. (2004). Is it possible to feel good and bad at the same time? New evidence on the bipolarity of mood-state dimensions. In K. V. Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments on structural equation modeling: Theory and applications* (pp. 197–220). Amsterdam: Kluwer Academic.

Steyer, R., Schwenkmezger, P., Eid, M., & Notz, P. (1991). *Befindlichkeitsmessung und Latent-State-Trait-Modelle* [Measurement of mood and latent-state-trait-models] (Tech. Rep.). Arbeitsbericht zum DFG-Projekt STE 411/3-1. Trier, Germany: University Trier.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)* [The multidimensional mood state questionnaire]. Göttingen, Germany: Hogrefe.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (2004). *Entwicklung des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). Primärdatensatz* [Development of the multidimensional mood state questionnaire. Data set]. Trier, Germany: Psychologisches Datenarchiv PsychData des Zentrums für Psychologische Information und Dokumentation ZPID.

Vautier, S., & Pohl, S. (2008). Do balanced scales assess bipolar constructs? The case of the STAI scales. *Psychological Assessment, 21,* 187–193.

Villar, P., Luengo, M. A., Gómez-Fraguela, J. A., & Romero, E. (2006). Assessment of validity of parenting constructs using the multitrait-multimethod model. *European Journal of Psychological Assessment, 22,* 59–68.

Watson, D., & Tellegen, A. (1999). Issues in the dimensionality structure of affect—effects of descriptors, measurement error, and response formats: Comments on Russell and Carroll (1999). *Psychological Bulletin, 125,* 601–610.

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

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

APPENDIX A

M*plus*-Syntax for the Method Effect Model with Common Trait Factors Model in the First Application

```
MODEL:
well1 BY good1@1 bad1@1;
well2 BY good2@1 bad2@1;
well3 BY good3@1 bad3@1;
well4 BY good4@1 bad4@1;
[well1*];
[well2*];
[well3*];
[well4*];
MEneg BY good1@-1 good2@-1 good3@-1 good4@-1 bad1@1 bad2@1 bad3@1 bad4@1;
MEneg*;
[MEneg*];
[good1@0];
[bad1@0];
[good2@0];
[bad2@0];
[good3@0];
[bad3@0];
[good4@0];
[bad4@0];
```

APPENDIX B

M*plus*-Syntax for the Method Effect Model with Common Trait Factors Model in the Second Application

```
MODEL:
calm BY relaxed1@1 nervous1@1;
calm BY relaxed2*(a);
calm BY nervous2*(a);
alert BY awake1@1 sleepy1@1;
alert BY awake2*(c);
alert BY sleepy2*(c);
Mcalm BY relaxed1@-1;
Mcalm BY nervous1@1;
Mcalm BY relaxed2*-1(a);
Mcalm BY nervous2*1(b);
Malert BY awake1@-1;
Malert BY sleepy1@1;
Malert BY awake2*-1(c);
Malert BY sleepy2*1(d);
[relaxed1@0];
[nervous1@0];
[relaxed2*](e);
[nervous1*](e);
[awake1@0];
[sleepy1@0];
[awake2*](f);
[sleepy2*](f);
calm*;
alert*;
Mcalm*;
Malert*;
[calm*];
[alert*];
[Mcalm*];
[Malert*];
MODEL CONSTRAINT:
0=a+b;
0=c+d;
```