Running head: COMMON TRAITS

Modeling common traits in multitrait-multimethod data

Steffi Pohl

Friedrich-Schiller-University Jena

Rolf Steyer

Friedrich-Schiller-University Jena

Address for correspondence:

Steffi Pohl

Department of Methodology and Evaluation Research

Institute of Psychology

Friedrich-Schiller-University Jena

Am Steiger 3, Haus 1

07743 Jena, Germany

Telephone: +49 3641 945237

E-mail: steffi.pohl@uni-jena.de

Abstract

Often, when constructs are measured by different methods, method effects occur. In recent models that account for method effects in multitrait-multimethod (MTMM) data, traits that are specific to a certain method are modeled. However, often we are not interested in a method specific trait but rather in a trait that is common to all methods. We here present the Method Effect model with common trait factors (MEcom), a new approach for modeling multitrait-multimethod data, where 'common' traits factors are modeled, and where the definition of the factors is derived from a theory. The common trait factors result through aggregating all variables measuring the same trait (by different methods) and are represented by the mean of the true outcomes. The definition of the method factors is derived from the theory of individual and average causal effects in the Neyman-Rubin tradition, allowing a causal interpretation of the method effects if alternative explanations may be ruled out. The model allows estimating (a) the mean method effects, (b) the correlation between method factors, and (c) the correlation between trait and method factors. The application of the model is demonstrated by an example investigating the effect of negative, as compared to positive, item formulation for the measurement of well-being.

Modeling common traits in multitrait-multimethod data

Introduction

Multitrait-multimethod research

In the social sciences, theories and hypotheses often involve constructs which may be measured by different methods, for example, self and peer rating for the measurement of personality constructs, questionnaire and behavior observation for the measurement of anxiety, or positively and negatively formulated items for the measurement of mood states. These different measures usually do not yield the same results. Instead, there are often systematic differences between different methods, and these method effects are person-specific, i.e. they differ between persons. In order to study the degree of convergence of the different measurement instruments, 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 (1959), convergent validity is achieved when different measurement methods yield similar results in measuring the same trait. Usually there is no method with undoubted validity. Therefore, in applied research, several methods are frequently used for the measurement of a construct. For example Holzbach (1978) and Mount (1984) used superior-, self- and peer ratings for the measurement of managerial performance. Similarly, Villar, Luengo, Gómez-Fraguela, and Romero (2006) measured parenting constructs using adolescent, mother and father reports, and McConnell and Leibold (2001) used implicit association tests, explicit measures and external ratings for the measurement of racial attitudes.

If systematic person-specific method effects exist, different measures are not unidimensional. Thus, in data analysis, ignoring systematic individual method effects often results in an unsatisfactory model fit. In many studies, method effects are regarded as nuisance effects that have to be taken into account in order to obtain a reasonable model fit (e.g. Cole, Martin, Powers,

& Truglio, 1996; Gignac, 2006; Motl & DiStefano, 2002). In other studies, method effects are not regarded as nuisance effects, but are the focus of research. For example, Conway and Huffcuff (1997) investigated the effect of the rater (self, supervisor, subordinate or peer) on ratings of job performance. Further studies have investigated the effect of item formulation (positive or negative) on ratings measuring different psychological constructs (e.g., Horan, DiStefano, & Motl, 2003; Marsh, 1996; Steyer & Riedl, 2004; Vautier, Steyer, Jmel, & Raufaste, 2005).

Since Campbell and Fiske (1959), many models for the analysis of multitrait-multimethod data have been proposed, most of them based on confirmatory factor analysis (CFA; e.g. Widaman, 1985; Marsh, 1989). More recently proposed are the correlated trait-correlated method minus one [CT-C(M-1)] model (Eid, 2000) and the Method Effect model with a reference method (MEref; Pohl, Steyer, & Kraus, 2008). In the CT-C(M-1) model, a reference method needs to be chosen and each trait factor then represents the true scores of the manifest variable, measuring this trait with the reference method. The method factors represent (a linear function of) residuals with respect to the regression of the true-score variable of a manifest variable measuring a trait by a certain method on the true-score variable of a manifest variable measuring the same trait by the reference method.

Since the Method Effect model with common trait factors (MEcom) introduced here is equivalent to the Method Effect model with a reference method (MEref), the MEref model is explained here in detail. Both in the new model and in the MEref model, method effects are defined as individual causal effects. Therefore, in the introduction we briefly describe the basic ideas of the theory of the individual and average causal effects with reference to Neyman (1923) and Rubin (1974, 1978). Then the MEref model is introduced and we conclude by discussing the advantages and limitations of the MEref model compared to traditional MTMM models.

The theory of individual causal effects

In previous MTMM models method effects are implicitly (correlated trait-correlated method model (Jöreskog, 1974; Widaman, 1985), correlated trait-correlated uniqueness model (Kenny, 1976; Marsh, 1989; Marsh & Craven, 1991)) or explicitly (CTC(M-1) model (Eid, 2000)) regarded as residuals. There are, however, substantive researchers in the field that do not regard method effects as residuals but rather as causal effects:

> "Methods are sets of causes and different sets (methods) contain different elements (causes). Causes as components of a method do not differ from causes that appear in psychological theories." (Schmitt, 2006, p. 24).

When subordinates rate the managerial performance of their superiors, not only the managerial performance of the superiors is measured, but also, for example, social desirability, need for approval or sympathy for the superior. Social desirability, need for approval and sympathy are rater-specific variables that are represented in the method effects and can be regarded as ‘causes’.

In the MEref and the MEcom models, method effects are regarded as causal effects. In this case we define a person-specific method effect as an individual causal effect (see Neyman, 1923; Rubin, 1974, 1978). The basic concepts of the theory of individual and average causal effects refer to the following random experiment, the *single-unit trial*: Draw a unit (here a person) u from a population of units, assign this unit to a treatment A ($X = 1$) or to a treatment B ($X = 0$), which may be a control condition, and then observe its outcome Y. It is presumed that there are two unit-specific conditional distributions of Y, one given treatment $X = 1$ and one given treatment $X = 0$. If there is a treatment effect, these intra-individual distributions will differ from each other. If there is no treatment effect for this unit, the two distributions will be the same. The intra-individual variation within treatment conditions may be due to measurement error, to fluctuations in alertness, tiredness or mood state, but also due to life events occurring between treatment and outcome if the outcome is assessed an adequate time span after treatment.

Considering the general case of J treatment conditions, we define the *true outcome* of unit u *under treatment* j, $\tau_j(u)$, as the expectation

$$\tau_j(u) \equiv E(Y \mid X = j, U = u) \tag{1}$$

of the unit-specific distribution given treatment j. Correspondingly, we call τ_j the *true-outcome variable*, referring to that random variable, the values of which are the true outcomes $\tau_j(u)$. [1]

Using the true outcomes defined above, several kinds of individual causal effects may be defined (see, for example Steyer, Partchev, Kröhne, Nagengast, & Fiege, 2008). First, we may define the individual causal effect of *treatment j compared to treatment k*,

$$ICE_{jk}(u) \equiv \tau_j(u) - \tau_k(u), \tag{2}$$

which is the difference between the true outcome of u under treatment j and under treatment k respectively. (Note that treatment k may also be an untreated control condition.) The corresponding average causal effect is defined by

$$ACE_{jk} \equiv E(ICE_{jk}), \tag{3}$$

the average of the individual causal effects, where $ICE_{jk} = \tau_j - \tau_k$ denotes the random variable, the values of which are the individual causal effects $ICE_{jk}(u)$.

Second, *a treatment j may be compared to all treatments*, i.e. an individual causal effect may also be defined as the difference between the individual true outcome $\tau_j(u)$ in treatment j, and the average of the individual true outcomes across all treatment conditions:

$$ICE_{j,total}(u) = \tau_j(u) - \tau_{total}(u) \qquad j = 1, 2, \ldots, J, \tag{4}$$

where

$$\tau_{total}(u) \equiv \frac{1}{J} \cdot \sum_{k=1}^{J} \tau_k(u) \qquad j = 1, 2, \ldots, J. \tag{5}$$

The corresponding average causal effect is defined as the average of these individual causal effects of treatment j across all units:

$$ACE_{j,total} = E(ICE_{j,total}). \tag{6}$$

The MEref and the MEcom model are specified using these different definitions of a causal effect. The MEref model uses the first definition of an individual causal effect, i.e. $ICE_{jk}(u)$, and the MEcom model introduced in this paper uses the second one, i.e. $ICE_{j,total}(u)$.

The Method Effect model with a reference method

Definition of the method effects.

In the MEref model, an individual or *unit-specific method effect*

$$M_{tj}(u) \equiv \tau_{tj}(u) - \tau_{t1}(u) \tag{7}$$

is defined as the difference between the true outcome of the measurement of trait t obtained under method j and the true outcome of the measurement of the same trait t obtained under method $k = 1$, the so-called reference method [2]. Choosing method $k = 1$ as the reference method does not restrict generality because each method can be chosen to be 'method 1'. Hence, the 'method' is considered a 'treatment' that may affect the measures (outcomes) observed.

According to this definition of a method effect, the *method-effect variable*

$$M_{tj} \equiv \tau_{tj} - \tau_{t1} \tag{8}$$

is defined by the difference between the corresponding true-outcome variables. In this equation, the difference $\tau_{tj} - \tau_{t1}$ between the two true-outcome variables represents the *systematic effects* of using method j instead of reference method 1, for measuring trait t. Thus, the manifest variable Y_{tj}, measuring trait t with method j, may be written as

$$\begin{aligned} Y_{tj} &= \tau_{tj} & & & + \quad & \varepsilon_{tj} \\ &= \tau_{t1} & + \quad & (\tau_{tj} - \tau_{t1}) & + \quad & \varepsilon_{tj} \\ &= \tau_{t1} & + \quad & M_{tj} & + \quad & \varepsilon_{tj}, \end{aligned} \tag{9}$$

where $\varepsilon_{tj} \equiv Y_{tj} - \tau_{tj}$. Note that these equations are *always true*.

Assumptions defining the MEref model.

Now let us consider several manifest variables Y_{tj}, measuring each of several traits by each of several methods, where t denotes the trait and j the method. In the simplest case, each of two traits is measured by each of two methods (see Figure 1). In this case, we may define two method-effect variables, $M_{12} \equiv \tau_{12} - \tau_{11}$ and $M_{22} \equiv \tau_{22} - \tau_{21}$, the first comparing method 2 to method 1 in measuring trait 1, and the second comparing method 2 to method 1 in measuring trait 2.

In order to define an identified model, some assumptions are introduced. The *first assumption* is

$$M_{12} = M_{22} \equiv M_2, \tag{10}$$

which states that the method-effect variables, and therefore also the individual, person-specific method effects, are the same for both traits. Therefore, we may drop the indices for the traits and call M_2 the *method factor*, the values of which represent the person-specific effects of using method 2 instead of method 1, in measuring each of the two traits.

If we investigate the effect of using a scale with positively formulated items (method 1), as compared to a scale with negatively formulated items (method 2), on the measurement of well-being (trait 1) and calmness (trait 2), this assumption implies that the difference between the true outcome variable of the positively formulated items and the true outcome variable of the negatively formulated items is the same for the measurement of well-being 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 seems quite plausible (see our application).

The respective structural equation model is depicted in Figure 1. The *trait factors* τ_{11} and τ_{21} in Figure 1 represent the *true-score variables* of the manifest variables, measuring the two traits *with the reference method* 1, e.g. with positively formulated items. Note that the trait factors will change (to some degree) if the reference method is changed. Hence, if we choose method 2, e.g. negatively formulated items, as the reference method, the trait factors will be τ_{12} and τ_{22} and these two trait factors will differ from the trait factors in Figure 1, i.e. τ_{11} and τ_{21}.

The values of the *method factor* M_2 in the ME model (see Figure 1) then represent the person-specific method effects of method 2, compared to the reference method 1, *for both traits*.

In the MEref model, the equations of the measurement model for the manifest variables are

$$Y_{tj} = \begin{cases} \tau_{t1} + \varepsilon_{t1} & \text{for } j = 1 \\ \tau_{t1} + M_j + \varepsilon_{tj} & \text{for } j \neq 1, \end{cases} \tag{11}$$

which shows that there is no method factor for the reference method.

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

$$Cov(\varepsilon_{tj}, \varepsilon_{sk}) = 0 \quad (t, j) \neq (s, k), \tag{12}$$

where t and s denote traits, whereas j and k denote methods. Furthermore, following from the definitions of true-score and error variables (e.g. Novick, 1966; Steyer, 1989, 2001; Zimmerman, 1975), the true-score variables τ_{t1}, and the method factor, M_j, are uncorrelated with the error variables:

$$Cov(\tau_{t1}, \varepsilon_{sj}) = Cov(M_2, \varepsilon_{sj}) = 0. \tag{13}$$

Note that there are no restrictions on the means, nor on the covariances of the latent variables. The method factors may have a mean different from zero; they may correlate with each other as well as with the trait factors.

Advantages of the MEref model.

The MEref model has many advantages. The model allows estimating (a) the average method effect, (b) the covariances of the method effects with the traits, (c) the covariances of the method effects among each other, and (d) the covariances of the trait factors among each other. Both the trait factors and the method factors have a clear substantive meaning. The trait factors are the true-score variables of the measures obtained with the reference method, and the method factors are the differences between the true-score variables of a variable measuring a trait with method j and a variable measuring the same trait with method 1. Since the method effects are defined according to a causal effect theory, they have the usual properties of an *effect*, i.e. an effect of zero indicates no difference and the size of the effect does not change with the direction of the comparison. In the example using self-report and subordinate measures, a method effect of zero indicates that the self and the subordinate rating do not differ in their true outcomes. Furthermore the size of the method effect is the same when comparing the self rating to the subordinate rating than when comparing the subordinate rating to the self rating. These properties do not hold for the CT-C(M-1) model (see Pohl et al., 2008, for a detailed explanation). Furthermore, the model fit of the MEref model is invariant with respect to the choice of the reference method. Further research regarded the latent change scores of a construct in longitudinal studies. Vautier and Pohl (in press) showed that even when method effects exist the latent change scores may be unidimensional if the method effects are temporally stable.

Limitations of the models for multitrait-multimethod designs

In traditional CFA-MTMM models (e.g. the correlated trait-correlated method model or the correlated trait-correlated uniqueness model; Marsh, 1989; Widaman, 1985), the trait factors are supposed to represent a common trait of all variables measuring the same trait. However, neither these common trait factors nor the method factors are well-defined in these models. Even using the joint and conditional distributions of all manifest variables involved (including the observational-unit variable U), we are not able to define the score of person u on such a trait factor. As shown above, this is different in the MEref model, where the latent variables are uniquely defined as true-outcome variables or as differences between true-outcome variables, and this is similar in the MEcom model and in the CT-C(M-1) model.

However, both the CT-C(M-1) model and the MEref model have a serious limitation. In both models, a reference method has to be chosen and the trait factor represents the true outcomes of the variable measuring this trait *by the reference method*. In some applications a reference-method specific trait may be appropriate, e.g. when measuring personality with self and peer ratings and when we are interested in the personality measured by the self rating as the trait. However, often we are 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 interested in well-being measured by positively (or negatively) formulated items. Instead, we usually aggregate these different measures in order to get a measure of well-being common to both methods. The same applies for the measurement of school performance by oral and by written exams, where we are usually interested in an aggregate of performance measures from both exam types. Besides interest in content, there is also a methodological advantage of modeling common traits. Epstein (1983, 1986) argued that aggregation over modes of measurement cancels out method-specificity and may increase the reliability and validity of the measurement. In the method effect model with common trait factors (MEcom model) we refine this idea combining both a clear-cut definition of the latent variables, which is derived from a theory, with the modeling of a common trait as an aggregate of variables, measuring the trait with different methods.

The Method Effect model with common trait factors

The MEcom model is a new parametrization of the MEref model, where each trait factor represents the mean of the true-score variables measuring this trait by different methods. We here start with the definition of method effects and then introduce the assumptions defining the model.

Definition of a method effect

In the theory of individual and average causal effects, a causal effect may be defined by comparing a treatment to all treatments. This definition may especially be of interest when there is no natural reference condition such as an untreated control.

The idea of comparing a treatment to all treatments is adapted to the ME model. Let Y_{11} and Y_{12} be two manifest variables measuring the same trait 1 with methods 1 and 2, respectively. Furthermore, let

$$\tau_1 \equiv \frac{\tau_{11} + \tau_{12}}{2} \tag{14}$$

be the mean of the true-score variables τ_{11} and τ_{12} of the two manifest variables Y_{11} and Y_{12}.

Hence τ_1 is no longer specific to one of the two methods. Instead, it represents a *common* trait. The *method-effect variables* may be defined accordingly as:

$$M_{11} \equiv \tau_{11} - \tau_1 \tag{15}$$

$$M_{12} \equiv \tau_{12} - \tau_1, \tag{16}$$

representing the difference between the true outcome variables of the manifest variables measuring a trait with a certain method and the common trait. If there are systematic person-specific method effects, the true-score variables of Y_{11} and Y_{12} will differ from each other and thus, also from the common trait factor τ_1. Note that the method-effect variables are still *trait-specific*, i.e. 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:

$$\begin{aligned} Y_{11} &= \tau_{11} + \varepsilon_{11} \\ &= \tau_1 + \tau_{11} - \tau_1 + \varepsilon_{11} \\ &= \tau_1 + M_{11} + \varepsilon_{11}, \end{aligned} \tag{17}$$

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

$$\begin{aligned} Y_{12} &= \tau_{12} + \varepsilon_{12} \\ &= \tau_1 + \tau_{12} - \tau_1 + \varepsilon_{12} \\ &= \tau_1 + M_{12} + \varepsilon_{12}, \end{aligned} \tag{18}$$

which is also always true.

The definition of the method-effect variables has some implications. Since 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 15 and 16), the following equations always hold:

$$\begin{aligned} M_{11} + M_{12} &= \tau_{11} - \tau_1 + \tau_{12} - \tau_1 \\ &= \tau_{11} + \tau_{12} - 2 \cdot \frac{\tau_{11} + \tau_{12}}{2} \\ &= \tau_{11} + \tau_{12} - (\tau_{11} + \tau_{12}) = 0, \end{aligned} \tag{19}$$

implying:

$$M_{11} = -M_{12}. \tag{20}$$

This leads to the following measurement equations:

$$Y_{11} = \tau_1 - M_{12} + \varepsilon_{11} \tag{21}$$

$$Y_{12} = \tau_1 + M_{12} + \varepsilon_{12}. \tag{22}$$

Considering m manifest variables Y_{1j}, measuring trait 1 by m methods, Equations 19 and 20 generalize as follows:

$$M_{11} + M_{12} + ... + M_{1m} = 0, \tag{23}$$

and thus

$$M_{11} = -M_{12} - ... - M_{1m}. \tag{24}$$

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 (MEcom), 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:

$$\tau_2 = \frac{\tau_{21} + \tau_{22}}{2} \tag{25}$$

$$M_{21} = \tau_{21} - \tau_2 \tag{26}$$

$$M_{22} = \tau_{22} - \tau_2 , \tag{27}$$

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

$$Y_{21} = \tau_2 - M_{22} + \varepsilon_{21} \tag{28}$$

$$Y_{22} = \tau_2 + M_{22} + \varepsilon_{22} \,, \tag{29}$$

which is always true.

Assumptions. 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:

$$M_{11} = M_{21} \equiv M_1, \tag{30}$$

saying that the method-effect variables M_{11} and M_{21} – therefore also the individual method effects – are the same for each of the two traits. This assumption implies the same as the respective assumption in the MEref model. For example, it is assumed that the difference between the true oucome of the negatively formulated items and the common trait is the same for the measurement of well-being as for the measurement of calmness. Since $M_{11} = -M_{12}$ and $M_{21} = -M_{22}$ (see Equation 20), the assumption in Equation 30 implies

$$M_{12} = M_{22} \equiv M_2. \tag{31}$$

and therefore

$$M_1 = -M_2. \tag{32}$$

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

$$Y_{11} = \tau_1 - M_2 + \varepsilon_{11} \tag{33}$$

$$Y_{12} = \tau_1 + M_2 + \varepsilon_{12} \tag{34}$$

$$Y_{21} = \tau_2 - M_2 + \varepsilon_{21} \tag{35}$$

$$Y_{22} = \tau_2 + M_2 + \varepsilon_{22}. \tag{36}$$

A path diagram representing these equations is depicted in Figure 2. The trait factors in Figure 2 represent the common trait variables τ_1 and τ_2. The *method factor* M_2 in Figure 2 represents the *method effects* of method 2 compared to the common trait *for both traits*. The *method effects* of method 1 compared to 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.

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

$$Cov(\varepsilon_{tj}, \varepsilon_{sk}) = 0 \quad (t, j) \neq (s, k); \quad t, s = 1, 2; \quad j, k = 1, 2, \tag{37}$$

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:

$$Cov(\tau_t, \varepsilon_{sj}) = Cov(M_j, \varepsilon_{sk}) = 0 \quad t, s = 1, 2; \quad j, k = 1, 2. \tag{38}$$

Note that the MEcom model is just a different parametrization of the MEref model. Thus, both models are equivalent, i.e. they imply the same mean and covariance structures and have the same model fit.

Generalization of the model

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 3. 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 3.

Each manifest variable, except for the first one, has a method factor on which its loading is 1. This method factor represents the person-specific effects of the corresponding method, as compared to all methods considered. All manifest variables measured by method 1 have the loading -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, i.e. $M_1 = -(M_2 + \ldots + M_m)$ (see Equation 24).

The general measurement equations for a MEcom model for variables measuring r traits by m methods are:

$$Y_{tj} = \begin{cases} \tau_t - M_2 - \ldots - M_m + \varepsilon_{t1} & \text{for } j = 1, \\ \tau_t + M_j + \varepsilon_{tj} & \text{for } j \neq 1, \end{cases} \qquad t = 1, \ldots, r, \quad j = 1, \ldots, m. \tag{39}$$

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 MEref and MEcom model is illustrated by an example examining the effect of item formulation on the measurement of mood states. The ME models are applied to data measuring well-being with positively and negatively formulated items (methods) at different measurement occasions. Note that this is a longitudinal study, where well-being at each measurement occasion represents a separate ‘trait’. Also note that since we only wish to illustrate the application of the model in this section, and not properly analyze the effect of item formulation, for simplicity, a small model with a simple structure is chosen.

Data

The data (Steyer, Schwenkmezger, Notz, & Eid, 2004) 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). For the analyses six cases with missing values were deleted listwise.

The participants were asked to complete some questionnaires on each of four occasions, with an interval of three weeks between measurements. Among others, the subjects rated their well-being on items from the multidimensional mood state questionnaire (MDBF; Steyer, Schwenkmezger, Notz, & Eid, 1997). The items comprising this scale are statements containing adjectives describing well-being which are either positively (e.g. 'I feel good') or negatively formulated (e.g. 'I feel bad'). The positive and negative formulated items are unsystematically distributed across the questionnaire. The subjects could agree or disagree to these items on a five point Likert scale ranging from (1) 'not at all' to (5) 'very much'. For further analyses, the scores of the negatively formulated items are reversed. Scale scores are yielded for both sets of items (positive-item scale and negative-item scale) by calculating the mean of all positive formulated items and of all negatively formulated 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 formulated items (Y_{bad1}, Y_{bad2}, Y_{bad3} and Y_{bad4}). The scale scores vary between 1 and 5, with high scores indicating high well-being. The distributions of all variables significantly differ from a normal distribution. The negatively formulated items are especially very skewed.

Analyses

All analyses were done using M*plus* 5 (Muthén & Muthén, 1998-2007). Because the variables are not normally distributed, the MLM estimator, using the Satorra-Bentler chi-square, was used for estimation (see recommendations of Olsson, Foss, Troye, & Howell, 2000). The MEref model was fitted twice to the data, using each kind of item formulation as a reference method. The MEref model with the positive item formulation as reference method is depicted in Figure 4(a). The models fit the data well ($\chi^2(16) = 21.301,\ p = 0.1672,\ RMSEA = 0.026$). The variances, covariances, correlations and means of both models are shown in Table 1.

On the negative-item scale the subjects rate their well-being on average higher (range from $\widehat{E}(\tau_{bad1}) = 4.062$ to $\widehat{E}(\tau_{bad4}) = 4.175$) than on the positive-item scale (range from $\widehat{E}(\tau_{good1}) = 3.329$ to $\widehat{E}(\tau_{good4}) = 3.443$). The subjects score on average $\widehat{E}(M_{neg}) = 0.733$ scale points higher on the well-being scale when rating their mood on negatively formulated items rather than on positively formulated ones.

The estimated correlations between the trait factors (range from $\widehat{\rho}(\tau_{good1}, \tau_{good2}) = 0.271$ to $\widehat{\rho}(\tau_{bad3}, \tau_{bad4}) = 0.515$) are higher for the negative-item than for the positive-item scale. Well-being as a state is not stable [3] across time, but more stable for the negative than for the positive-item scale.

The estimated variances of the trait factors are larger for the negative (range from $\widehat{\sigma}^2(\tau_{bad4}) = 0.711$ to $\widehat{\sigma}^2(\tau_{bad2}) = 0.865$) than for the positive-item scale (range from $\widehat{\sigma}^2(\tau_{good4}) = 0.612$ to $\widehat{\sigma}^2(\tau_{good2}) = 0.724$, $\chi^2_{diff}(4) = 19.454$, $p = 0.0006$ [4]). Hence, the persons differ more from each other in their well-being ratings on negatively than on positively formulated items. Compared to these variances, the variance of the method factor ($\widehat{\sigma}^2(M_{neg}) = \widehat{\sigma}^2(M_{pos}) = 0.152$, $p < 0.001$) is considerable, indicating that the subjects differ in the size of their individual method effects. The difference in the ratings on positively and negatively formulated items is higher for some subjects than for others. Thus, the negative-item and the positive-item scale are not unidimensional measurements of the same construct. Note that the method effects are due to method specificity of both item-scales.

The individual method effects do not correlate significantly with the true outcomes of the positive-item scale (between $\widehat{\rho}(\tau_{good2}, M_{neg}) = -0.017$ and $\widehat{\rho}(\tau_{good1}, M_{neg}) = -0.090$). However, they correlate significantly with the true outcomes of the negative-item scale (between $\widehat{\rho}(\tau_{bad1}, M_{pos}) = -0.374$ to $\widehat{\rho}(\tau_{bad2}, M_{pos}) = -0.403$). The higher the subjects score on the negative-item scale, the larger the method effect (meaning large negative scores on the factor M_{pos}). Thus, subjects with a high score on well-being measured by the negative-item scale tend to differ more in their ratings on the two item scales than subjects with a low score.

In a second analysis, the MEcom model was fitted to the data (see Figure 4(b)). The respective M*plus* (and LISREL) syntax for this model can be found in Appendix A. Since this model is equivalent to the MEref model it has the same model fit ($\chi^2(16) = 21.301,\ p = 0.1672,\ RMSEA = 0.026$). The variances, covariances, correlations and means of the MEcom model in Figure 4(b) are shown in Table 2.

The trait-factors τ_{well1} to τ_{well4} represent the mean of the true well-being ratings across both item formulations. The subjects rate their well-being on average above the center of the scale (the estimated means range from 3.696 to 3.809). On average, the subjects rate their well-being 0.366 scale scores higher than the average trait score, when asked with negatively formulated items, and, therefore 0.366 scale scores lower than the average trait score, when asked with positively formulated items (see estimated mean of the method factor). The difference between the answers to positively and negatively formulated items is on average 0.366 + 0.366 = 0.732 scale points. Note that this difference is represented by the mean of the method factor in the MEref model.

The estimated correlations between the trait factors in the MEcom model (range from $\widehat{\rho}(\tau_{well1,well4}) = 0.291$ to $\widehat{\rho}(\tau_{well3,well4}) = 0.447$) are just slightly higher than the correlations of the trait factors in the MEref model with the positive item formulation as reference method. Clearly, well-being is not stable over time but varies with occasions.

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 variances of the trait factors in the MEcom model are smaller than the variances of the trait factors in the MEref model, when employing the negative item formulation as the reference method. The subjects differ more in their well-being ratings on the negative-item scale than on the common well-being scores. Although the variance of the method factor ($\widehat{\sigma}^2(ME_{neg}) = 0.038$, $p < 0.001$) is small compared to the variances of the trait factors, it is still substantive. Note that the method factor 'only' represents the difference in the true outcomes of the variables measured by the negatively formulated items and the *average true outcomes of both methods*. The difference in the true outcomes *between both methods* is larger (see method factors in the MEref model). M_{neg} of the MEref model is twice the magnitude of M_{neg} in the MEcom model because, in the MEcom model $2 \cdot M_{neg} = M_{neg} + (-M_{pos}) = \tau_{t,bad} - \tau_t - (\tau_{t,good} - \tau_t) = \tau_{t,bad} - \tau_{t,good}$. Thus, according to calculation rules for variances, the variance of M_{neg} in the MEref model is four times as large as the variance of M_{neg} in the MEcom model (see for example, Steyer, 2003, for calculation rules). So, even if the variance of the method factor in the MEcom model is not that large, i.e. one quarter of the variance of the method factor in the MEref model, the variance of the difference in the answers to positive and negative items is substantive. The subjects thus differ significantly in the size of the method effect.

The method factors correlate significantly (range from $\widehat{\rho}(\tau_{well1}, ME_{neg}) = 0.155$ to $\widehat{\rho}(\tau_{well2}, ME_{neg}) = 0.207$) with the trait factors. Note that the correlations between method and trait factors in the MEcom model are located in between the correlations between method and trait factors in the two MEref models. The higher the subjects rate their well-being, the more they differ in the answer to positively and negatively formulated items. These correlations are, however, small.

Summarizing the results, there is a considerable method effect of item formulation for the measurement of well-being. There is a significant mean method effect, considerable interindividual differences in the method effects and the method factors correlate with the trait.

Discussion

Multitrait-multimethod data occur when several traits are measured by the same set of methods. For example, different raters (e.g. self, parent, superior, teacher) are used for the measurement of psychological constructs, written and oral exams for the measurement of school performance or positively and negatively formulated items for the measurement of self-esteem. The MEcom model can be used to model such data, and is applicable to the same data constellations to which traditional MTMM models can be applied.

Advantages and limitations of the MEcom model

Previous models, like the correlated trait-correlated method model or correlated trait-correlated uniqueness model, aim at modeling ‘common’ trait factors. However, in these models the factors are not clearly defined. In new models, like the CT-C(M-1) and MEref model, the factors are clearly defined (e.g. as true outcomes). In the MEref model, the definition of the factors is even derived from a theory. However, in both models, the CT-C(M-1) and the MEref model, no common trait factors, but reference-method-specific trait factors are modeled. In this article we introduced a new model, the MEcom model, which integrates both, a clear definiton of the factors, which is derived from a theory and the modeling of a common trait factor.

The MEcom model has many advantages. As in the MEref model, the definition of the trait and method factors is derived from the theory of individual causal effects (see Neyman, 1923; Rubin, 1974, 1978). The mean and the variance of the method factors, as well as their correlation among each other and with the traits, may be estimated. In the ME models, the variance of the manifest variables can not 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 comparing the variance of the method factors to the variance of the trait factors. When the variance of the method factor is small compared to the variance of the trait factors, convergent validity can be assured.

There are also some limitations of the ME models. The models, as they are presented here, are very restrictive. First, they presumes the same scaling for the manifest variables. However, in practice, methods with different scaling, e.g. questionnaires, behavior observations and physiological measures are often used. Second, the models assume invariant method effects across different traits. In practice, these method effects are not necessarily the same for each trait. The effect of item formulation may be different for the measurement of well-being than for the measurement of calmness. Extensions of the ME models with less restrictive assumptions are possible (1) by either introducing unrestricted factor loadings if the method effects are assumed to be unidimensional or (2) by using multiple indicators for measuring each trait-method unit if trait-specific method effects are assumed (in analogy to Eid, Lischetzke, Nussbeck, & Trierweiler, 2003 or Eid et al., 2008 who extended the CTC(M-1) model allowing for trait-specific method effects). The elaboration of these extensions for the ME models is one of the future research tasks.

Choice of a model

In both models, the MEref and the MEcom model, a method effect of a person is defined as an individual causal effect. In the MEref model we regard the individual effect of using method j as compared to method 1, whereas, in the MEcom model, a method effect is defined by comparing method j to the mean of the true outcomes of all methods measuring the same trait. 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 a trait common to all methods. In the MEref model, depending on the chosen reference method, different trait definitions are possible and one has to chose between them; for example, well-being measured by the positive item-scale or well-being measured by the negative item-scale. The MEref model may, thus, be convenient when there is a clear reference method, e.g. when subordinate and superior ratings are compared to 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 formulated items for the measurement of well-being, we are not concerned with a reference-method specific trait, e.g. 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.

Causal interpretation

Although the method effects in the ME models are defined according to the theory of individual causal effects, the method effects may not necessarily be interpreted as causal effects of the method. The true outcomes that we refer to in the theory of individual and average causal effects, are the true outcomes of one unit receiving treatment j but not receiving treatment k *and* of the same unit receiving treatment k but not receiving treatment j. In the multitrait-multimethod designs we consider the method as the treatment. Here we do observe the outcome of each unit u in *all* treatment conditions, i.e. the outcome of every person is observed for each method. However, since we may only apply one method *after* the other we can not be sure that the conditions are the same for each treatment (method). A person may be more fatigued when measured with a method applied later in the session than when measured at the beginning. Thus, we can not be sure whether we really measure the outcome variable we intend to measure according to the theory, i.e. the outcome that is obtained when treating a unit u in *only one* treatment condition and not in one after the other. In this case, the individual method effects may not necessarily be interpreted as causal effects of the method. In the example of using different item formulations, the method effects may not (only) be due to using different methods (item formulations), but could, for example, also be due to sequence effects, when one method is systematically applied *before* the other method. For instance, Orthner (2004) has shown that the answer to an item may depend on the response to previous items. Effects of test length might also be an alternative interpretation for a method effect. For example, Kraut, Wolfson, and Rothenberg (1975) found an increasing use of modal (vs. extreme) answer categories at the end of a test compared to the beginning of the test. Only when alternative interpretations (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), such as sequence, order and test length effects in questionnaires with positively and negatively formulated items, are ruled out, can the method effects be interpreted as causal effects of the method. Since, in the application presented here, the positively and negatively formulated items are not presented in a systematic sequence (e.g. first the positive-item, then the negative-item scale), but are unsystematically allocated to the questionnaire, we can rule out that these sequence effects are responsible for the interindividual differences in the individual method effects. Obviously, a substantive interpretation of the method effects is only meaningful when they can be regarded as causal effects of the method.

Thus, if method effects are not only introduced in the model in order to receive a better model fit, but are the focus of research, emphasis should be put on the design of the study. If, as in the application presented here, the positive- and negative-item scales are allocated unsystematically in the questionnaire, sequence and test length effects will be ruled out as alternative explanations for the method effects. If, in contrast, sequence effects are a viable alternative interpretation, then method effects can not be interpreted as causal effects of the method. Instead, the method effects may only be interpreted as latent difference scores, which may be due to many causes, e.g. item formulation, sequence and test length effects. However, even in these cases the ME models may be applied. The trait factors and their relationship with other variables can still be investigated. In such cases however, the values of the method factors should not be interpreted as effects of the method and are only relevant for investigating convergent validity and for improving the model fit.

Explaining the individual method effects

In order to properly interpret the scores of the constructs we measure with a certain measurement method, and in order to choose appropriate methods, the researcher should know the components of the method effects. Method effects may have many components (e.g. response style or social desirability). For example, Horan et al. (2003) 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 relation of method effects of 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 MMPI. However, although method effects are often modeled in applications, not many studies deal with their explanation. Therefore, further research should not only focus on accounting for method effects, but also on explaining them. Since the Method Effect models allow (under certain conditions) a causal interpretation of the method effects, they are appropriate models for the investigation of explanatory variables for the method effects.

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.

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

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasiexperimental designs for research on teaching. In N. Gage (Ed.), *Handbook on research on teaching* (p. 171-246). Chicago, IL: Rand McNally.

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.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin Company.

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

Eid, M., Lischetzke, T., Nussbeck, F., & 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., Nussbeck, F., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230-253.

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.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*, 579-588.

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. (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, p. 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.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, *60*, 774-776.

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 muliple dimensions of preadolescent self-concept: Inferences by teacher, mothers, and fathers. *Journal of Educational Psychology*, *83*, 393-404.

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. A Festschrift in honor of Karl Joereskog.* Lincolnwood: 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. Fifth edition.* Los Angeles, CA: Muthén & Muthén.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (reprint 1990). *Statistical Science*, *5*, 465-472.

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*, 557Ű595.

Orthner, T. M. (2004). On changing the positions of items in personality questionnaires analysing effects of item sequence using IRT. *Psychology Science*, *46(4)*, 446-476.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34-58.

Schmitt, M. (2006). Conceptual, theoretical and historical foundations of multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (p. 9-25). Washington DC: American Psychological Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

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* (p. 481-520). Pergamon, Oxford.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression [Probability and regression].* Berlin: Springer.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*, No.1.

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., Kröhne, U., Nagengast, B., & Fiege, C. (2008). Causal Effects in

Between-Group experiments and Quasi-Experiments: Theory. (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* (p. 197-220). Amsterdam: Kluwer Academic Press.

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.

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: University Trier.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF) [The multidimensional mood state questionnaire].* 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: Psychologisches Datenarchiv PsychData des Zentrums für Psychologische Information und Dokumentation ZPID.

Vautier, S., & Pohl, S. (in press). Bipolarity of latent change in STAI scores. *Psychological Assessment*.

Vautier, S., Steyer, R., Jmel, S., & Raufaste, E. (2005). Imperfect or perfect dynamic bipolarity? The case of antonymous affective judgments. *Structural Equation Modeling*, *12*, 391-410.

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.

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* and LISREL syntax of the MEcom model in the application

*M*plus *syntax*

```
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*;
[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];

well1 WITH well2 well3 well4;

well2 WITH well3 well4;

well3 WITH well4;

MEneg WITH well1 well2 well3 well4;
```

LISREL syntax

```
MO NY=8 NE=5 LY=FU,FI PS=SY,FR TE=DI,FR TY=FI AL=FR

VA 1 LY(1,1) LY(2,1)                          ! well1

VA 1 LY(3,2) LY(4,2)                          ! well2

VA 1 LY(5,3) LY(6,3)                          ! well3

VA 1 LY(7,4) LY(8,4)                          ! well4

VA -1 LY(1,5) LY(3,5) LY(5,5) LY(7,5)         ! ME2

VA 1 LY(2,5) LY(4,5) LY(6,5) LY(8,5)          ! ME2
```

Footnotes

[1]Note that the theory deals with a random experiment, the *single-unit trial*, and not with empirical data resulting from repeating this random experiment many times.

[2]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).

[3]according to the definition of latent-state stability in the Latent-State-Trait theory (Steyer, Ferring, & Schmitt, 1992; Steyer, Schmitt, & Eid, 1999)

[4]results of the Wald test

Table 1

Application of the MEref model: variances (diagonal), covariances (lower triangle), correlations (upper triangle) and means (last column) of the latent variables (|z| > 1.96*)*

Reference method: positive item formulation						
variables	τ_{good1}	τ_{good2}	τ_{good3}	τ_{good4}	M_{neg}	means
τ_{good1}	0.627*	0.271*	0.285*	0.278*	-0.090	3.329*
τ_{good2}	0.183*	0.724*	0.321*	0.390*	-0.017	3.380*
τ_{good3}	0.189*	0.230*	0.707*	0.432*	-0.040	3.344*
τ_{good4}	0.173*	0.260*	0.284*	0.612*	-0.087	3.443*
M_{neg}	-0.028	-0.006	-0.013	-0.026	0.152*	0.733*

Reference method: negative item formulation						
variables	τ_{bad1}	τ_{bad2}	τ_{bad3}	τ_{bad4}	M_{pos}	means
τ_{bad1}	0.724*	0.380*	0.387*	0.377*	-0.374*	4.062*
τ_{bad2}	0.301*	0.865*	0.428*	0.484*	-0.403*	4.113*
τ_{bad3}	0.300*	0.363*	0.832*	0.515*	-0.390*	4.077*
τ_{bad4}	0.270*	0.379*	0.396*	0.711*	-0.382*	4.175*
M_{pos}	-0.124*	-0.146*	-0.138*	-0.125*	0.152*	-0.733*

Table 2

Application of the MEcom model: variances (diagonal), covariances (lower triangle), correlations (upper triangle) and means (last column) of the latent variables ($|z| > 1.96$*)*

variables	τ_{well1}	τ_{well2}	τ_{well3}	τ_{well4}	M_{neg}	means
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*

Figure Captions

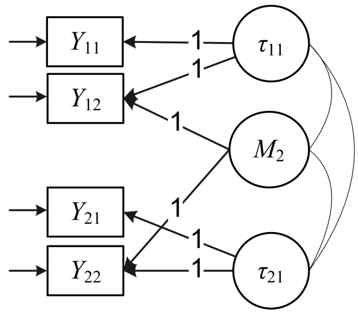
Figure 1. The MEref model for four variables measuring two traits by two methods. Arcs denote correlations.

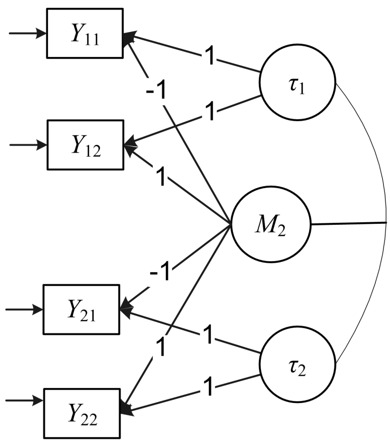
Figure 2. The MEcom model for four variables measuring two traits by two methods. Arcs denote correlations.

Figure 3. 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.

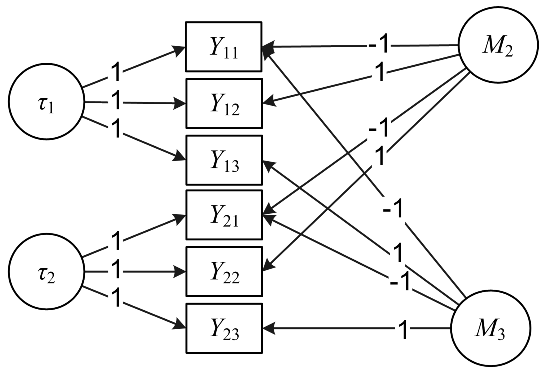
Figure 4. (a) MEref model with the positive item scale as the reference method and (b) the MEcom model for eight variables measuring well-being with positively and negatively formulated items at four measurement occasions. All factor loadings, that are not displayed are set to one. There are no restrictions on the means of and correlations between the latent variables.

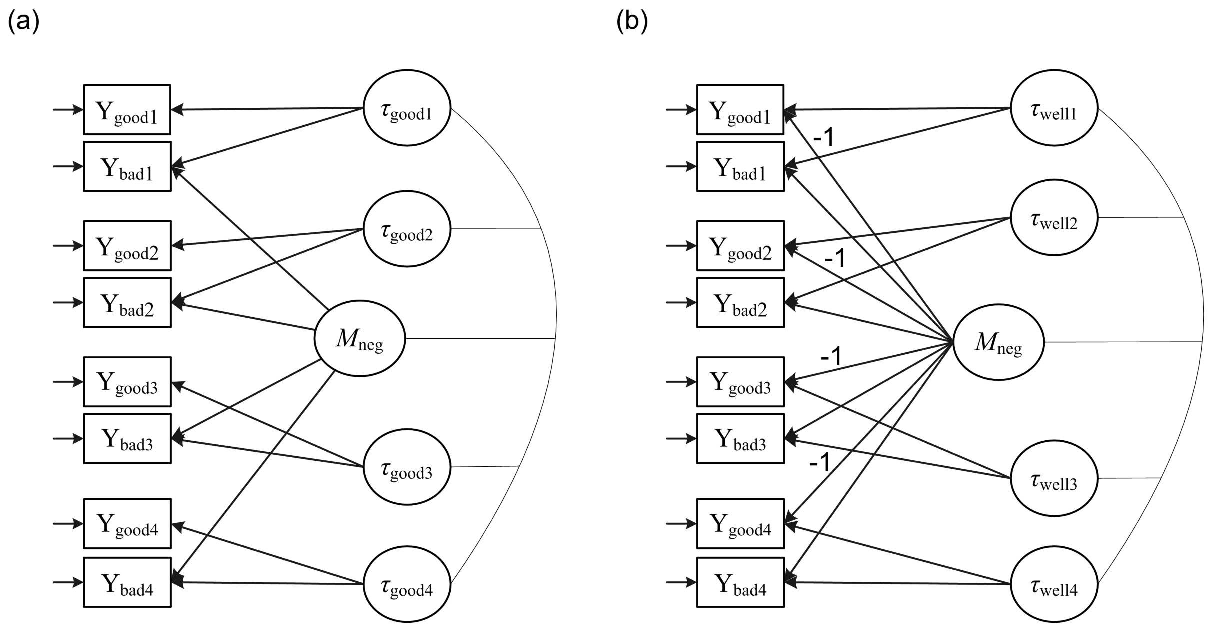
Common traits, Figure 1




Y_{11}
Y_{12}
Y_{21}
Y_{22}
τ_1
M_2
τ_2
1
-1
1
1
-1
1
1
1

Common traits, Figure 3




(a)
(b)
Y_{good1}
Y_{bad1}
Y_{good2}
Y_{bad2}
Y_{good3}
Y_{bad3}
Y_{good4}
Y_{bad4}
τ_{good1}
τ_{good2}
τ_{good3}
τ_{good4}
M_{neg}
τ_{well1}
τ_{well2}
τ_{well3}
τ_{well4}
-1
-1
-1
-1