J. R. Statist. Soc. A (2008)
171, *Part* 1, *pp.* 41–63

Modelling method effects as individual causal effects

Steffi Pohl and Rolf Steyer

Friedrich-Schiller-Universität, Jena, Germany

and Katrin Kraus

University of Uppsala, Sweden

[Received August 2006. Final revision August 2007]

Summary. Method effects often occur when different methods are used for measuring the same construct. We present a new approach for modelling this kind of phenomenon, consisting of a definition of method effects and a first model, the *method effect model*, that can be used for data analysis. This model may be applied to multitrait–multimethod data or to longitudinal data where the same construct is measured with at least two methods at all occasions. In this new approach, the definition of the method effects is based on the theory of individual causal effects by Neyman and Rubin. Method effects are accordingly conceptualized as the individual effects of applying measurement method j instead of k. They are modelled as latent difference scores in structural equation models. A reference method needs to be chosen against which all other methods are compared. The model fit is invariant to the choice of the reference method. The model allows the estimation of the average of the individual method effects, their variance, their correlation with the traits (and other latent variables) and the correlation of different method effects among each other. Furthermore, since the definition of the method effects is in line with the theory of causality, the method effects may (under certain conditions) be interpreted as causal effects of the method. The method effect model is compared with traditional multitrait–multimethod models. An example illustrates the application of the model to longitudinal data analysing the effect of negatively (such as 'feel bad') as compared with positively formulated items (such as 'feel good') measuring mood states.

Keywords: Causality; Method effect; Multitrait–multimethod; Negative item formulation; Structural equation modelling

1. Introduction

In the social sciences, theories and hypotheses often involve constructs, e.g. intelligence, school performance, depression or attitude. To test hypotheses, such constructs need to be measured via observable or 'manifest' variables. Usually, there are different ways to measure such a construct. For example, intelligence can be measured via various tests, school performance via oral and written examinations, depression via self-rating and the rating of a psychologist or attitude via questionnaires and observation of behaviour. Since each single-measurement method has its own specific effects on the resulting scores, multiple methods are often used in research. Holzbach (1978), for example, measured managerial performance with three methods: superior, self- and peer rating. Villar *et al.* (2006) also used different raters for the measurement of a construct. They measured parenting constructs with adolescent, mother and father reports.

Address for correspondence: Steffi Pohl, Lehrstuhl für Methodenlehre und Evaluationsforschung, Friedrich-Schiller-Universität Jena, Am Steiger 3, Haus 1, 07743 Jena, Germany.
E-mail: steffi.pohl@uni-jena.de

 0964–1998/08/171041

McConnell and Leibold (2001), in contrast, used not only different raters but also different tests for the measurement of a construct. They measured racial attitudes via self-rating using explicit measures and the implicit association test. Additionally, racial attitude was measured by external ratings of behaviour.

Campbell and Fiske (1959) proposed multitrait–multimethod designs, where several *traits* are measured by different *methods*. For example, *three* aspects of managerial performance, like 'administrative abilities', the 'ability to give feedback' and 'consideration', may be measured by *three* methods such as self-, supervisor and subordinate ratings. Another example is measuring a mood state such as wellbeing of individual members of a panel at *four* time points with *two* different tests. The term 'trait' is just a technical term with different substantive meanings in different applications and the same is true for the term 'method'. Particularly note that trait refers neither to the notion of a trait in differential psychology (see, for example, Matthews and Deary (1998)) nor to the trait concept that was introduced in latent state trait theory (see, for example, Steyer *et al.* (1992) or Steyer *et al.* (1999)).

According to Campbell and Fiske (1959), convergent validity is achieved when different measurement methods yield similar results in measuring the same trait. However, in most applications the measurement methods do not yield the same results. Instead, *individual method effects*, i.e. systematic differences between the measurements of a construct with different methods, occur. These method effects are person specific, i.e. may be different for each person. As an example we regard the measurement of mathematical performance with oral and with written examinations. Some subjects might score higher on mathematical performance in oral examinations than in written examinations, whereas for other subjects it might be vice versa. The difference in the mathematical performance that is measured with the two test modes is person specific. The various measures of mathematical performance are then not unidimensional. In addition to mathematical performance oral mathematics examinations may also measure communication skills and written mathematics examinations may also measure writing skills. The method effect, i.e. the difference between the two measurements, is due to the method specificity of both methods.

In data analysis, ignoring systematic individual method effects often results in an unsatisfactory model fit. Especially in longitudinal studies, when the same construct is measured by the same set of methods at different occasions, individual method effects need to be accounted for to obtain a reasonable model fit (e.g. Cole *et al.* (1996), Gignac (2006), Marsh (1996) and Motl and DiStefano (2002)). Although, in many applications, individual method effects are regarded as nuisance effects that have to be taken into account to reach a reasonable model fit, these effects are the focus of research in other applications. Examples are the differences between self- and peer ratings (e.g. Lewin *et al.* (1993) and Marsh and Byrne (1993)) or the effects of negative compared with positive item formulation (e.g. Horan *et al.* (2003) and Steyer and Riedl (2004)).

1.1. Models for multitrait–multimethod designs

Since Campbell and Fiske (1959), many models for the analysis of multitrait–multimethod data have been proposed. Many of them are based on confirmatory factor analysis (see, for example, Widaman (1985) and Marsh (1989)). The most frequently applied models are the correlated trait–correlated uniqueness (CTCU) model (Kenny, 1976; Marsh, 1989; Marsh and Craven, 1991) and the correlated trait–correlated method (CTCM) model (Jöreskog, 1974). A more recent model is the CTCM minus one model (Eid, 2000) (the CT-C(M-1) model) and its extension for multiple indicators (Eid *et al.*, 2003). These three models will be explained in more detail by using an example that is based on a study by Mount (1984). In this study, three aspects of

managerial performance, 'administrative abilities', the 'ability to give feedback to subordinates' and 'consideration', were rated by the employee himself, by a supervisor and by a subordinate. Thus there are nine measured variables Y_{tj}, where t denotes the trait (aspect of performance) and j the method (rater). The path diagrams of the three multitrait–multimethod models for this example are depicted in Fig. 1. They represent systems of regression equations. The arrows indicate which variables are the regressands (dependent variables) and which are the regressors (independent variables) in these regressions.

1.1.1. Correlated trait–correlated uniqueness model

In the CTCU model (Kenny, 1976; Marsh, 1989; Marsh and Craven, 1991), each trait (see T_a, T_f or T_c in Fig. 1(a)) represents the common source of the three ratings that are obtained in measuring the same aspect of managerial performance. Each manifest variable (rectangle) is regressed on its respective trait. There is no explicit latent variable representing the individual (person-specific) effects of the rater. However, these effects are implicitly accounted for by correlations (which are represented by arcs) between error terms (indicated by arrows on the left-hand side of the rectangles) of variables measured by the same rater. These correlations are assumed to be due to the systematic individual effects of the rater affecting the ratings of all three traits (performance aspects). Joe's subordinate, for example, might give higher ratings (than Joe himself) of all three aspects of Joe's performance assuming that a high rating is socially desired, whereas Carl's subordinate might rate Carl's performance lower than Carl himself, because he shows less socially desired behaviour. The variances of the manifest variables are only decomposed into

(a) a trait component representing the score of the rated manager on the performance dimension to be assessed and
(b) an error term.

The error variances are due to measurement error as well as rater-specific effects (the individual method effects).

In general, this model does not suffer from estimation or identification problems (Marsh, 1989). However, it has some limitations. The individual effects of the rater are neither explicitly defined, or explicitly represented in the model, nor assumed to be unidimensional within a given rater. Instead they are part of the error terms. Correlations of the individual effects between different raters are not allowed, because the corresponding error terms are assumed to be uncorrelated. The model, therefore, does not allow that different raters of one employee systematically show the similar response tendencies. Instead it implicitly assumes that these rater effects are independent of each other. The model also does not allow the rater effects to depend on the trait scores (otherwise, the error terms would correlate with the traits). Furthermore, reliability coefficients are underestimated, because the error terms contain systematic rater-specific effects. Finally, the rater effects cannot be related to external variables such as social desirability and, last but not least, the assumption of uncorrelated rater effects can lead to biased estimates for the trait variances and covariances (Conway *et al.*, 2004).

1.1.2. Correlated trait–correlated method model

In the CTCM model (see Fig. 1(b); Jöreskog (1974) and Widaman (1985)), a latent variable is modelled for each trait and also for each rater, i.e. each manifest variable is regressed on one trait and one rater factor. The trait factors (T_a, T_f and T_c) represent the common sources of the ratings of the three raters due to the common trait that is considered, whereas the scores of the method factors (M_{se}, M_{sp} and M_{sb}) represent the individual effects on the ratings due to the

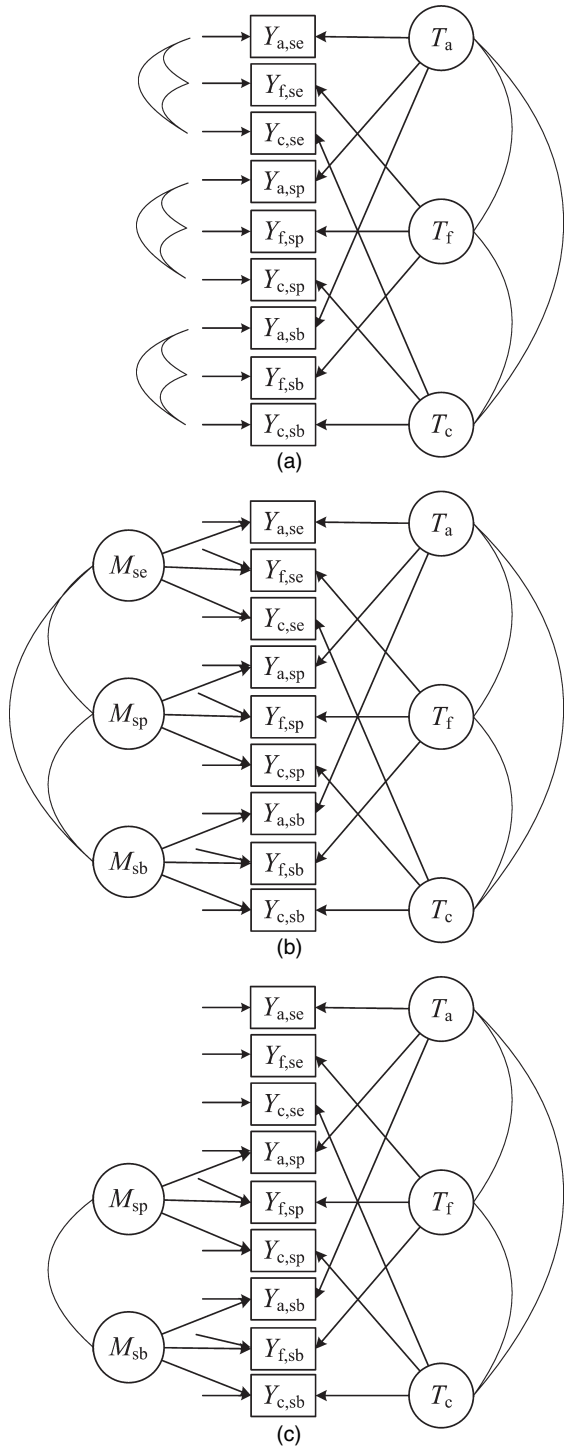

Fig. 1. (a) CTCU model, (b) CTCM model and (c) CT-C(M-1) model (for simplicity, loadings of the manifest variables on the factors are not depicted): *T*, trait; *M*, method; a, administrative abilities; f, ability to give feedback; c, consideration; se, self-rating; sp, supervisor rating; sb, subordinate rating

same rater. The trait factors as well as the method (rater) factors are allowed to correlate among each other. Correlations between trait and method factors, however, are not permitted, i.e. the individual effects of the raters are supposed to be uncorrelated with the trait scores. Thus, the model allows, for instance, no tendency for higher method effects (e.g. due to socially desired behaviour) of the subordinate when the managerial performance of his superior is low than when the managerial performance is high.

One convenient aspect of this model is that the variance of the manifest variables can be additively decomposed into a trait-specific part, a rater-specific part and an error term. The rater effects are represented by a separate factor for each rater, and they may be regressed on explanatory variables. Disadvantages are identification and estimation problems, especially when all rater factors are correlated (Marsh, 1989; Marsh and Grayson, 1995). Furthermore, like in the CTCU model, neither the trait factors nor the method factors are clearly defined, which makes their interpretation difficult. This critique will become more lucid when the CTCM model is compared with the models that are introduced in the next sections.

1.1.3. Correlated trait–correlated method minus one model

The CT-C(M-1) model (Fig. 1(c); Eid (2000)) overcomes some of the problems of the other models. As in the two other models, there are three trait factors (T_a, T_f and T_c). The number of method (rater) factors, however, is one less than the number of methods. A reference method (rater) needs to be chosen to which all the other methods are compared. In this example, the self-rating is chosen as the reference method.

Eid used the concepts of classical test theory to define the factors. Each of the trait factors T in the CT-C(M-1) model represents the true score variable τ_{t1} of the manifest variable Y_{t1} measuring trait t by the reference method 1. Choosing the first method as the reference does not restrict generality. Eid regressed τ_{tj}, the true score variable of a variable Y_{tj} measuring trait t by method j, on τ_{t1}, i.e. $E(\tau_{tj}|\tau_{t1})$. The residual of this regression is then defined as the *method effect variable* M_{tj}, i.e.

$$M_{tj} \equiv \tau_{tj} - E(\tau_{tj}|\tau_{t1}). \tag{1}$$

The *method factors* occurring in Fig. 1(c) result from assuming that each residual that is defined in equation (1) is a specific deterministic function of a common latent variable M_j, namely

$$M_{tj} = \kappa_{tj} M_j, \tag{2}$$

where κ_{tj} is a real number. For reasons of identification κ_{1j} is set to 1. Defining a method factor M_j in this way implies that its expected value is 0 and that it is uncorrelated with the regressor τ_{t1}. Note again that τ_{t1} represents the true score variable of the manifest variable Y_{t1} assessing trait t by the reference method 1.

In the Mount example, the trait factors represent the true score variables of the manifest variables $Y_{t,se}$, measuring the three traits by self-ratings (the reference method). For identification, the factor loading $\kappa_{a,j}$ of the manifest variables measuring administrative abilities on the respective method factors are set to 1. The rater factor M_{sp} then represents the residual of the regression $E(\tau_{a,sp}|\tau_{a,se})$ of the true scores of the supervisor ratings of administrative abilities on the respective true scores of the self-ratings. The scores of the rater factor M_{sb} are the residuals of the regression $E(\tau_{a,sb}|\tau_{a,se})$. The method effects of the subordinate rating compared with the self-rating on the traits feedback f and consideration c are related to the method factor M_{sb} by equation (2).

In the model that is described above, each trait–method combination is measured by only one manifest variable (a single-indicator model). It should be mentioned that Eid *et al.* (2003)

also extended the CT-C(M-1) model to a multiple-indicator model, where each trait–method combination is measured by at least two variables. This allows the introduction of trait-specific method effects.

The CT-C(M-1) model has several advantages compared with the other two models that were described above. In the CT-C(M-1) model, the latent variables are uniquely defined as true score variables or as residuals of a regression. The trait factors are not nebulous 'sources' of covariation, but mathematically well-defined true score variables (see, for example, Lord and Novick (1968), Steyer (2001) and Zimmerman (1975)) and the same also pertains to the definition of the method factors as residuals with respect to specific regressions. This is not only a mathematical virtue but it also helps in interpreting the latent variables and in deciding whether or not it is in fact these variables that we are interested in. In the CT-C(M-1) model the variance of a manifest variable can additively be decomposed into components due to trait, method and error variances. Exceptions are the manifest variables that are measured with the reference method. They do not have a method-specific component, at least as long as no different method is chosen as a reference. Finally, it should be noted that the model does not suffer from identification problems.

Nevertheless, there are also limitations to the CT-C(M-1) model. First, it is not symmetric and the definition of the method effect (equation (1)) has strange implications. For example, the absolute value of the individual effect $M_{\mathrm{sp}-\mathrm{se}}(u)$ for a rated person u of using superior rating instead of self-rating is not necessarily equal to the absolute effect $M_{\mathrm{se}-\mathrm{sp}}(u)$ of using self-rating instead of superior rating. Second, the variances of the method factors as well as the model fit are not invariant to the choice of the reference method.

1.2. Shortcomings of the multitrait–multimethod models

All three models that have been described above have one thing in common: the method effects are implicitly or explicitly regarded as residuals. Jöreskog (1971) defined the method factors as 'what is left over after all trait factors have been eliminated' (page 128) and being 'independent of the particular traits that the method is used to measure' (page 128). Eid (2000) explicitly stated that the CT-C(M-1) model is based on the assumption that the 'residual indicates the method-specific effect of a method j with respect to a method 1 that is chosen as a comparison method' (page 245). Regarding method effects as residuals may cast some doubts on the meaningfulness of this definition for an *effect*. In usual effect definitions (e.g. in the theory of individual causal effects; Neyman (1990) and Rubin (1974, 1978)) an effect is zero when there is no difference between the compared scores. Furthermore the size of the effect is the same when comparing treatment A with treatment B as when comparing treatment B with treatment A. However, in the CT-C(M-1) model the method effects do not have these properties. A method effect of zero in the CT-C(M-1) model does not necessarily mean that the raters in the Mount example agree in their scorings (see Eid *et al.* (2007)). The method factor for the subordinate rating does not represent the amount of overestimation or underestimation of the subordinate rating compared with the self-rating on managerial performance. Furthermore, as described in Section 1.1.3, the model and therefore the size of the method effect is not symmetric. The individual method effect of comparing the subordinate with the self-rating is not necessarily the same as the individual method effect of comparing the self- with the subordinate rating. This is not how we typically define an effect. The interpretation of the method effects as well as of the relationship of the method effects with explanatory variables depends on the reference method that is chosen.

There are, however, substantive researchers in the field who 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), page 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. The subordinate rating is due to a set of causes. Social desirability, need for approval and sympathy are rater-specific variables that are represented in the method factor. They are psychological constructs and can, like other psychological constructs, be regarded as 'causes'. Thus each method measures a set of constructs.

Defining the method effects as residuals implies that they have all the properties of residuals: the means of the method effects are defined to be 0 (the CT-C(M-1) model) and usually they are not analysed at all (the CTCU and CTCM models). But why should the mean of the effects of one measurement method compared with another be 0? Since this is implied by the definition of method effects in the CT-C(M-1) model, this property casts some doubts on the meaningfulness of the definition itself. We might, for example, be interested in how much, on average, subordinates overestimate or underestimate their superior's managerial performance compared with the self-rating. Or it might be of interest how much, on average, the answer to a positively formulated item differs from the answer to a negatively formulated item.

Furthermore, the assumption of method effects as residuals implies that the method effects are uncorrelated with the trait factors. Marsh and Grayson (1995) have 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'

(page 181). This is especially true for the CTCU and the CTCM model. In fact, it might be quite reasonable to allow for correlations between trait and method factors. It might, for example, be possible that the overestimation or underestimation of managerial performance by the subordinates covaries with the level of the managerial performance (as measured by the self-rating). For low managerial performance, subordinate and self-ratings might differ more than for high performance. The effect of scoring higher on negatively formulated items might be larger for a low mood level than for a high level. Although in the CT-C(M-1) model there is no correlation between trait and method factors allowed, the underestimation or overestimation may covary with the level of the trait factor (see Eid *et al.* (2007)).

There are new developments, where method effects are not defined as residuals. A recent approach that was presented by Lischetzke *et al.* (2002) (also see Eid *et al.* (2007)) proposed to use the differences between the true scores of manifest variables by measuring the same trait with two different methods to define a method factor. This approach is inspired by the so-called 'true change models' (Steyer, Eid and Schwenkmezger, 1997) in which the differences between two true score variables pertaining to two different time points are modelled as latent change variables. In the same vein, a method effect variable is conceptualized as the difference between true score variables pertaining to two different methods.

The aim of this paper is to introduce a model in which the definition of the method effects is not based on technical considerations, but on theoretical reasonableness. Schmitt (2006) has already convincingly elaborated the theoretical reasonableness of regarding methods as 'sets of causes' (page 24). In this paper the latent difference approach (Steyer, Eid and Schwenkmezger, 1997) of modelling method effects (see also Lischetzke *et al.* (2002)) is combined with the theory of individual and average causal effects going back to Neyman (1990) and Rubin (1974, 1978)

(see, for example, Steyer *et al.* (2007) for a comprehensive introduction). Modelling method effects as latent difference scores allows the estimation of the mean method effects as well as covariation of the method effects with the trait. Conceptualizing method effects as individual causal effects may allow a causal interpretation of the method effects and invites a search for variables explaining the interindividual differences in the individual causal method effects.

In the following sections, the method effect model is introduced, explicitly defining individual method effects and introducing assumptions specifying the model. The model is illustrated by an example on measuring wellbeing with positively and negatively formulated items at four measurement occasions. Finally strengths and limitations of the model are discussed and compared with the other models that were mentioned above.

2. The method effect model

In this section we shall first introduce an individual method effect as the difference between the two true scores of the manifest variables Y_{tj} and Y_{tk} for an observational unit (person) u measuring trait t with methods j and k respectively. We then show that these individual method effects can be interpreted as individual causal effects in the Neyman–Rubin tradition if alternative interpretations can be ruled out. Finally, the method effect model is defined by introducing some assumptions.

2.1. Definition of method effects

According to classical test theory, a manifest variable Y_{t1} measuring trait t with method 1 can always be decomposed into a true score variable τ_{t1} and an error term ε_{t1}:

$$Y_{t1} = \tau_{t1} + \varepsilon_{t1}, \tag{3}$$

provided that Y_{t1} has a finite expected value. Now suppose that the manifest variables Y_{t1} and Y_{tj} measuring the same trait t with methods 1 and j respectively measure exactly the same latent variable in the sense that their true score variables are identical, i.e. $\tau_{t1} = \tau_{tj}$, implying that

$$Y_{tj} = \tau_{t1} + \varepsilon_{tj}. \tag{4}$$

In this case, there would be no systematic method effects, and in classical test theory we would say that the two variables Y_{t1} and Y_{tj} are τ equivalent. However, if there *are* systematic method effects, then equation (4) will not hold and we must replace it by

$$Y_{tj} = \tau_{t1} + (\tau_{tj} - \tau_{t1}) + \varepsilon_{tj}, \tag{5}$$

which is always true. In this equation, the difference $\tau_{tj} - \tau_{t1}$ between the two true score variables represents the systematic effects of using method j instead of method 1 for measuring trait t. Hence, we call

$$M_{tj} \equiv \tau_{tj} - \tau_{t1} \tag{6}$$

the *method effect variable of* Y_{tj}. Its scores

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

are the effects of unit (or individual) u being measured on trait t with method j as compared with being measured on the same trait t with the reference method 1. Choosing method 1 as the reference method does not restrict generality, because each method can be chosen to be 'method 1'.

Note that equation (6) defines a method effect variable that is *specific for each trait t*. Later, we shall introduce the assumption

$$M_{tj} = M_{sj} \equiv M_j \tag{8}$$

that the method effect variables do not differ between different traits t and s (see Section 2.2).

Also note that the method effect that is defined in equation (6) is not necessarily the *causal* effect of using method j instead of method 1 for measuring trait t, even though our definition of a method effect is in line with the concept of individual causal (treatment) effects in the Neyman–Rubin tradition. In this tradition, the individual causal effect of treatment j compared with treatment 1 for unit u is defined as the difference $\tau_j(u) - \tau_1(u)$ in the true scores of u under treatment j and under treatment 1 (see, for example, Steyer (2005)).

Although this definition refers to a random experiment in which unit u is *either* in treatment j *or* in treatment 1, the method effect $M_{tj}(u) = \tau_{tj}(u) - \tau_{t1}(u)$ of Y_{tj} refers to a random experiment in which the trait t of unit u is measured *with both methods*, j and 1. Therefore, a causal interpretation of the individual method effects $M_{tj}(u)$ might not be valid if there are any effects (such as learning or other sequence effects) due to *repeatedly* measuring the same units (for a comprehensive presentation of the theory of causal effects see Steyer *et al.* (2007)). However, if all alternative explanations (see Campbell and Stanley (1963)) of the method effects M_{tj} can be ruled out, the method effects *do* have a causal interpretation.

Unless the alternative explanations for the individual differences in equation (7) can be ruled out, the term 'individual method effect' is just a technical term with no substantive meaning other than referring to the difference between two true scores. If, for instance, method j is applied to measure the trait t *after* applying method 1, this difference may be also due to a sequence effect, and not only to the difference in the two methods. As an example, let us consider the assessment of wellbeing. If Joe is first asked to rate whether he feels 'good' and a short time later whether he feels 'bad', he might be irritated by the fact of being asked the same (but inverted) question twice. His response to the item 'feel bad' might have been systematically different if he had been asked the same question without being asked whether he feels 'good' before. In this case, the systematic difference between the true scores of the two response variables may, from a substantive point of view, not only be a method effect, owing to the differences in the items 'feel good' and 'feel bad'; instead this difference may also be due to the sequence in which the two items (measurement methods) are applied. The difference between the true scores of the two response variables is then due to a combination of method and sequence effects. The proportion of method and sequence effect cannot be determined. Obviously, this causality issue is of crucial importance when it comes to substantive interpretations and explanations of the method effects.

2.2. Assumptions defining the single-indicator method effect model

In the preceding paragraphs we just have dealt with the definition of a method effect variable for a single trait t. To introduce model assumptions allowing us to identify the theoretical parameters, let us now consider four manifest variables measuring *two* traits with *two* methods. For Y_{11} measuring the first trait with the reference method 1, the following equation is always true:

$$Y_{11} = \tau_{11} + \varepsilon_{11}. \tag{9}$$

Hence, the manifest variable Y_{11} is decomposed into the true score variable τ_{11} and a measurement error variable ε_{11}. For Y_{12}, measuring the same trait with method 2, the following equations are always true:

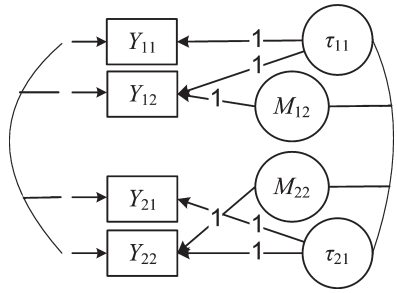

Fig. 2. Path diagram of the method effect model for four variables measuring two traits by two methods: arcs denote correlations; assumptions have not yet been introduced

$$\begin{aligned} Y_{12} &= \tau_{12} + \varepsilon_{12}, \\ &= \tau_{11} + (\tau_{12} - \tau_{11}) + \varepsilon_{12}, \\ &= \tau_{11} + M_{12} + \varepsilon_{12}, \end{aligned} \tag{10}$$

with the method effect variable $M_{12} \equiv \tau_{12} - \tau_{11}$ defined as the difference between the true score variables of the manifest variable Y_{12} measuring trait 1 with method 2 and the manifest variable Y_{11} measuring trait 1 with (the reference) method 1. The scores of the random variable M_{12} are the trait-1-specific effects of using method 2 instead of method 1.

The same applies to the second trait. The manifest variable measuring the second trait with method 1 can be decomposed as

$$Y_{21} = \tau_{21} + \varepsilon_{21}. \tag{11}$$

Again, the following equations for the manifest variable Y_{22}, measuring the second trait with method 2, are always true:

$$\begin{aligned} Y_{22} &= \tau_{22} + \varepsilon_{22}, \\ &= \tau_{21} + (\tau_{22} - \tau_{21}) + \varepsilon_{22}, \\ &= \tau_{21} + M_{22} + \varepsilon_{22}, \end{aligned} \tag{12}$$

with the variable $M_{22} \equiv \tau_{22} - \tau_{21}$ defined as the difference in the true score variables of the manifest variable Y_{22} measuring the second trait with the second method and the true score variable of the manifest variable Y_{21} measuring the same trait with method 1. The scores of M_{22} represent the individual trait-2-specific effects of using method 2 instead of method 1. Note that each person u may have a different score on M_{22}. These values of the method effect variable are the individual method effects $M_{22}(u) = \tau_{22}(u) - \tau_{21}(u)$. They are true *intraindividual difference scores*. The same holds for the scores of M_{12}. The model corresponding to these measurement equations is depicted in Fig. 2. Note that this model is not yet identified. For the identification of the model, assumptions need to be introduced.

2.2.1. Assumptions

Until now, only definitions have been given, but no assumptions have been made. To identify the theoretical parameters such as the variance of the method factors, their covariance with the trait factors, and the variances of the error variables, we must introduce appropriate assumptions. The *first assumption* is

$$M_{12} = M_{22} \equiv M_2. \tag{13}$$

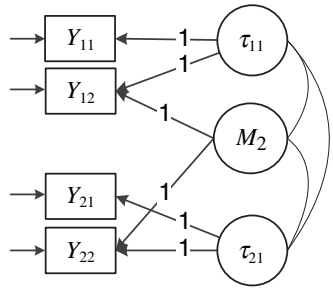

Fig. 3. Path diagram of the identified method effect model with four variables measuring two traits by two methods

With this equation we assume that the method effect variables, and therefore also the individual method effects, are the same for each trait. In longitudinal designs, this assumption is quite reasonable. The difference in the true scores of the answers to positively and negatively formulated items may be the same at each measurement occasion (see Section 3). The *method factor* M_2 in the method effect model (Fig. 3) represents the *method effects* of method 2 compared with the reference method 1 *for both traits*. Thus, the equations of the measurement model for all manifest variables simplify as follows:

$$Y_{11} = \tau_{11} + \varepsilon_{11}, \tag{14}$$

$$Y_{12} = \tau_{11} + M_2 + \varepsilon_{12}, \tag{15}$$

$$Y_{21} = \tau_{21} + \varepsilon_{21}, \tag{16}$$

$$Y_{22} = \tau_{21} + M_2 + \varepsilon_{22}. \tag{17}$$

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

$$\operatorname{cov}(\varepsilon_{tj}, \varepsilon_{sk}) = 0 \qquad (t, j) \neq (s, k), \quad t, s = 1, 2, \quad j, k = 1, 2. \tag{18}$$

Following from the definitions of true score and error variables (see, for example, Novick (1996), Steyer (1989, 2001) and Zimmerman (1975)), the true score variables τ_{j1} and the method factor M_2 are uncorrelated with the error variables:

$$\operatorname{cov}(\tau_{t1}, \varepsilon_{sj}) = \operatorname{cov}(M_2, \varepsilon_{sj}) = 0 \qquad t, s = 1, 2, \quad j = 1, 2. \tag{19}$$

It should be emphasized that, unlike equations (13) and (18), equations (19) are not assumptions. Instead they are logical consequences from the definitions of true score and error variables in classical test theory as specific conditional expectations and their residuals (see the appendix of Steyer *et al.* (2007) for the mathematical background).

The correlations between the trait factors as well as between the method factors are not restricted to be 0. Since, in contrast with the definition of method effects in the CT-C(M-1) model (equation (1)), the method factors in the method effect model represent differences in true score variables (equation (6)), they will usually correlate with the true score variables themselves. Thus, the trait factors and the method factors may also covary with each other. There is no restriction on the means of the trait and of the method factors. The identification of the method effect model for this example (Fig. 3) is presented in Appendix A.

2.3. Generalizing the method effect model

The model can be extended to the measurement of more than two traits measured by more than two methods. In the general method effect model with fixed loadings, the measurement model for the manifest variable Y_{tj} measuring trait t with method j is

$$Y_{tj} = \begin{cases} \tau_{t1} + \varepsilon_{t1} & \text{for } j = 1, \\ \tau_{t1} + M_j + \varepsilon_{tj} & \text{for } j \neq 1, \end{cases} \tag{20}$$

where $j = 1$ denotes the reference method. The error variables are assumed to be uncorrelated with each other:

$$\text{cov}(\varepsilon_{tj}, \varepsilon_{sk}) = 0, \qquad (t, j) \neq (s, k), \tag{21}$$

and all error variables are uncorrelated with all true score variables τ_{t1} and with all method factors M_j,

$$\text{cov}(M_j, \varepsilon_{tk}) = \text{cov}(\tau_{t1}, \varepsilon_{sj}) = 0, \tag{22}$$

where the first index always refers to a trait, and the second to a method. (Again, equations (22) follow from the definitions of true score and error variables as specific conditional expectations and their residuals.)

Note that the covariance matrix of the latent variables (traits and method factors) is unrestricted by these model assumptions, i.e. the trait factors may correlate among each other and with the method factors M_j, and the method factors M_j may correlate among each other as well. In fact, since the method factors are defined as differences between true score variables (see equations (6) and (8)), a correlation between method factors and trait factors is to be expected. Furthermore, the expected values of the trait factors and the method factors M_j are unrestricted. It should be emphasized that, even though all latent variables, i.e. method factors and trait factors, may correlate among each other and no restrictions have been made on the expectations of the latent variables, the method effects model is already identified for four manifest variables measuring two traits by two methods (see Appendix A).

3. Application

In this section, the method effect model will be illustrated by an example on measuring mood states with two scales (methods). The first scale consists of positively and the second of negatively formulated items. We shall present the estimated parameters and model fit statistics and extend the model by an additional variable explaining the interindividual differences in the individual method effects.

Many researchers have been puzzled by the empirical evidence that measures of psychological constructs that are assessed by positively formulated items (e.g. 'I feel good') differ far more than we might expect from the measurement of the same construct assessed by negatively formulated items (e.g. 'I feel bad'). Some studies focused on correlations between the factor of the positively and the factor of the negatively formulated items (Russell and Carroll, 1999; Barrett and Russell, 1998), whereas others used confirmatory factor analysis multitrait–multimethod models (Horan *et al.*, 2003; Motl and DiStefano, 2002) to analyse this phenomenon. Although this kind of study shows evidence of the divergence of the methods using positively *versus* negatively formulated items, Steyer and Riedl (2004) used correlations between latent state residuals (i.e. the deviations of the latent actual mood state from the latent habitual mood state) showing a high degree of convergence. These findings have been corroborated by Vautier *et al.* (2005) by

using correlations between latent change scores, i.e. correlations between the true change (over different time points) in the scales of the positively and negatively formulated items.

In the present application, we do not consider latent change scores and their correlation between positively and negatively formulated items. Instead, we focus on latent differences between the scales of positively and negatively formulated items: the individual method effects. We shall estimate the average method effect, the variance of the individual method effects, the correlations of the individual method effects with the latent states and the regression of the method effects on explanatory variables.

3.1. Data

The data (Steyer *et al.*, 2004) that were used for this analysis stem from a validation study of a questionnaire measuring mood states. The sample consists of 503 subjects. 291 are female and 212 are male. Their age ranges from 17 to 77 years with a mean age of 31.2 years. For a more detailed description of the sample, see Steyer *et al.* (1991). For the analyses four cases with missing values were deleted listwise.

The participants were asked to complete a couple of questionnaires at four occasions with 3 weeks between measurements. Among others, mood states and personality traits were measured. The personality traits were measured by the Freiburg personality inventory (Fahrenberg *et al.*, 1984). One scale in the Freiburg personality inventory measures openness. Subjects with a high score on openness self-critically admit weaknesses, whereas subjects with a low score are strongly oriented on norms and answer as is socially desired. Another scale measures emotionality. Subjects with a high score on emotionality have many problems and inner conflicts. They are described as depressed, anxious and nervous. Subjects with a low score on emotionality are satisfied with their life and even tempered. The scales consist of 12 and 14 items respectively. The subjects could agree or disagree to item statements on a two-point scale. The scores of the items were aggregated across all items and all measurement occasions by calculating the mean.

The subjects also rated their mood states on items from the wellbeing scale of the German version of the multidimensional mood state questionnaire (Steyer, Schwenkmezger, Notz and Eid, 1997). The items of this scale are statements containing adjectives describing the mood (e.g. 'I feel well'). The subjects are asked to rate their actual mood states on a five-point Likert scale ranging from 'not at all' to 'very much'. Four items are formulated positively (e.g. 'I feel good'), and four items are formulated negatively (e.g. 'I feel bad').

For the further analyses the scores of the negatively formulated items were reversed. The positively as well as the negatively formulated items were then aggregated, yielding scale scores by calculating their mean within each occasion. Altogether eight manifest variables (two scales at four occasions) were analysed: for each measurement occasion one scale score of positively ($Y_{\text{good1}}, Y_{\text{good2}}, Y_{\text{good3}}$ and Y_{good4}) and one scale score of negatively formulated items ($Y_{\text{bad1}}, Y_{\text{bad2}}, Y_{\text{bad3}}$ and Y_{bad4}). The scale scores vary between 1 and 5. The items are unsystematically distributed across the questionnaire. Positively and negatively formulated items are located at various positions in the questionnaire, ensuring that there is no systematic sequence of positively and negatively formulated items such as a block of positively formulated items followed by the negatively formulated items.

3.2. Establishing the individual method effects

All analyses were conducted by using LISREL 8.71 (Jöreskog and Sörbom, 2004). The manifest variables were treated as continuous variables and maximum likelihood was used for estimation. For more information on structural equation modelling, see Bollen (1989). The first model

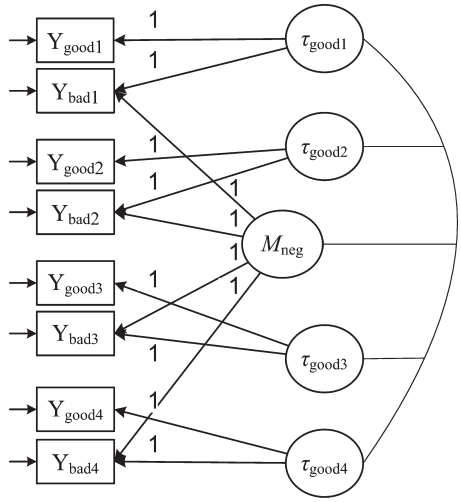

Fig. 4. Method effect model for two scales of wellbeing: the scale of positively formulated items is the reference method; all latent variables are allowed to correlate

that was analysed was a multistate model without method factors, where, for each occasion, one latent variable was modelled with unrestricted factor loadings and no restrictions on the mean structure. This model, ignoring individual method effects, did not fit ($\chi^2(14) = 238.44$; $p < 0.001$; root-mean-square error of approximation RMSEA $= 0.179$). RMSEA is a standard goodness-of-fit index in structural equation modelling. Rules of thumb are that RMSEA < 0.05 indicates a close fit, $0.05 <$ RMSEA < 0.08 a reasonable fit and $0.10 <$ RMSEA a poor fit (Browne and Cudeck, 1993).

The second and third models that were analysed were method effect models, one of which is depicted in Fig. 4 (see Appendix B, for the LISREL syntax). The two method effect models differ only in the choice of the reference method. As expected, the two models have exactly the same model fit ($\chi^2(16) = 24.52$; $p = 0.079$; RMSEA $= 0.033$) and fit the data well. The variances, covariances, correlations and means of both models are shown in Table 1. On the scale of positively formulated items, the subjects rate their wellbeing, on average, slightly above the centre of the rating scale (the estimated means ranging from 3.33 for τ_{good1} to 3.46 for τ_{good4}). On the scale of the negatively formulated items the estimated means range from 4.07 for τ_{bad1} to 4.20 for τ_{bad4}. The mean method effect of using the negative item instead of the positive item scale is 0.75 ($t = 36.21$). Hence, the subjects score on average 0.75 scale points higher on the wellbeing scale when rating their mood on negatively formulated items rather than on positively formulated items. Note that the means of the method factors differ only in sign, but not in magnitude, between the two models.

The estimated correlations between the latent state variables range from $\hat{\rho}(\tau_{\text{good1}}, \tau_{\text{good4}}) = 0.27$ to $\hat{\rho}(\tau_{\text{bad3}}, \tau_{\text{bad4}}) = 0.49$, where the correlations between the latent state variables of the negative item scale seem to be systematically higher than the corresponding correlations between the latent state variables of the positive item scales. Wellbeing is not stable over time. However, the negative item scale is more stable over time than the positive item scale.

The estimated variances of the latent state variables range from $\hat{\sigma}^2(\tau_{\text{good4}}) = 0.64$ to $\hat{\sigma}^2(\tau_{\text{good2}}) = 0.73$ for the positively and from $\hat{\sigma}^2(\tau_{\text{bad1}}) = 0.74$ to $\hat{\sigma}^2(\tau_{\text{bad2}}) = 0.85$ for the negatively formulated variables. Hence, the subjects differ from each other more in their wellbeing ratings on negatively than on positively formulated items.

Table 1. Application of the method effect model: variances (diagonal), covariances (lower triangle), correlations (upper triangle) and means (last column) of the latent variables†

Variable	τ_{good1}	τ_{good2}	τ_{good3}	τ_{good4}	M_{neg}	*Mean*
Reference method: positive item formulation						
τ_{good1}	0.66	0.28	0.29	0.27	−0.10	3.33
τ_{good2}	0.19	0.73	0.32	0.38	−0.04	3.38
τ_{good3}	0.20	0.23	0.72	0.42	−0.07	3.34
τ_{good4}	0.18	0.26	0.29	0.64	−0.10	3.46
M_{neg}	−0.03	−0.01	−0.02	−0.03	0.15	0.75
	τ_{bad1}	τ_{bad2}	τ_{bad3}	τ_{bad4}	M_{pos}	
Reference method: negative item formulation						
τ_{bad1}	0.74	0.38	0.38	0.36	−0.36	4.07
τ_{bad2}	0.29	0.85	0.42	0.46	−0.39	4.13
τ_{bad3}	0.29	0.35	0.82	0.49	−0.36	4.09
τ_{bad4}	0.26	0.37	0.39	0.74	−0.36	4.20
M_{pos}	−0.12	−0.14	−0.13	−0.12	0.15	−0.75

†With the clear exception of the quite small correlations and covariances of M_{neg} with each of the τ_{good1}, τ_{good2}, τ_{good3} and τ_{good4}, all the estimates are significantly different from 0 at the 0.1% level. This significance testing was performed through t-statistics of estimates divided by their standard errors. The t-statistics are approximately normal (Bollen (1989), page 108, and Jöreskog and Sörbom (2004), page 103).

The variance $\hat{\sigma}^2(M) = 0.15$ ($t = 11.11$) of the method factor is smaller than the variances of the trait factors. Nevertheless, this variance indicates that the subjects differ considerably in the size of their individual method effects. The difference in the ratings on positively and negatively formulated items is for some subjects much higher than for others. This means that the two methods are definitely not equivalent measures of the same trait. In other words, each method measures its own method-specific construct, 'wellbeing as measured by positively formulated items' and 'wellbeing as measured by negatively formulated items'. Note that the variance of the individual method effects is invariant to the choice of the reference method.

The individual method effects do not correlate significantly with the true scores of the positive item scale (between $\hat{\rho}(\tau_{good2}, M_{neg}) = -0.04$ and $\hat{\rho}(\tau_{good4}, M_{neg}) = -0.10$). However, they correlate significantly with the true scores of the negative item scale (between $\hat{\rho}(\tau_{bad1}, M_{pos}) = -0.36$ to $\hat{\rho}(\tau_{bad2}, M_{pos}) = -0.39$). The higher the subjects score on the negative item scale, the larger (meaning large negative scores on the factor M_{pos}) the method effect. Thus, subjects with a high score on the negative item scale tend to differ more in their ratings to the two item scales than subjects with a low score. However, there is no correlation between the scores on the positive item scale and the scores on M_{neg}. The size of the method effect is not (linearly) related to the level of the mood measured by the positive item scale.

The analyses in the preceding section establish that there are considerable interindividual differences in the individual method effects and that the factor M_{pos} correlates with the latent state variables of the negative item scale. Since the positively and negatively formulated items are unsystematically allocated in the questionnaire, we can rule out that sequence effects are responsible for the interindividual differences in the individual method effects. We are not aware of alternative explanations that could be responsible for the method effects and have not been ruled out by the unsystematic distribution of positively and negatively formulated items.

Therefore, in this application 'method effects' is not just a technical term for the true score differences that are described in equation (7). Instead, these differences can in fact be interpreted as effects of assessing wellbeing with the negative item instead of the positive item scale. In other words, the method effects may be interpreted as individual causal effects. Hence, it is meaningful to look for variables explaining the interindividual differences in the individual method effects.

3.3. Explaining the interindividual differences in the individual method effects

To investigate the 'nature' of the method effects, explanatory variables were included in the model. Note that this part of the analysis is not theory driven. The main purpose here is to demonstrate the inclusion of explanatory variables in the model. Here we present only empirical findings. First, for 'openness', a scale of the Freiburg personality inventory was included in the model as a manifest explanatory variable for the trait factors and the method effect. The path diagram of the model is depicted in Fig. 5. In Fig. 5 the presentation of the measurement models has been omitted.

The model was fitted twice, using each kind of item formulation as the reference method. The model fit is the same for both models ($\chi^2(19) = 25.05$; $p = 0.159$; RMSEA $= 0.025$). The standardized regression coefficients of the regression of the trait and method factors on openness are presented in Table 2.

Openness explains a small amount (3%) of the variance of the individual method effects (standardized $\beta = 0.18$). The more open someone is, the larger is the individual method effect, i.e. the greater is the difference between his answers to the items of the two different item formulations. However, the effect is quite small.

Therefore, we added another predictor: 'emotionality', a scale of the Freiburg personality inventory. The model with openness and emotionality (see Fig. 6) as explanatory variables was fitted twice: once using the positive item scale and once using the negative item scale as the reference method. Again, the model fit is the same for both models ($\chi^2(19) = 29.90$; $p = 0.121$;

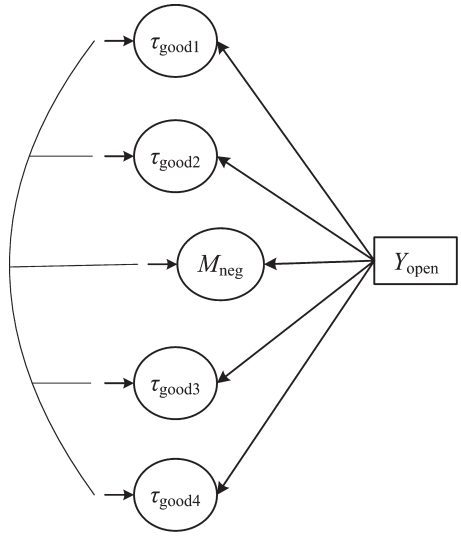

Fig. 5. Method effect model for two scales of wellbeing with the explanatory variable openness (open): the scale of the positively formulated items is the reference method; for simplicity the measurement model has been omitted

Table 2. Method effect model with the explanatory variable openness: standardized regression coefficients†

Results for reference method positive item formulation		*Results for reference method negative item formulation*	
Variable	*Openness*	*Variable*	*Openness*
τ_{good1}	0.10‡	τ_{bad1}	0.17§
τ_{good2}	0.14§§	τ_{bad2}	0.20§
τ_{good3}	0.10	τ_{bad3}	0.17§
τ_{good4}	0.04	τ_{bad4}	0.11‡
M_{neg}	0.18§	M_{pos}	−0.18§

†The standard normal approximation is used to indicate a significant deviation from 0 (Bollen (1989), page 108; Jöreskog and Sörbom (2004), page 103).
‡$|t| > 1.96$, where t is the parameter estimate divided by its standard error.
§$|t| > 2.58$.
§§$|t| > 3.29$.

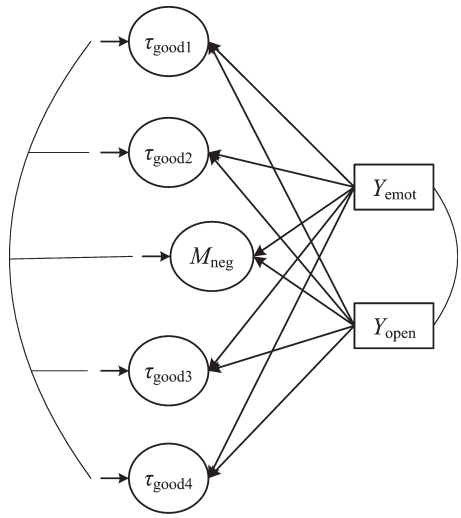

Fig. 6. Method effect model for two scales of wellbeing with the explanatory variables openness (open) and emotionality (emot): the scale of the positively formulated items is the reference method; for simplicity, the measurement model has been omitted

RMSEA = 0.027). The two personality scales correlate positively with each other ($\hat{\rho} = 0.42$). Subjects with high scores on emotionality also tend to have high scores on openness. The standardized partial regression coefficients of the regression of the trait and method factors on openness and emotionality are presented in Table 3.

Controlling for emotionality, openness does not explain the individual method effects ($\beta =$ 0.07, which is not statistically significant at the 0.05 level). Controlling for emotionality, openness also does not explain the trait factors (the standardized partial regression coefficients range from $\beta = -0.10$ to $\beta = 0.03$, all of them being non-significant at the 0.05 level).

Table 3. Method effect model with the explanatory variables openness and emotionality: standardized partial regression coefficients†

Results for reference method positive item formulation			*Results for reference method negative item formulation*		
Variable	*Openness*	*Emotionality*	*Variable*	*Openness*	*Emotionality*
τ_{good1}	−0.05	0.36‡	τ_{bad1}	−0.01	0.45‡
τ_{good2}	0.00	0.34‡	τ_{bad2}	0.03	0.41‡
τ_{good3}	−0.03	0.34‡	τ_{bad3}	0.00	0.40‡
τ_{good4}	−0.10	0.34‡	τ_{bad4}	−0.06	0.42‡
M_{neg}	0.07	0.26‡	M_{pos}	−0.07	−0.26‡

†The standard normal approximation is used to indicate a significant deviation from 0 (Bollen (1989), page 108; Jöreskog and Sörbom (2004), page 103).
‡$|t| > 3.29$, where t is the parameter estimate divided by its standard error.

However, controlling for openness, emotionality turned out to explain both method factor and trait factors. It explains significantly the variance of the method factor (standardized $\beta = 0.26$). The higher a subject scores on emotionality, the larger is his individual method effect. The standardized partial regression coefficients in the regression of the true scores of the negative item scale on emotionality (ranging from $\beta = 0.40$ to $\beta = 0.45$) are larger than those of the true scores of the positive item scale (ranging from $\beta = 0.34$ to $\beta = 0.36$). The negative item scale is more related to emotionality than the positive item scale.

Summarizing the results, there are considerable method effects. The subjects score their well-being on average much higher on the negative item than on the positive item scale. There are considerable interindividual differences in the individual effects. The size of these method effects is related to the true scores on the negative item scale. The smaller the method effect (i.e. the higher the negative scores on M_{pos} and the smaller the positive scores on M_{neg}), the higher the subjects score on the negative item scale. However, the method effects are not correlated with the scores on the positive item scale. Because alternative explanations of the individual method effects can be ruled out by design, the method effects may be interpreted as individual causal effects of using the negative item instead of the positive item scale. Emotionality explained a significant amount of the variance of the method factor. Note that, since the inclusion of the explanatory variables was exploratory, a cross-validation needs to be performed to verify the results.

4. Discussion

In this paper, we defined a method effect of individual u as the difference $\tau_{tj}(u) - \tau_{t1}(u)$ between his true scores of Y_{tj} and Y_{t1}, where Y_{tj} and Y_{t1} are the manifest variables measuring trait t with method j and method 1 respectively. This idea has also been presented by Lischetzke *et al.* (2002) (also see Eid *et al.* (2007)) and emerged from the 'true change models' (Steyer, Eid and Schwenkmezger, 1997) in which the differences between two true score variables have been used to introduce latent change variables into structural equation modelling. Defining individual method effects as latent difference scores not only clearly defines a method effect, but it also allows us to estimate mean method effects as well as correlations between method effects and traits. In contrast with the latent difference approach of Lischetzke *et al.* (2002), in the method effect model the method effects are related to a causality theory and may (under certain conditions) be interpreted as causal effects, allowing a substantial interpretation.

4.1. Interpreting the method effects as causal effects

Just like changes in true scores over time, differences in true scores between different methods can be due to many causes. To be able to attribute a true change to an *intervention between two time points*, Steyer (2005) introduced a control group with no intervention and argued that all alternative explanations of the change in the true scores in the intervention group can be ruled out if there is no change in the true scores of the (untreated) control group. Only in this case can the changes in the true scores in the intervention group be interpreted as the individual causal effects of the intervention.

In a similar vein, we suggest in this paper that method effects should be regarded as *causal effects* of the methods on the manifest measures. Conceptually, if method j is used *instead of* method k, the person-specific distribution of the manifest variable, including its expected value (the true score) might be different. It is exactly that difference in the individual expected values of the outcome variable between treatment j and treatment k which defines an *individual causal effect* in the Neyman–Rubin approach to causality (see Steyer *et al.* (2007) for details).

This theory of causal effects refers to using method j *instead of* method k. It does not refer to using *both* methods for the same subject, method j *and* method k. Nevertheless, in repeated measures designs (such as repeatedly assessing a trait with different methods), the scores of the method factor may be interpreted as individual *causal* effects of using method j instead of method k, provided that we are ready to make assumptions, ruling out *alternative explanations* (Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish *et al.*, 2002).

Differences between the true scores of two manifest variables Y_{tj} and Y_{tk} that are assessed by methods j and k respectively might, for example, not be due to using different methods but to *sequence effects*. For instance, Orthner (2004) has shown that the answer to an item may depend on the response to previous items. Also effects of the length of the test might be an alternative explanation, especially when the different methods are applied in sequence. Kraut *et al.* (1975) found an increasing use of modal (*versus* extreme) answer categories at the end of a test compared with the beginning of the test.

Hence, emphasis should be put on the design of the study, to rule out these alternative explanations. If, as in the application that was presented here, the positive and negative item scales are not presented in a systematic sequence (e.g. first the positive item; then the negative item scale), sequence and test length effects will be ruled out as alternative explanations for the method effects, and it makes sense to search for variables explaining the interindividual differences in the method effects. If, in contrast, sequence effects are a viable alternative interpretation of the interindividual differences in the true scores, then these effects will call for completely different explanations.

As mentioned before, even if the method effects cannot be interpreted as individual causal effects, the method effect model can still be applied. However, in this case the method effects may only be interpreted as latent difference scores; not as causal effects. This is unproblematic as long as the main focus is not the method effects, but the relationships between the trait factors and other latent or manifest variables. In these applications, the method factors are only useful for improving the model fit.

4.2. Strengths and limitations of the method effect model

In contrast with previous models, both the method effect model and the CT-C(M-1) model provide a unique definition of the method factors, and in both models the trait factors represent the true scores of the variables measuring a trait by the reference method.

In the method effect model, in contrast with other models, the definition of the method effects is based on theoretical considerations of the nature of method effects and is fully in line with the definition of individual causal effects in the Neyman–Rubin tradition of causality. We may estimate mean method effects as well as the covariation between trait and method factors. Furthermore, in the method effect model the model fit is invariant to the choice of the reference method.

Although in the method effect model the variance cannot additively be decomposed into trait-, method- and error-specific variance, convergent validity can still be evaluated. Convergent validity is assured when the mean and the variance of the method effects is small or even 0. The larger the relationship of the variance of the method effects to the variance of the trait factors and the larger the mean of the method effects (relative to the standard deviation of the traits), the less convergent validity is obtained.

Although the method effect model has many advantages, there are also limitations. First, a reference method must be chosen and the trait factor represents the true scores of the manifest variables measuring the trait *with the reference method*. This means that the trait factors in the method effect model do not represent *common* trait factors. Instead, the trait factors are *specific to the reference method*. This may be convenient for some applications, e.g. when comparing self-ratings with peer, parent or superior ratings, where there is a 'natural' reference method. But even in such a case we strive for a trait factor representing the *common trait*, not the self-rating-specific trait. Remember, the most important reason for using different methods is to increase the construct validity of the traits! Therefore, a reference-method-specific trait is not really what we are interested in. In most applications, there is no 'natural' reference method with undoubted validity. Instead, we must conceptualize the trait as some kind of aggregate of the different methods representing the trait. This is not possible with the method effect model that is presented in this paper.

The method effect model with fixed loadings that is presented here is very restrictive. It presumes the same scaling for the manifest variables and invariant method effects across different traits. Extensions of the method effect model with less restrictive assumptions are possible. The elaboration of these extended models is one of the future research tasks.

Finally, the applicability of the method effect model in empirical research needs further tests. First applications and extensions have been presented by Vautier *et al.* (2007), Vautier and Pohl (2007) and Kraus (2006). Vautier *et al.* (2007) presented a model with both true change variables *and* method effects. Vautier and Pohl (2007) investigated the bipolarity of true change scores of anxiety measured with different methods. Kraus (2006) applied the method effect model to the Mount (1984) data and conducted a first simulation study investigating the convergence behaviour of the method effect model.

The causal interpretation of the method effects in the method effect model, that may be attempted if alternative explanations can be ruled out, invites further investigation of the nature of these method effects. Explanatory variables may be included in the model to explain the interindividual differences in the method effects. Further empirical research should pay regard to the explanation of the method effects to obtain an insight into the 'causes' that they represent.

Appendix A: Identification of the method effect model

The method effect model for the example in Section 2.2 implies the following variance, covariance and mean structures:

$$\mathrm{var}(Y_{11}) = \sigma^2_{\tau_{11}} + \sigma^2_{\varepsilon_{11}}, \tag{23}$$

$$\mathrm{var}(Y_{12}) = \sigma^2_{\tau_{11}} + \sigma^2_{M_2} + 2\sigma_{\tau_{11},M_2} + \sigma^2_{\varepsilon_{12}}, \tag{24}$$

$$\mathrm{var}(Y_{21}) = \sigma^2_{\tau_{21}} + \sigma^2_{\varepsilon_{21}}, \tag{25}$$

$$\mathrm{var}(Y_{22}) = \sigma^2_{\tau_{21}} + \sigma^2_{M_2} + 2\sigma_{\tau_{21},M_2} + \sigma^2_{\varepsilon_{22}}, \tag{26}$$

$$\mathrm{cov}(Y_{11}, Y_{12}) = \sigma^2_{\tau_{11}} + \sigma_{\tau_{11},M_2}, \tag{27}$$

$$\mathrm{cov}(Y_{11}, Y_{21}) = \sigma_{\tau_{11},\tau_{21}}, \tag{28}$$

$$\mathrm{cov}(Y_{11}, Y_{22}) = \sigma_{\tau_{11},\tau_{21}} + \sigma_{\tau_{11},M_2}, \tag{29}$$

$$\mathrm{cov}(Y_{12}, Y_{21}) = \sigma_{\tau_{11},\tau_{21}} + \sigma_{\tau_{21},M_2}, \tag{30}$$

$$\mathrm{cov}(Y_{12}, Y_{22}) = \sigma_{\tau_{11},\tau_{21}} + \sigma_{\tau_{11},M_2} + \sigma_{\tau_{21},M_2} + \sigma^2_{M_2}, \tag{31}$$

$$\mathrm{cov}(Y_{21}, Y_{22}) = \sigma^2_{\tau_{21}} + \sigma_{\tau_{21},M_2}, \tag{32}$$

$$E(Y_{11}) = \mu_{\tau_{11}}, \tag{33}$$

$$E(Y_{12}) = \mu_{\tau_{11}} + \mu_{M_2}, \tag{34}$$

$$E(Y_{21}) = \mu_{\tau_{21}}, \tag{35}$$

$$E(Y_{22}) = \mu_{\tau_{21}} + \mu_{M_2}. \tag{36}$$

The assumption that the method effects are the same across all traits (see equation (13)) leads to the following restriction on the mean structure:

$$E(Y_{22}) - E(Y_{21}) = E(Y_{12}) - E(Y_{11}) = \mu_{M_2}. \tag{37}$$

Looking at equations (23)–(36), it can be easily seen how to identify the theoretical parameters from the empirical variances, covariances and means. For example, the covariance $\sigma_{\tau_{11},\tau_{21}}$ can be identified by the covariance between the manifest variables Y_{11} and Y_{21} (see equation (28)) and the covariance σ_{τ_{11},M_2} by the difference $\mathrm{cov}(Y_{11}, Y_{22}) - \mathrm{cov}(Y_{11}, Y_{21})$ (see equations (28) and (29)). Similar identification equations can easily be found for all 13 theoretical parameters in the model. Suffice it to say that altogether the model has 13 theoretical parameters (three variances and three covariances of the latent variables, four measurement error variances and three expected values of the latent variables), which can be identified by 14 empirical variances, covariances and means. There are no restrictions on the variances and covariances and one restriction on the means in the design with two traits and two methods that is presented in Fig. 3. Note that in models with more than two traits or more than two methods there *will* be restrictions also on the covariances of the manifest variables. In a design with three traits and two methods or with two traits and three methods, for example, there will be 20 theoretical parameters and 27 empirical variances, covariances and means. In these designs there are five restrictions on the variances and covariances, and two restrictions on the expected values.

Appendix B: LISREL syntax for the method effect model in the application

```
TI Method effect model (Reference method: positively formulated items)
DA NI=8 NO=499 MA=CM

CM=good.cm
ME=good.me

LA good1 bad1 good2 bad2 good3 bad3 good4 bad4

MO NY=8 NE=5 LY=FU,FI PS=SY,FR TE=DI,FR TY=FI AL=FR
LE good1 good2 good3 good4 ME
```

```
VA 1 LY(1,1) LY(2,1)                    !good1
VA 1 LY(3,2) LY(4,2)                    !good2
VA 1 LY(5,3) LY(6,3)                    !good3
VA 1 LY(7,4) LY(8,4)                    !good4
VA 1 LY(2,5) LY(4,5) LY(6,5) LY(8,5)    !ME

PD OU AD=OFF ALL
```

References

Barrett, L. F. and Russell, J. A. (1998) Independence and bipolarity in the structure of current affect. *J. Personlty Socl Psychol.*, **74**, 967–984.

Bollen, K. A. (1989) *Structural Equations with Latent Variables*. Oxford: Wiley.

Browne, M. W. and Cudeck, R. (1993) Alternative ways of assessing model fit. In *Testing Structural Equation Models* (eds K. A. Bollen and J. S. Long), pp. 136–162. Newbury Park: Sage.

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

Campbell, D. T. and Stanley, J. C. (1963) Experimental and quasiexperimental designs for research on teaching. In *Handbook on Research on Teaching* (ed. N. Gage), pp. 171–246. Chicago: Rand McNally.

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

Conway, J. M., Lievens, F., Scullen, S. E. and Lance, C. E. (2004) Bias in the correlated uniqueness model for MTMM data. *Struct. Equn Modlng*, **11**, 535–559.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

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

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

Eid, M., Nussbeck, F., Geiser, C., Cole, D. A., Gollwitzer, M. and Lischetzke, T. (2007) Analyzing multitrait-multimethod data with structural equation modeling: some guidelines for selecting an appropriate model. *Manuscript*. To be published.

Fahrenberg, J., Hampel, R. and Selg, H. (1984) *Das Freiburger Persönlichkeitsinventar (FPI und FPI-R): Handbuch*, 4th edn. Göttingen: Hogrefe.

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. *J. Appl. Psychol.*, **63**, 579–588.

Horan, P. M., DiStefano, C. and Motl, R. W. (2003) Wording effects in self-esteem scales: methodological artifact or response style? *Struc. Equn Modlng*, **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 *Contemporary Developments in Mathematical Psychology*, vol. 2 (eds R. Atkinson, D. Krantz, R. Luce and P. Suppes), pp. 1–56. San Francisco: Freeman.

Jöreskog, K. G. and Sörbom, D. (2004) *LISREL 8.7 for Windows*. Lincolnwood: Scientific Software International.

Kenny, D. A. (1976) An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *J. Expertl Socl Psychol.*, **12**, 247–252.

Kraus, K. (2006) Analysis of multitrait-multimethod data based on the theory of individual causal effects. *Master's Thesis*. Friedrich-Schiller-Universität, Jena.

Kraut, A. I., Wolfson, A. D. and Rothenberg, A. (1975) Some effects of position on opinion survey items. *J. Appl. Psychol.*, **60**, 774–776.

Lewin, L. M., Hops, H., Davis, B. and Dishion, T. J. (1993) Multimethod comparison of similarity in school adjustment of siblings and unrelated children. *Devlpmntl Psychol.*, **29**, 963–969.

Lischetzke, T., Eid, M. and Nussbeck, F. (2002) Unterschiedliche Definitionen von Methodeneffekten in MTMM Modellen und ihre Implikationen für die Analyse der Validität. *43rd Meet. German Psychological Association, Berlin*.

Lord, F. M. and Novick, M. R. (1968) *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.

Marsh, H. W. (1989) Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions. *Appl. Psychol. Measmnt*, **13**, 335–361.

Marsh, H. W. (1996) Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *J. Personlty Socl Psychol.*, **70**, 810–819.

Marsh, H. W. and Byrne, B. M. (1993) Do we see ourselves as others infer: a comparison of self-other agreement on multiple dimensions of self-concept from two continents. *Aust. J. Psychol.*, **45**, 49–58.

Marsh, H. W. and Craven, R. G. (1991) Self-other agreement on multiple dimensions of preadolescent self-concept: inferences by teacher, mothers, and fathers. *J. Educ. Psychol.*, **83**, 393–404.
Marsh, W. H. and Grayson, D. (1995) Latent variable models of multitrait-multimethod data. In *Structural Equation Modeling: Concepts, Issues and Applications* (ed. R. Hoyle), pp. 177–198. Thousand Oaks: Sage.
Matthews, G. and Deary, I. J. (1998) *Personality Traits*. Cambridge: Cambridge University Press.
McConnell, A. R. and Leibold, J. M. (2001) Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *J. Exptl Socl Psychol.*, **37**, 435–442.
Motl, R. W. and DiStefano, C. (2002) Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Struct. Equn Modlng*, **9**, 562–578.
Mount, M. K. (1984) Psychometric properties of subordinate ratings of managerial performance. *Personn. Psychol.*, **37**, 687–702.
Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, sect. 9. *Statist. Sci.*, **5**, 465–472.
Novick, M. R. (1966) The axioms and principal results of classical test theory. *J. Math. Psychol.*, **3**, 1–18.
Orthner, T. M. (2004) On changing the positions of items in personality questionnaires analysing effects of item sequence using IRT. *Psychol. Sci.*, **46**, 446–476.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
Russell, J. A. and Carroll, J. M. (1999) On the bipolarity of positive and negative affect. *Psychol. Bull.*, **125**, 3–30.
Schmitt, M. (2006) Conceptual, theoretical and historical foundations of multimethod assessment. In *Handbook of Multimethod Measurement in Psychology* (eds M. Eid and E. Diener), pp. 9–25. Washington DC: American Psychological Association.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
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 *International Encyclopedia of the Social and Behavioural Sciences: Logic of Inquiry and Research Design* (eds C. Ragin and T. Cook), pp. 481–520. Oxford: Pergamon.
Steyer, R. (2005) Analyzing individual and average causal effects via structural equation models. *Methodology*, **1**, 39–54.
Steyer, R., Eid, M. and Schwenkmezger, P. (1997) Modeling true intraindividual change: true change as a latent variable. *Meth. Psychol. Res. Online*, **2**, no. 1.
Steyer, R., Ferring, D. and Schmitt, M. (1992) States and traits in psychological assessment. *Eur. J. Psychol. Assessmnt*, **8**, 79–98.
Steyer, R., Partchev, I., Kröhne, U., Nagengast, B. and Fiege, C. (2007) Causal effects in between-group experiments and quasi-experiments: theory. *Manuscript*. Friedrich-Schiller-Universität Jena, Jena. (Available from `http://www.causal-effects.de/`.)
Steyer, R. and 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 *Recent Developments on Structural Equation Modeling: Theory and Applications* (eds K. V. Montfort, H. Oud and A. Satorra), pp. 197–220. Amsterdam: Kluwer.
Steyer, R., Schmitt, M. and Eid, M. (1999) Latent state-trait theory and research in personality and individual differences. *Eur. J. Personlty*, **13**, 389–408.
Steyer, R., Schwenkmezger, P., Eid, M. and Notz, P. (1991) Befindlichkeitsmessung und Latent-State-Trait-Modelle. *Arbeitsbericht, DFG-Projekt STE 411/3-1*. University of Trier, Trier.
Steyer, R., Schwenkmezger, P., Notz, P. and Eid, M. (1997) *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)*. Göttingen: Hogrefe.
Steyer, R., Schwenkmezger, P., Notz, P. and Eid, M. (2004) *Entwicklung des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF): Primärdatensatz*. Trier: Psychologisches Datenarchiv PsychData des Zentrums für Psychologische Information und Dokumentation.
Vautier, S. and Pohl, S. (2007) Bipolarity and method effects in STAI scores. *Manuscript.* To be published.
Vautier, S., Steyer, R. and Boomsma, A. (2007) A true-change model with individual method effects: reliability issues. *Br. J. Math. Statist. Psychol.*, to be published.
Vautier, S., Steyer, R., Jmel, S. and Raufaste, E. (2005) Imperfect or perfect dynamic bipolarity?: the case of antonymous affective judgments. *Struct. Equn Modlng*, **12**, 391–410.
Villar, P., Luengo, M. A., Gómez-Fraguela, J. A. and Romero, E. (2006) Assessment of validity of parenting constructs using the multitrait-multimethod model. *Eur. J. Psychol. Assessmnt*, **22**, 59–68.
Widaman, K. F. (1985) Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Measmnt*, **9**, 1–26.
Zimmerman, D. W. (1975) Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, **40**, 395–412.