Erschienen in: Multivariate Behavioral Research ; 51 (2016), 1. - S. 67-85
https://dx.doi.org/10.1080/00273171.2015.1101367

A Cross-Classified CFA-MTMM Model for Structurally Different and Nonindependent Interchangeable Methods

Tobias Koch[a], Martin Schultze[b], Minjeong Jeon[c], Fridtjof W. Nussbeck[d], Anna-Katharina Praetorius[e], and Michael Eid[b]

[a]Leuphana Universität Lüneburg; [b]Freie Universität Berlin; [c]The Ohio State University; [d]Universität Bielefeld; [e]German Institute for International Educational Research

ABSTRACT
Multirater (multimethod, multisource) studies are increasingly applied in psychology. Eid and colleagues (2008) proposed a multilevel confirmatory factor model for multitrait-multimethod (MTMM) data combining structurally different and multiple independent interchangeable methods (raters). In many studies, however, different interchangeable raters (e.g., peers, subordinates) are asked to rate different targets (students, supervisors), leading to violations of the independence assumption and to cross-classified data structures. In the present work, we extend the ML-CFA-MTMM model by Eid and colleagues (2008) to cross-classified multirater designs. The new C4 model (Cross-Classified CTC[M-1] Combination of Methods) accounts for nonindependent interchangeable raters and enables researchers to explicitly model the interaction between targets and raters as a latent variable. Using a real data application, it is shown how credibility intervals of model parameters and different variance components can be obtained using Bayesian estimation techniques.

KEYWORDS
Bayesian analysis; cross-classification; MTMM modeling; structurally different and interchangeable methods

Introduction

A growing body of research is devoted to multirater (multimethod, multisource) measurement designs (Campbell & Fiske, 1959; Eid & Diener, 2006; Kenny, 1995). In organizational psychology, reports from multiple informants (e.g., supervisors, subordinates, colleagues) are commonly used to gather information of a target's behavior (Conway & Huffcutt, 1997; Ghorpade, 2000; Mahlke et al., 2015; Yammarino, 2003; Yammarino & Atwater, 1997). In educational and developmental psychology, reports from teachers, students, and peers are often collected to obtain deeper insights into a child's ability or behavior (Bull, Schultze, & Scheithauer, 2009; Pham et al., 2012; Schultze, 2012).

Multirater measurement designs bear many advantages. For example, they are more informative than single-rater (method) designs as they allow researchers to quantify the amount of (dis)agreement among different types of raters, model traits, and rater effects as latent variables and relate them to external variables to identify potential causes of trait and method effects (see also Eid, Lischetzke, & Nussbeck, 2006; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Koch, Eid, & Lochner, in press).

Despite the growing interest in modeling multirater data, researchers sometimes struggle choosing an appropriate multirater-multimethod (MTMM) model. One important factor for analyzing multirater data properly is the type of raters used in the particular MTMM design (see Eid et al., 2008). According to Eid et al. (2008), measurement designs can incorporate (a) interchangeable (or random) raters, (b) structurally different (fixed) raters, or (c) a combination of structurally different and interchangeable raters. Interchangeable raters are raters that stem from a common rater pool for each target. Consider, for example, multiple peer ratings of student empathy or multiple colleague ratings of supervisors' leadership quality. Due to the sampling procedure, measurement designs with interchangeable raters imply a multilevel data structure (i.e., interchangeable raters are nested within targets, see Eid et al., 2008). By contrast, structurally different raters cannot easily be replaced by one another, given that they do not belong to a common pool of raters, but differ with respect to their role or relation with the target (e.g., student self-reports, parent reports). Hence, they may have fundamentally different perspectives and information about the target's behavior (e.g., physiological measures vs. self-reports vs. implicit measures). Eid et al. (2008) proposed different CFA-MTMM models for measurement designs with structurally different, interchangeable, and a combination

CONTACT Tobias Koch tobias.koch@uni-leuphana.de Center for Methods, Leuphana Universität Lüneburg, Scharnhorststr. 1, D-21335 Lüneburg, Germany.

Table 1. Multirater data structure of fully nested and cross-classified interchangeable ratings.

	(a) Fully nested interchangeable ratings								
	Rater 1A	Rater 2A	Rater 3A	Rater 1B	Rater 2B	Rater 3B	Rater 1C	Rater 2C	Rater 3C
Target A	×	×	×	—	—	—	—	—	—
Target B	—	—	—	×	×	×	—	—	—
Target C	—	—	—	—	—	—	×	×	×
	(b) Cross-classified interchangeable ratings								
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9
Target A	×	—	×	—	—	×	—	×	×
Target B	—	×	—	×	×	—	×	×	—
Target C	×	—	×	—	×	—	×	—	×

Note: Panel (a) refers to a fully nested multirater data structure. Each target (A, B, and C) is rated by three interchangeable raters (e.g., 1A, 2A, and 3A), which are randomly drawn from a target-specific rater pool. Panel (b) refers to a cross-classified multirater data structure. Each target (A, B, and C) is rated by five interchangeable raters, which are randomly drawn from a common rater pool containing nine raters in total. The ratings are crossed with targets and raters. × = observation; — = no observation.

of structurally different and independent interchangeable raters.

In this article, we focus on the general multiple indicator CFA-MTMM model for the combination of structurally different and interchangeable raters. The model is called the Multilevel-Correlated-Trait-Correlated-Methods-Minus-One (ML-CTC[M-1]) model (Eid et al., 2008). This model and has been successfully applied to various areas of psychology (Carretero-Dios, Eid, & Ruch, 2011; Danay & Ziegler, 2011; Pham et al., 2012). The ML-CTC(M-1) model enables researchers to specify method factors on a rater-specific and a target-specific level and study the convergent validity (or rater consensus), the method specificity (or rater-specific effects), as well as the reliability of the given measures. Another advantage of the ML-CTC(M-1) model is that the latent variables are defined according to stochastic measurement theory and have a clear and unambiguous meaning (Eid et al., 2008; Koch, Schultze, Eid, & Geiser, 2014, Koch et al., in press).

Notwithstanding, the ML-CTC(M-1) model cannot be used for all MTMM designs. In particular, the model assumes a fully nested multirater data structure (Meiser & Steinwascher, 2014; Schultze, Koch, & Eid, 2015), meaning that the rating of (or scores from) the interchangeable raters are assumed to be independent across different targets. However, when different targets are rated by the same raters (e.g., students being rated by the same teachers), this independence assumption is violated (Meiser & Steinwascher, 2014; Schultze et al., 2015). Table 1 shows the sampling procedure of multirater data with fully nested and cross-classified interchangeable ratings.

Multirater data structures, as represented in Table 1, Panel (b), occur often in practice. For example, in educational and developmental research, it is quite common to collect multiple peer ratings of students' empathy level (see, e.g., Bull, Schultze, Geiser, & Scheithauer, 2013) or multiple students' ratings for the assessment of teaching quality (see, e.g., Marsh & Roche, 1997). In organizational research, ratings from several subordinates or colleagues are used for the assessment of supervisors' leadership quality (see, e.g., Mahlke et al., 2015). In cases like this, it is often not possible to sample from a target-specific set of interchangeable raters, but rather from a common population of interchangeable raters. Thus, it is likely that different targets (e.g., students or supervisors) will be rated by the same persons (e.g., peers or subordinates), resulting in partially crossed structures. Fully crossed structures are often found in experimental designs, where each target is rated by the same set of raters (see, e.g., Shrout, 1995).

Cross-classified multirater data violate the independence assumption made in conventional multilevel CFA-MTMM models and therefore require specific modeling approaches that have not yet been presented. In a recent simulation study by Schultze et al. (2015), the effect of nonindependent interchangeable raters on parameter estimates and standard errors in ML-CTC(M-1) models was examined. The results of this simulation study indicated that the parameter estimates are well recovered in the ML-CTC(M-1) models even in cases in which the assumption of independent interchangeable raters is violated. This means that the estimates of convergent validity, discriminant validity, method specificity, and reliability are trustworthy. However, the standard errors of the level-1 covariance matrix and the level-2 mean structure are biased under such circumstances. In particular, the standard error bias is higher in cases of few level-1 units (i.e., fewer than 10 raters per target). Oftentimes, multirater designs incorporate only a few interchangeable raters per target (e.g., two to three team member ratings, colleague ratings, or friend ratings) or are carried out to investigate treatment effects or trait changes over time, as for example, in longitudinal MTMM designs (Koch, 2013). In these cases, the accuracy of the standard error estimates (especially for the mean structure) is critical.

Current approaches for analyzing cross-classified and MTMM data

Research devoted to the analysis of cross-classified data has mainly focused on methods such as analysis of variances (e.g., Gaugler & Akritas, 2011; Sahai & Ojeda, 2005; Searle, 2006; Searle, Casella, & McCulloch, 2009), hierarchical linear models (e.g., Beretvas, 2011; Fielding & Goldstein, 2006; Goldstein, 1994, 2011; Raudenbush & Bryk, 2002; Raudenbush, 1993), models based on generalizability theory (Hoyt, 2000), and latent variable approaches such as confirmatory factor analysis or structural equation models (Asparouhov & Muthén, 2012). The two latter approaches using latent variables have also been suggested for the analysis of multitrait-multimethod (MTMM) data. Today, a multitude of different CFA-MTMM models exist, including traditional, single-level MTMM models such as the correlated traits–correlated uniqueness model (CT-CU; Kenny, 1976), the correlated traits–uncorrelated methods model (CT-UM; Marsh & Grayson, 1995), or the correlated traits–correlated methods model (CT-CM, Marsh & Grayson, 1995), as well as more recently developed approaches such as the correlated traits–correlated methods-minus-one (CT-C[M-1]) model (Eid, 2000), the latent difference model (Pohl, Steyer, & Kraus, 2008), or the latent means model (Pohl & Steyer, 2010). These single-level (traditional) CFA-MTMM models have been extensively discussed in the literature (Castro-Schilo, Widaman, & Grimm, 2013; Dumenci, 2000; Eid, 2000; Eid et al., 2003, 2008; Geiser, Eid, & Nussbeck, 2008, Geiser, Eid, West, Lischetzke, & Nussbeck, 2012, Geiser, Koch, & Eid, 2014; Kenny & Kashy, 1992; Koch et al., in press; Pohl & Steyer, 2010; Widaman, 1985).

In this study, we concentrate on multilevel (or multirater) modeling approaches that became increasingly popular over the past two decades. In fact, several multilevel CFA-MTMM models have been proposed so far (see, e.g., Bauer et al., 2013; Eid et al., 2008; Hox & Maas, 2006; Koch et al., 2014; Maas, Lensvelt-Mulders, & Hox, 2009; Mahlke et al., 2015). However, these models have been primarily developed for analyzing purely hierarchical (or fully nested) data structures, but not for complex cross-classified MTMM measurement designs.

This study complements previous research by combining cross-classified multilevel models and CFA-MTMM analysis. Up to this point, there are only a few studies addressing the conceptual similarities between MTMM designs and cross-classified designs (see Hoyt, 2000; Jeon & Rijmen, 2014). Jeon and Rijmen (2014) clarified that classical MTMM designs can be conceived as a form of cross-classified designs, in which traits and method are fully crossed, and discussed different robust (maximum likelihood–based) estimation approaches. Hoyt (2000) presented a model based on multivariate generalizability theory. Hoyt's approach allows researchers to adjust for different types of rater bias in cross-classified multirater designs by defining variance coefficients that represent target (called universe variance), rater (called rater variance), and target-rater interaction (called dyadic variance) effects. However, Hoyt (2000) and Jeon and Rijmen (2014) did not show how cross-classified MTMM measurement designs combining structurally different and interchangeable methods can be properly analyzed.

Here, we will present the extension of a multilevel CFA-MTMM model to a between-person level cross-classified structure. The model will be defined on the basis of stochastic measurement theory (Steyer, 1989; Steyer & Eid, 2001; Zimmermann, 1975), which enables researchers to specify target, rater, and target-rater interaction effects as random variables. Moreover, we will focus on the ML-CTC(M-1) model proposed by Eid et al. (2008), which has been particularly designed for the analysis of MTMM designs combining structurally different and interchangeable methods.

Aims of the present study

The aim of the present work is to extend the ML-CTC(M-1) model to MTMM designs combining structurally different and nonindependent interchangeable raters. The new model is called the C4 model (cross-classified CTC[M-1] combination of methods). The C4 model allows researchers to explicitly model the true (i.e., free of measurement error) rater-target interaction effects as latent residual variables in addition to trait effects and two-rater effects. Moreover, the C4 model allows researchers to study different variance components (i.e., consistency, method specificity, rater-target interdependency, and reliability). The new model will be formulated for continuous observed variables and will be applied to real data from an educational intervention study in which multiple (interchangeable) teachers were asked to rate multiple students. Due to the measurement design, the ratings (i.e., teacher reports for each student) were nested within students and teachers. For reasons of brevity, we do not discuss all possible extensions of the C4 model (e.g., to three-level cross-classified data, longitudinal data, or categorical outcomes). Instead, we consider cross-sectional multirater designs with continuous observed variables, in which multiple interchangeable raters are assumed to stem from one common rater pool as this increases the likelihood that interchangeable raters rate multiple targets. The advantages and limitations of the new model are addressed, and suggestions for future research are made.

Extending the ML-CTC(M-1) model to cross-classified structures

In this section, we describe how the ML-CTC(M-1) model by Eid et al. (2008) can be extended to multirater data structures combining structurally different and non-independent interchangeable raters. The new C4 model enables researchers to model the interaction between a target and a rater as latent variable. In contrast to traditional modeling approaches that define latent interaction effects as multiplicative effects (see Kelava, 2009; Marsh, Wen, & Hau, 2004; Moosbrugger, Schermelleh-Engel, & Klein, 1997; Schumacker & Marcoulides, 1998, for an overview), the latent interaction variable in the C4 model will be defined as a zero-mean residual variable that is uncorrelated with the target and the rater variables. In the following, the basic steps of the definition of the original ML-CFA-MTMM model by Eid et al. (2008) are repeated before the new C4 model is introduced.

Basic decomposition in fully nested multirater designs

The starting point of the model definition is the decomposition of the observed variables into a true score variable and a latent error variable. The decomposition of the observed variables pertaining to structurally different raters (e.g., self-reports) is given by

$$Y_{tijk} = \tau_{tijk} + \epsilon_{tijk} \quad \text{(structurally different raters)}. \quad (1)$$

Equation (1) states that the observed variables Y_{tijk} of target t, item i, construct j, and method k (e.g., 1 = self-reports) is decomposed into a true score variable τ_{tijk} and a measurement error variable ϵ_{tijk}. The true score variable τ_{tijk} is defined as the conditional expectation of Y_{tijk} given the target variable (i.e., p_T) and can be interpreted as the true self-rating of a target t on item i and construct j [i.e., $\mathrm{E}(Y_{tijk}|p_T)$]. The target variable p_T is a random variable, and its values are the targets. The measurement error variable ϵ_{tijk} is defined as a residual with respect to the target's true score τ_{tijk}. Hence, the error variable ϵ_{tijk} has an expectation (mean) of zero and is uncorrelated with the target's true score variable τ_{tijk}. It is noteworthy that the true score variables τ_{tijk} and the latent error ϵ_{tijk} variables are measured at the target level. This means that the self-reports of a target can only vary across targets, not across different raters (e.g., peers or colleagues). The decomposition described above (see Equation [1]) does not differ from the original ML-CTC(M-1) model proposed by Eid et al. (2008) and is in line with basic principles of classical test theory (CTT).

Next, we decompose the observed variables Y_{rtijk} pertaining to the set of independent interchangeable raters. The observed variables Y_{rtijk} are measured at level 1 (i.e., rater level) and therefore contain an additional index r for a rater. In the original ML-CTC(M-1) model (with independent sets of interchangeable raters), these observed variables are decomposed as follows:

$$Y_{rtijk} = \tau_{rtijk} + \epsilon_{rtijk} \quad \text{(independent interchangeable raters)}, \quad (2)$$

$$\tau_{rtijk} = T_{tijk} + UM_{rtijk} \quad \text{(independent interchangeable raters)}. \quad (3)$$

Equation (2) states that the observed variables are decomposed into a rater- and target-specific true score variable τ_{rtijk} and a latent error variable ϵ_{rtijk}. The rater- and target-specific true score variable τ_{rtijk} is defined as conditional expectation of the Y_{rtijk} given the target variable p_T and the rater variable p_R [i.e., $E(Y_{rtijk}|p_T, p_R)$]. The rater variable p_R is also a random variable, and its values are the raters.

The latent trait variables T_{tijk} (see Equation [3]) are defined as true expectations of all interchangeable ratings for a particular target [i.e., $E(Y_{rtijk}|p_T)$]. The unique (rater-specific) method UM_{rtijk} variables are defined as residuals with respect to the latent trait variables [i.e., $E(Y_{rtijk}|p_T, p_R) - E(Y_{rtijk}|p_T)$]. Thus, the UM_{rtijk} variables have an expectation (mean) of zero and are uncorrelated with the latent trait variable T_{tijk}. The UM_{rtijk} variables capture rater-specific effects and/or interaction effects between the target and the rater. As Eid et al. (2008) pointed out, the separation of measurement error influences ϵ_{rtijk} from unique method (rater-specific) influences UM_{rtijk} is only possible because multiple indicators i per trait method unit (TMU) are used. To be more specific, multiple indicator models allow the specification of trait-specific and unidimensional unique method factors: $UM_{rtijk} = \lambda_{ijk}^{UM} UM_{rtjk}$. In addition, it is assumed that the latent unique method factors UM_{rtjk} and the error variables ϵ_{rtijk} are independently and identically distributed (IID) across targets.

Basic decomposition in cross-classified multirater designs

In this section we explain how the original ML-CTC(M-1) model by Eid et al. (2008) can be extended to cross-classified multitrait-multirater designs. In cross-classified multitrait-multirater designs, the UM_{rtijk} variables can be decomposed further into a rater-specific variable R_{rijk} and a rater-target-specific interaction variable Int_{rtijk} (cf. Gaugler & Akritas, 2011; Hoyt, 2000):

$$UM_{rtijk} = R_{rijk} + Int_{rtijk}. \quad (4)$$

Equation (4) states that the unique method variables UM_{rtijk} contain both rater-specific components $R_{rijk} = E(Y_{rtijk}|p_R)$ and target-rater interaction $Int_{rtijk} =$

$E(Y_{rtijk}|p_T, p_R) - E(Y_{rtijk}|p_T) - E(Y_{rtijk}|p_R)$ components. In classical (pure) hierarchical data structures, however, these effects cannot be separated empirically from one another because interchangeable raters are assumed to be independent across different targets. On the contrary, cross-classified multirater designs enable researchers to model these components explicitly because multiple interchangeable raters are allowed to rate multiple targets. The above decompositions (see Equation [4]) are aligned with the measurement equation of two-way crossed random effects models for unbalanced data (see Gaugler & Akritas, 2011; Sahai & Ojeda, 2005; Searle, 2006; Searle et al., 2009). One important assumption of crossed random effect models is that the levels of each random factor (here, raters and targets) are sampled independently from one another. That is, cross-classified multirater models imply that the selection of targets is independent of the selection of raters and vice versa. According to this independence assumption, the true rating τ_{rtijk} can be decomposed into a true target T_{tijk} variable, a true rater R_{rijk} variable, and a true interaction Int_{rtijk} variable:

$$\tau_{rtijk} = T_{tijk} + R_{rijk} + Int_{rtijk} \quad \text{(nonindependent interchangeable raters).} \tag{5}$$

A proof for the additive decomposition of the true score variables has been demonstrated by Gaugler and Akritas (2011) for a model with two random factors. Our proposed model can be seen as a multiple indicator and multiple trait extension of the two-factor crossed random effects model (with two main effects and their interaction effects). Therefore, their derivation can be applied to our proposed model. The target's trait $T_{tijk} = E(Y_{rtijk}|p_T)$ can be interpreted as the expected true rating across all interchangeable raters of that particular target. Positive values indicate that a target is generally rated higher by his peers (or colleagues) than other targets. The rater-specific method $R_{rijk} = E(Y_{rtijk}|p_R)$ variable represents the expected true rating of a particular rater across all targets. Positive values indicate that a particular rater tends to overrate the target's trait as compared to all other interchangeable raters. The latent interaction effect $Int_{rtijk} = E(Y_{rtijk}|p_T, p_R) - E(Y_{rtijk}|p_T) - E(Y_{rtijk}|p_R)$ captures the part of the true ratings that can be explained neither by the target's trait T_{tijk} nor by the rater-specific method R_{rijk} effect. A value of the interaction variable reflects the expected over- or underestimation of a particular rater–target combination that is not due to the expected target and not due to the expected rater effects.

The interaction variable contains information about the specific rater–target combination because it depicts the deviation of the true score for this combination from what would be expected due to the trait score of the target and the true method effect of the rater. Consider, for example, that a rater (e.g., Dave, the class teacher) rates multiple targets (e.g., students) and it is found that he tends to have a somewhat positive bias, rating all of his students 2 points higher than the average teacher. This is captured as $R_{1ijk} = 2$. One of the students that Dave is rating is John, who has a trait score of 3 across all raters (i.e., all teachers), indicated by $T_{1ijk} = 3$. A purely additive approach would then lead to the conclusion that the true score of Dave rating John should be 5 ($\tau_{rtijk} = R_{1ijk} + T_{1ijk}$). Assume that Dave had more negative interactions with John than with other students. His personal experience might have resulted in Dave rating John lower than we would expect according to (1) John's trait score of 3 and (2) Dave's generally positive bias of +2. This dislike may lead to Dave rating John 2 points below John's trait score of 3. The value of the interaction variable thus shows that Dave's specific rating for John is 4 points below what would be expected according to John's trait score of 3 and Dave's generally positive rating bias of +2. In contrast to the general bias of a given rater r that applies to all targets (and is characterized by the values on the variable R_{tijk}), the values on the interaction variable Int_{rtijk} thus characterize additional, target-specific rater biases of a specific rater r.[1]

To separate measurement error influences from true interaction effects, it is necessary to assume unidimensional interaction Int_{rtijk} factors:

$$Int_{rtijk} = \lambda_{ijk}^{INT} Int_{rtjk}. \tag{6}$$

Equation (6) implies that the latent interaction variables Int_{rtijk} are positive linear functions of each other, respectively. Hence, it is assumed that the latent interaction effects are perfectly correlated across different indicators i and i' of the same trait method unit (TMU). According to this assumption (see Equation [6]), it is possible to replace the indicator-specific latent interaction variables Int_{rtijk} with a general latent interaction factor Int_{rtjk} weighted by a factor-loading λ_{ijk}^{INT} parameter. Note that the subscripts r and t were dropped for the factor-loading parameters (λ_{ijk}^{INT}), assuming that they are identical across raters and targets. In addition, it is assumed that the interaction variables (Int_{rtijk}) are independently and identically distributed (IID) random variables with zero means.

In summary, the measurement model of the observed variables pertaining to the set of nonindependent interchangeable raters is given by

$$Y_{rtijk} = T_{tijk} + R_{rijk} + \lambda_{ijk}^{INT} Int_{rtjk} + \epsilon_{rtijk}. \tag{7}$$

[1] Throughout the present work, we use the term "bias" as deviations from conditional expectations (e.g., true or trait scores). That is, the term "bias" should not be misinterpreted as a kind of "false" evaluation.

The model expressed in Equation (7) is an extension of two-way (un)balanced random effects models to multiple indicator MTMM designs with nonindependent interchangeable raters. To be fully in line with the description of two-way (un)balanced random effects models, we include the overall grand mean μ_{ijk} for each indicator i, construct j, and method k in the model:

$$Y_{rtijk} = \mu_{ijk} + T_{tijk} + R_{rijk} + \lambda_{ijk}^{INT} Int_{rtjk} + \epsilon_{rtijk}. \quad (8)$$

Note that the latent variables (T_{tijk}, R_{rijk}, Int_{rtjk}, and ϵ_{rtijk}) have an expectation (mean) of zero in Equation (8). The models described (see Equations [7] and [8]) assume indicator-specific latent target and rater variables, instead of common target and rater factors. This means that the models represent the least restrictive variant of a MTMM measurement model for nonindependent interchangeable raters.

Combination of structurally different and nonindependent interchangeable raters

In the next step, measurement models of both structurally different methods (see Equation [1]) and nonindependent interchangeable methods (see Equation [7]) are combined. To differentiate between structurally different raters (e.g., students' self-reports) and the sets of nonindependent interchangeable raters (e.g., multiple teacher reports), we use the following notation: $k = 1$: structurally different raters, e.g. self-reports; $k = 2$: nonindependent interchangeable raters, e.g. multiple teacher reports. Then,

$$Y_{tij1} = T_{tij1} + \epsilon_{tij1}, \quad (9)$$

$$Y_{rtij2} = T_{tij2} + R_{rij2} + \lambda_{ij2}^{INT} Int_{rtj2} + \epsilon_{rtij2}. \quad (10)$$

Equation (9) states that the observed variables of the structurally different raters (e.g., self-reports) measure a target-specific trait T_{tij1} and a latent error ϵ_{tij1} variable. By contrast, Equation (10) states that the observed variables of the nonindependent interchangeable raters measure a target-specific trait T_{tij2} variable, a rater-specific method R_{rij2} variable, a weighted rater-target interaction $\lambda_{ij2}^{INT} Int_{rtj2}$ variable, and a latent error ϵ_{rtij2} variable.

Following the CTC(M-1) modeling approach (cf. Eid, 2000; Eid et al., 2003, 2008), it is possible to define target-specific method variables by choosing a reference method (e.g., self-reports, $k = 1$) and by conducting linear latent regression analysis on the level of the trait variables (T_{tij1}, T_{tij2}). Generally, the most prominent method should be selected as a reference method. Guidelines on choosing an appropriate reference method can be found in the work by Geiser et al. (2008).

With respect to the example above, either the structurally different raters or the set of interchangeable raters may serve as reference method. For simplicity, we choose the first method (structurally different raters, self-reports) as reference method. The set of non-interchangeable raters serves as the non-reference method. According to the CTC(M-1) modeling approach (Eid, 2000; Eid et al., 2003, 2008), the true scores of the nonreference method (dependent variables) are predicted by the true score pertaining to the reference method (independent variables). Formally, these linear latent regression analyses can be expressed as follows:

$$E(T_{tij2}|T_{tij1}) = \mu_{ij2} + \lambda_{ij2} T_{tij1}. \quad (11)$$

The residuals of these latent regression analyses are then defined as (target-specific) common method CM_{tij2} variables:

$$\begin{aligned} CM_{tij2} &= T_{tij2} - E(T_{tij2}|T_{tij1}) \\ &= T_{tij2} - (\mu_{ij2} + \lambda_{ij2} T_{tij1}). \end{aligned} \quad (12)$$

The common method CM_{tij2} variable represents the part of the target's trait T_{tij2} measured by the nonreference method that cannot be predicted by the trait T_{tij1} variable measured by the reference method. The common method variables CM_{tij2} capture the part of true expected ratings that is shared by all interchangeable raters, but not shared with the reference method (i.e., self-report). Thus, the target effect T_{tij2} measured by the set of nonindependent interchangeable raters can be further decomposed into

$$T_{tij2} = E(T_{tij2}|T_{tij1}) + [T_{tij2} - E(T_{tij2}|T_{tij1})] \quad (13)$$

$$= \mu_{ij2} + \lambda_{ij2} T_{tij1} + CM_{tij2}. \quad (14)$$

For parsimony, we recommend specifying unidimensional common method factors: $CM_{tij2} = \lambda_{ij2}^{CM} CM_{tj2}$. In practice, the common method variables will often be strongly correlated across different indicators pertaining to the same TMU and thus should be replaced by common method CM_{tj2} factors. Note that this assumption is not necessary in order to identify and estimate the model.

In summary, the measurement model of the C4 model can be expressed as follows:

$$Y_{tij1} = T_{tij1} + \epsilon_{tij1}, \quad (15)$$

$$\begin{aligned} Y_{rtij2} = \mu_{ij2} &+ \lambda_{ij2}^{T} T_{tij1} + \lambda_{ij2}^{CM} CM_{tj2} \\ &+ R_{rij2} + \lambda_{ij2}^{INT} Int_{rtj2} + \epsilon_{rtij2}. \end{aligned} \quad (16)$$

Equation (15) states that the observed variables of the reference method (e.g., self-reports) are decomposed into a reference latent trait variable T_{tij1} and a measurement error variable ϵ_{tij1}. According to Equation (16), the observed variables of the interchangeable raters (e.g., peer reports) measure a latent indicator-specific reference trait factor (T_{tij1}), a weighted target-specific method factor ($\lambda_{ij2}^{CM} CM_{tj2}$), a latent indicator-specific rater-specific method factor (R_{rij2}), a weighted rater-target-specific

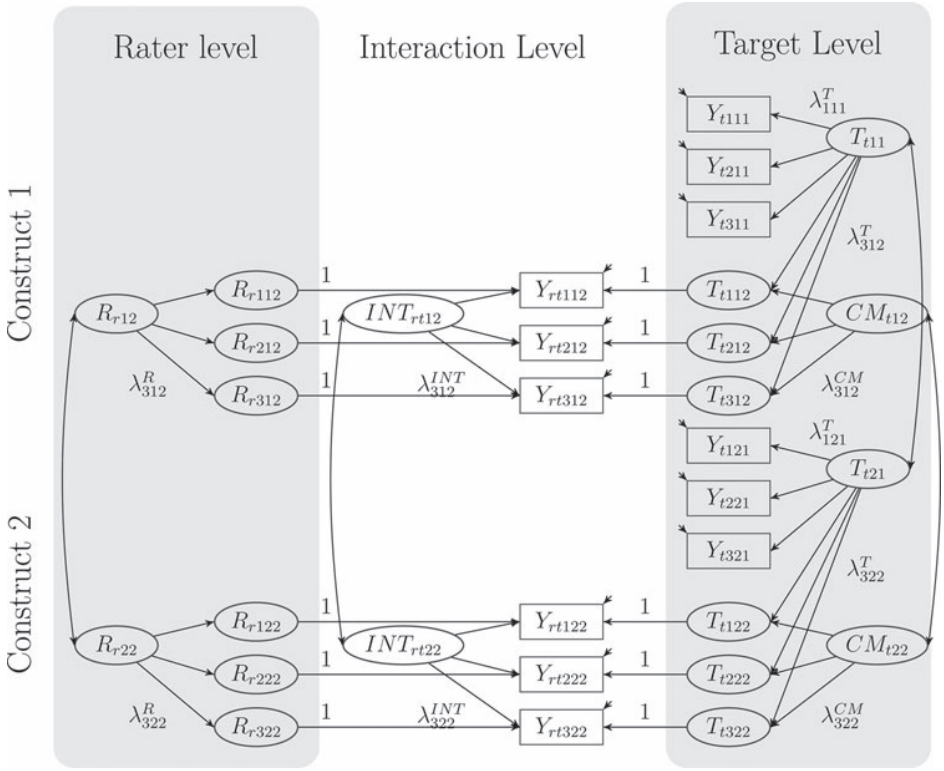

Figure 1. Path diagram of the C4 model with three indicators ($i = 3$), two constructs ($j =$ constructs), and two types of raters ($k = 1$: structurally different raters; $k = 2$: set of nonindependent interchangeable raters). The rater and target level (level 2) are represented in two gray rectangles. The rater-target level (level 1) is represented with a white background. For simplicity, the first (i.e., structurally different) method was chosen to serve as the reference method in the figure.

interaction factor ($\lambda^{INT}_{ij2} Int_{rtj2}$), and a measurement ($\epsilon_{rtij2}$) error variable.

A more restrictive variant of the C4 model with unidimensional latent reference trait (T_{tj1}) and unidimensional latent rater-specific (R_{rj2}) factors is given by

$$Y_{tij1} = \mu_{ij1} + \lambda^T_{ij1} T_{tj1} + \epsilon_{tij1}, \tag{17}$$

$$Y_{rtij2} = \mu_{ij2} + \lambda^T_{ij2} T_{tj1} + \lambda^{CM}_{ij2} CM_{tj2} + \lambda^R_{ij2} R_{rj2} + \lambda^{INT}_{ij2} Int_{rtj2} + \epsilon_{rtij2}. \tag{18}$$

This model (see Equations [17] and [18]) is illustrated as a path diagram in Figure 1. In case of heterogeneous observed variables (e.g., items assessing different aspects of a construct), the model (see Equations [17] and [18]) can be too restrictive. In such cases, we recommend using the less restrictive variant of the C4 model (see Equations [15] and [16]). Researchers may compare the fit of both models by using Bayesian model fit criteria (e.g., deviance information criteria, DIC).

The correlations between latent trait factors T_{tj1} and $T_{tj'1}$ pertaining to different constructs j and j' can be interpreted as an indicator of discriminant validity. High correlations indicate low discriminant validity. Correlations between method factors (CM_{tj2}, R_{rj2}, and Int_{rtj2}) pertaining to different constructs ($CM_{tj'2}$, $R_{rj'2}$, and $Int_{rtj'2}$, where $j \neq j'$) indicate to which degree the method effects can be generalized across different constructs. For example, high correlations between R_{rj2} and $R_{rj'2}$ suggest that rater-specific effects generalize across different constructs j and j'.

Variance coefficients

According to the independent sampling assumption, the latent rater, target, and interaction variables are mutually uncorrelated (see Gaugler & Akritas, 2011). Thus, it possible to decompose the total variance of the observed variables (Y_{tijk} and Y_{rtijk}) as follows:

$$Var(Y_{tij1}) = Var(T_{tij1}) + Var(\epsilon_{tij1}), \tag{19}$$

$$Var(Y_{rtij2}) = \left(\lambda^T_{ij2}\right)^2 Var(T_{tij1}) + \left(\lambda^{CM}_{ij2}\right)^2 Var(CM_{tj2}) + \left(\lambda^R_{ij2}\right)^2 Var(R_{rj2}) + \left(\lambda^{INT}_{ij2}\right)^2 Var(Int_{rtj2}) + Var(\epsilon_{rtij2}). \tag{20}$$

The level-1 consistency coefficient represents the proportion of true variance of a nonreference method indicator

that is determined by the reference method. It is defined as

$$L1Con(\tau_{rtij2}) = \frac{(\lambda_{ij2})^2 Var(T_{tij1})}{Var(Y_{rtij2}) - Var(\epsilon_{rtij2})}. \quad (21)$$

The level-2 consistency coefficient represents the proportion of true target-specific variance (i.e., free of rater-specific and interaction effects) of a nonreference method indicator that is determined by the reference method

$$L2Con(\tau_{rtij2}) = \frac{(\lambda_{ij2})^2 Var(T_{tij1})}{(\lambda^{T}_{ij2})^2 Var(T_{tij1}) + (\lambda^{CM}_{ij2})^2 Var(CM_{tj2})}. \quad (22)$$

In addition, three different method specificity coefficients can be defined. The common method specificity coefficient reflects the proportion of rater agreement among all interchangeable raters (common view) that is not shared with the reference method (e.g., self-report):

$$CMS(\tau_{rtij2}) = \frac{(\lambda^{CM}_{ij2})^2 Var(CM_{tj2})}{Var(Y_{rtij2}) - Var(\epsilon_{rtij2})}. \quad (23)$$

The rater-specific method specificity coefficient captures the proportion of rater-specific effects that is not shared with other interchangeable raters, but is only due to the specific view of a particular interchangeable rater. The rater-specific method specificity coefficient is defined as

$$RMS(\tau_{rtij2}) = \frac{(\lambda^{R}_{ij2})^2 Var(R_{rj2})}{Var(Y_{rtij2}) - Var(\epsilon_{rtij2})}. \quad (24)$$

The interdependency coefficient represents the proportion of true variance of that is due to interaction effects between targets and raters. The interdependency coefficient is defined as follows:

$$IMS(\tau_{rtij2}) = \frac{(\lambda^{INT}_{ij2})^2 Var(Int_{rtj2})}{Var(Y_{rtij2}) - Var(\epsilon_{rtij2})}. \quad (25)$$

Finally, the reliability coefficients of the observed variables pertaining to the reference method (here, structurally different method) or the nonreference method (e.g., set of nonindependent interchangeable raters) are defined as the proportion of the variance of the observed variables that is not due to measurement error influences:

$$Rel(Y_{tij1}) = 1 - \frac{Var(\epsilon_{tij1})}{Var(Y_{tij1})}, \quad (26)$$

$$Rel(Y_{rtij2}) = 1 - \frac{Var(\epsilon_{rtij2})}{Var(Y_{rtij2})}. \quad (27)$$

Parameter estimation and credibility intervals

The C4 model is a two-way crossed random effects model for (un)balanced data. Estimating cross-classified multilevel structural equation models is challenging (see Cho, Partchev, & De Boeck, 2012; Cho & Rabe-Hesketh, 2011; Jeon & Rabe-Hesketh, 2012; Jeon & Rijmen, 2014; Rabe-Hesketh, Skrondal, & Pickles, 2004, 2005; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). One of the difficulties of maximum-likelihood-based approaches is that the integrals in the likelihood function are crossed and cannot be simplified (Rabe-Hesketh et al., 2005). Although it may be possible to estimate such models using maximum-likelihood-based approaches (see Jeon & Rijmen, 2014, for a discussion), one alternative solution is to use Bayesian estimation techniques incorporating Markov chain Monte Carlo (MCMC) simulation. Bayesian estimation techniques do not rely on numerical integration, which can be computationally demanding (especially in case of CFA-MTMM models with many latent variables), but instead enable researchers to generate the entire posterior distribution of each model parameter by MCMC simulation.

Moreover, Bayesian estimation techniques allow researchers to estimate credibility intervals for model parameters or functions of model parameters. This aspect seems particularly convenient with regard to MTMM analysis, allowing researchers to estimate credibility intervals for different coefficients representing convergent validity, discriminant validity, method specificity, and reliability. Hence, Bayesian estimation techniques allow researchers to provide more information concerning the psychometric properties of the given measures (e.g., a plausible range of convergent and discriminant validity). In subsequent studies, this information can be used for the specification of more informative priors in future applications.

Empirical application

Data set

The C4 model with indicator-specific (see Equations [15–16]) as well as unidimensional latent trait and rater factors (see Equations [17–18] and Figure 1) was fitted to data from a German educational intervention study examining the academic interest and academic self-concept via student self-reports (Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013) as well as multiple teacher reports for each student. In total, 7,828 student reports and 389 teacher reports were collected. Teachers from different subjects (i.e., math, German, English, biology, music/arts) were asked to rate a portion of the students attending

their classes. Each student was rated by 2–14 teachers. However, due to the measurement design, some teachers rated multiple students (1–66), leading to a violation of the independence assumption of interchangeable raters (here, teachers) for each student. Moreover, students were asked to rate themselves with respect to both constructs (i.e., academic interest and academic self-concept). All observed variables were centered at the class mean (i.e., centering at the group mean). Using group-mean centering, each variable can be divided into two orthogonal components (i.e., within and between component). We aimed to model "pure" within-class variation. The empirical application is presented for illustrative purposes and to explain how the model parameters of the C4 model can be interpreted.

Research objectives

From a substantive perspective, the main research questions of the present study were as follows:

1. To what degree do teacher reports of academic interest and academic self-concept relate to students' self-reports? This research question concerns teachers' general ability to judge students' noncognitive characteristics or the convergent validity between teachers' and students' ratings, which is a key goal of MTMM analysis. In the C4 model, the level-1 and level-2 consistency coefficients can be studied for that purpose.
2. To what degree do individual or common teacher effects exist? This research question relates to the issue of whether teachers deviate in their specific evaluations of students' competencies or whether teachers generally (i.e., as an entire rater group) deviate from students' self-reports. In the C4 model, the common and rater-specific method specificity coefficients can be compared for that purpose.
3. To what degree can these method effects be generalized across different constructs? Given that both constructs (i.e., academic interest and academic self-concept) are related on a theoretical level, it can be expected that the two-method (teacher) effects can be generalized across both constructs as well. In the C4 model, the correlations between the common and rater-specific method factors can be investigated for that purpose.
4. Do students differentiate between academic interest and academic self-concept? This research question concerns the discriminant validity of students' self-reports, which is another classical goal of MTMM analysis. In the C4 model, the correlations between the latent trait factors pertaining to the reference method can be examined for that purpose.
5. Do teachers judge students' characteristics differently depending on which student they rate? What proportion of the true teacher ratings is attributable to these rater-target-interaction effects? In the C4 model, the interdependency coefficient can be examined for that purpose.
6. Finally, how reliable are students' and teachers' evaluations? In the C4 model, this can be examined by calculating the reliability coefficients for the observed variables.

Measures

Two short questionnaires were used for the assessment of academic interest and academic self-concept. Academic interest was measured by three items on a 4-point Likert scale (1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree). The items (e.g., "School is important to me personally.") were taken from two German scales also used in international, large-scale assessments such as PISA (Kunter et al., 2003; Ziegler et al., 2008). Academic self-concept was assessed by three items on a 5-point Likert scale. The items (e.g., "Learning new things in school is" … 1 = difficult; 2 = rather difficult; 3 = neither/nor; 4 = rather easy; 5 = easy) were taken from a German scale for the assessment of academic self-concept (SESSKO; Schöne, Dickhäuser, Spinath, & Stiensmeier-Pelster, 2002). For simplicity reasons, the items were treated as continuous indicators.

Estimation and model selection

The C4 model was fitted to the data set using the freely accessible software R 3.1.2 (R Core Team, 2014), JAGS 3.4.0 (Plummer, 2003), R2jags (Su & Yajima, 2015), and mcmcplots (Curtis, 2012). A detailed description of the model and prior specification can be found in Appendices A and B. A syntax for estimating the C4 model is provided in the web-based supplementary materials (S.2 and S.3). Convergence diagnostics for the MCMC chains were assessed using the Gelman-Rubin convergence criterion (Gelman et al., 2004; Gelman & Rubin, 1992) and by visual inspection of trace plots and autocorrelation plots. The Gelman-Rubin convergence criterion is based on the potential scaling reduction (PSR) factor. The PSR reflects a comparison of the within and between variation for each model parameter across multiple MCMC chains. A PSR close to 1.00 indicates that multiple chains have converged

Table 2. Model fit statistics of the C4 model with indicator-specific and unidimensional factors on the rater and target level.

C4 model	Indicator-specific factors	Unidimensional factors
Deviance	53,139.396	59,828.824
pD	1,100,285.4	907,097.0
DIC	1,153,424.8	966,925.8

Note: $N = 7,828$. pD = effective number of parameters; DIC = deviance information criterion. Lower DIC indicates better fit. The same prior specifications were used to evaluate the fit of both models.

to a stationary distribution. In the present study, the PSR ranged between 1.00 and 1.03.

According to the Gelman-Rubin convergence criterion and the visual inspection of the trace plots and the autocorrelation plots (see web-based supplementary material S.1), the C4 model converged after 10,000 MCMC iterations. These first 10,000 iterations served as burn-in period and were discarded. To extract the parameter estimates from the final posterior distributions, an additional 10,000 MCMC iterations were used in which every 10th iteration was recorded (thinning). The estimation took 86 minutes on a Macintosh system with 1.7 GHz Intel Core i7.

The deviance information criteria (DIC), the effective number of parameters (pD), and the deviance were used for model comparison. The DIC is a hierarchical generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Like the AIC and BIC, the DIC is a combined measure of the deviance and the complexity of the model. According to Spiegelhalter et al. (2002), the DIC is the sum of the expectation of posterior deviance ($\bar{D}$) and the effective number of parameters (p_D), where p_D is an approximate measure of model complexity. Models with lower DIC should be preferred.

Results

Table 2 presents information of the fit of the C4 model with indicator-specific and unidimensional latent factors. According to these results, the C4 model with unidimensional latent factors should be preferred.

The unstandardized factor-loading parameters and error variances with 95% credibility intervals from the C4 model are given in Table 3. Although the intercepts μ_{ijk} were estimated, they are not displayed in Table 3, given that they were close to zero a as result of the group-mean centering. The results of Table 3 show that the unstandardized factor loadings of the same TMU were close to 1, indicating that the observed variables loaded similarly on the corresponding latent factors. The trait factor loadings of the nonreference method (i.e., the latent regressions) were relatively low (.15 - .17 for academic interest and .21 - .22 for academic self-concept), indicating a rather low level of rater agreement (convergent validity) between student self-reports and teacher reports. The 95% credibility intervals revealed that all parameter estimates differed from zero.

Table 4 presents the variance coefficients of consistency, method specificity, and reliability of each single observed variable with a 95% Bayesian credibility interval. It is interesting that the reliabilities of the teacher report measures (*Rel* = .69–.79) were generally higher than those of the students' self-report measures (*Rel* = .42–.60).

The first research question concerned the rater agreement (convergent validity) between teacher and student reports. The (level-1 and level-2) consistency coefficients in Table 4 can be interpreted as indicators of the convergent validity. The level-1 consistency coefficients capture the amount of shared variance between true student self-reports and true single (or individual) teacher ratings. In the present study, the level-1 consistency coefficients ranged between 2% and 3%. That means that only 2%–3% of the variability in students' true self-ratings could be explained by true individual teacher ratings. Thus, the convergent validity between single student ratings and single teacher ratings was rather low. The square root of the consistency coefficients can be seen as latent correlations between the student self-reports and teacher reports. In this study the latent correlations (i.e., square root of the level-1 consistency coefficients) ranged between .14 and .17. Again, this finding suggests that teachers were not very successful in predicting students' self-reported characteristics.

On a descriptive level, the level-2 consistency coefficients were comparably higher, ranging between 6% and 10% or, in terms of latent correlations, ranging between .24 and .32. The level-2 consistency coefficients represent the rater agreement between students' self-reports and the common teacher perspective. Although the convergent validity between teacher and student reports was comparably high at level 2, the results are consistent with past findings showing that self- and other reports often share only little in common (see, e.g., Koch, 2013; Koch et al., in press; Nussbeck, Eid, & Lischetzke, 2006).

Another research goal was to examine the amount of method influences due to rater-specific, common rater, and target-rater-interaction effects. In the last three columns of Table 4, the method specificity coefficients (CMS, RMS, and IMS) are given.

The amount of true common method (CMS) influences ranged between 24% and 33% for academic interest and between 23% and 27% for academic self-concept. Given that the CMS coefficients represent the common view of the teachers that was not shared with students' self-reports, the CMS coefficients can also be interpreted

Table 3. C4 Model with common latent trait factors for structurally different and nonindependent interchangeable raters: Unstandardized factor loadings and error variances on the rater level.

		Trait	Method			Error
Rater	Indicator	λ^{T}_{ijk}	λ^{R}_{ijk}	λ^{CM}_{ijk}	λ^{INT}_{ijk}	*Var*(ϵ)
			Academic interest			
Students	Y_{t111}	1.00				0.28 (0.24; 0.32)
	Y_{t211}	0.87 (0.79; 0.98)				0.35 (0.31; 0.39)
	Y_{t311}	1.06 (0.95; 1.17)				0.41 (0.37; 0.46)
Teachers	Y_{rt112}	0.15 (0.12; 0.19)	1.00	1.00	1.00	0.12 (0.11; 0.13)
	Y_{rt212}	0.17 (0.13; 0.21)	1.00 (0.95; 1.07)	1.07 (1.02; 1.11)	0.88 (0.84; 0.92)	0.15 (0.14; 0.15)
	Y_{rt312}	0.15 (0.12; 0.18)	1.35 (1.29; 1.40)	0.95 (0.92; 1.00)	0.91(0.87; 0.95)	0.12 (0.12; 0.13)
			Academic self-concept			
Students	Y_{t121}	1.00				0.36 (0.27; 0.29)
	Y_{t221}	1.11 (1.00; 1.22)				0.48 (0.41; 0.51)
	Y_{t321}	1.08 (1.00; 1.19)				0.25 (0.22; 0.29)
Teachers	Y_{rt122}	0.21 (0.17; 0.26)	1.00	1.00	1.00	0.17 (0.16; 0.18)
	Y_{rt222}	0.22 (0.17; 0.27)	1.16 (1.08; 1.24)	1.08 (1.03; 1.12)	1.13 (1.10; 1.16)	0.18 (0.17; 0.19)
	Y_{rt322}	0.22 (0.18; 0.27)	1.05 (0.98; 1.13)	1.10 (1.05; 1.14)	1.04 (1.01; 1.07)	0.16 (0.15; 0.17)

Note. Y_{tijk} = observed variables (students' self-reports); Y_{rtijk} = observed variables (teacher reports); i = indicator; j = construct (1 = academic self-interest; 2 = academic self-concept); k = raters. λ^{T}_{ijk} = trait factor-loading parameter on the target level; λ^{R}_{ijk} = unique method factor-loading parameter on the rater level; λ^{CM}_{ijk} = common method factor-loading parameter on the target level; λ^{INT}_{ijk} = factor-loading parameter of the latent interaction factor on level 1; *Var*(ϵ) = variance of the latent error variables. Values of the 95% highest posterior density interval (i.e., credibility interval) are given in parentheses.

as an indicator of rater consensus among multiple interchangeable raters (here, teachers) that is not shared with the reference method (here, the students' perspective). The CMS coefficients correspond to latent correlations ranging from .48 to .57, indicating a medium level of rater agreement among teachers.

By contrast, the amount of true teacher-specific method (RMS) influences ranged between 23% and 28% for academic interest and between 13% and 14% for academic self-concept. While the CMS coefficients can be interpreted as a kind of rater consensus, the RMS coefficients reflect unique (rater-specific) method influences. With regard to Hoyt (2000), the RMS coefficients can also be interpreted as magnitude of the rater effects. Rater effects reflect "how *r* [a rater] generally perceives people on the attribute of interest" (Hoyt, 2000, p. 67). An interesting finding is that the rater-specific influences were lower for students' self-concept than for students' academic interest. This means that individual teacher effects were less present with regard to the assessment of students' self-concept. With regard to research question 2, these findings suggest that the main source of disagreement between teacher ratings and students' self-reports were attributable to common, instead of individual, teacher effects.

Nevertheless, the major part of method-specific influences was due to interaction effects between targets and raters. The interdependency coefficients ranged between

Table 4. C4 Model with common latent trait factors for structurally different and nonindependent interchangeable raters: Reliability, consistency, and method specificity.

			Consistency		Method Specificity		
Rater	Indicator	Reliability	L1Con	L2Con	CMS	RMS	IMS
			Academic interest				
Students	Y_{t111}	.54 (.48; .60)					
	Y_{t211}	.42 (.37; .48)					
	Y_{t311}	.47 (.41; .53)					
Teachers	Y_{rt112}	.75 (.73; .77)	.02 (.01; .03)	.06 (.04; .10)	.27 (.24; .30)	.19 (.17; .25)	.52 (.49; .55)
	Y_{rt212}	.69 (.67; .70)	.02 (.01; .04)	.07 (.04; .10)	.33 (.30; .36)	.21 (.17; .25)	.43 (.40; .47)
	Y_{rt312}	.75 (.74; .77)	.02 (.01; .03)	.07 (.04; .10)	.24 (.21; .27)	.33 (.30; .37)	.41 (.38; .44)
			Academic self-concept				
Students	Y_{t121}	.55 (.50; .61)					
	Y_{t221}	.46 (.41; .51)					
	Y_{t321}	.60 (.54; .65)					
Teachers	Y_{rt122}	.76 (.74; .77)	.03 (.02; .04)	.10 (.06; .14)	.25 (.22; .28)	.13 (.11; .16)	.59 (.56; .62)
	Y_{rt222}	.79 (.77; .80)	.02 (.01; .03)	.09 (.05; .13)	.23 (.21; .26)	.14 (.12; .17)	.60 (.57; .63)
	Y_{rt322}	.79 (.78; .80)	.03 (.02; .04)	.09 (.06; .13)	.27 (.24; .30)	.13 (.11; .16)	.57 (.54; .60)

Note: Y_{tijk} = observed variables (students' self-reports); Y_{rtijk} = observed variables (teacher reports); i = indicator; j = construct (1 = academic self-interest; 2 = academic self-concept); k = rater. Rel = Reliability; L1Con = level-1 consistency; L2Con = level-2 consistency; CMS = common method (rater) specificity; RMS = unique method (rater) specificity; IMS = interdependency (rater-target specificity). The coefficients of consistency and method specificity were standardized on the true variance of an indicator. Values of the 95% highest posterior density interval (i.e., credibility interval) are given in parentheses.

Table 5. Latent variances (diagonal) and correlations of the trait and method factors in the C4 model.

	1	2	3	4	5	6	7	8
T_{t11}	.33 (.28; .38)							
T_{t21}	.47 (.41; .53)	.32 (.27; .37)						
CM_{t12}	0	0	.11 (.10; .13)					
CM_{t22}	0	0	.84 (.81; .85)	.13 (.12; .15)				
R_{r12}	0	0	0	0	.08 (.07; .09)			
R_{r22}	0	0	0	0	.41 (.31; .51)	.07 (.05; .09)		
INT_{rt12}	0	0	0	0	0	0	.21 (.20; .23)	
INT_{rt22}	0	0	0	0	0	0	.60 (.58; .62)	.32 (.30; .34)

Note: T_{tjk} = latent trait factor; CM_{tjk} = common rater factor; R_{rjk} = unique rater factor; INT_{rtjk} = rater-target-interaction factor; j = construct (1 = academic self-interest; 2 = academic self-concept); k = rater. Values of the 95% highest posterior density interval (i.e., credibility interval) are given in parentheses.

41% and 52% for academic interest and between 57% and 60% for academic self-concept. These findings reveal that teachers judge students' characteristics differently than expected from only the general teacher (rater) effect and the general student (target) effect. "A dyadic [interaction] effect is present when observed r rates target t either higher or lower than one would predict given r's rater effect and t's target effect" (Hoyt, 2000, p. 67). Hence, these findings suggest that teacher evaluations of students' noncognitive characteristics depended on which student they rated (see research question 5).

This information on the dyadic or interaction effects could not be obtained if the original ML-CTC(M-1) model had been applied. Although the reliability, consistency, and common method specificity coefficients in the original CTC(M-1) model will always match those of the C4 model, the unique method specificity coefficients represent a compound of the rater-specific and the interdependency effects (i.e., UMS = RMS + IMS). Hence, the UMS coefficients can always be derived by adding the RMS and the IMS coefficients of the C4 model.

In Table 5, the variances of the latent factors as well as their correlations are given. The correlation between the latent trait factors can be interpreted as an indicator of discriminant validity. In this study, the correlation of the latent trait factors was .47 [CI: .41–.53], indicating that self-reported academic interest was positively associated with self-reported academic self-concept. This result means that students who reported higher scores in academic interest also tended to report higher scores in academic self-concept. With regard to research question 4, these findings provide some evidence of discriminant validity.

The correlations between latent method factors pertaining to different constructs reflect to which degree method-specific effects generalize across different constructs. In this study, all method factors correlated positively with each other, displaying that individual and common teacher effects as well as teacher–student interaction effects generalized across both constructs (see research question 3). The highest correlation was found among the common method factors ($r_{(CM_{t12},CM_{t22})} = .84$), revealing that teachers who tended to over- or underestimate students' academic interest also tended to over- or underestimate students' academic self-concept in a similar way.

Discussion

The main goal of this study was to extend the ML-CTC(M-1) model by Eid et al. (2008) to cases of nonindependent sets of interchangeable raters (i.e., fully or partially cross-classified multirater data). Cross-classified multirater data originate whenever some or all interchangeable raters (e.g., peers, students, colleagues) are allowed to rate multiple targets. This is a common issue in practice.

In the present study, we proposed a new multiple indicator multilevel CFA-MTMM model (C4 model) for the analysis of cross-classified multirater designs combining structurally different and multiple nonindependent interchangeable raters. The C4 model bears several advantages and overcomes many limitations of current modeling approaches of cross-classified data.

First, the C4 model accounts for the dependency among different sets of interchangeable raters that are fully or partly overlapping. If the additional clustering is not modeled, the model is essentially underspecified, and the standard errors of the parameter estimates are likely to be biased (see Luo & Kwok, 2009; Schultze et al., 2015). Second, the C4 model enables researchers to explicitly model rater-target-interaction effects as a latent variable (i.e., free of measurement error influences). With regard to the classical cross-classified multilevel models or the original ML-CTC(M-1) model, this is not possible. By relating explanatory variables (e.g., rater's age, target's sex, quality of the rater-target relationship) to the latent factors in the C4 model, researchers can investigate potential predictors of rater effects, target effects, and/or rater-target-interaction effects. Third, the C4 model allows researchers to analyze complex multirater designs combining structurally different and interchangeable methods. In this

respect, the C4 model represents an extension of current modeling approaches such as Eid et al. (2008), Gaugler and Akritas (2011), and Hoyt (2000) to complex cross-classified MTMM designs combining different types of methods.

To illustrate the new model and facilitate the interpretation of the model parameters, we applied the model to real data from an educational intervention study. In our application, the rater agreement (convergent validity) between student and teacher reports was rather low. A great amount of true variance was due to teacher as well as student-teacher-interaction effects. These results support previous findings of multirater studies, in which the rater–target agreement was relatively low, whereas the amount of method-specific influences was high (Koch, 2013; Koch et al., 2014). Our study also suggested that students' self-rated academic interest and academic self-concept were positively associated, but still showed some evidence of discriminant validity. Furthermore, the method factors correlated positively with each other. The strongest association was found among the common (teacher) method factors. This means that teachers who tended to overestimate students' self-reported academic interest also tended to overestimate students' self-reported education self-concept and vice versa. Using Bayesian estimation techniques, it was possible to obtain 95% credibility intervals for all model coefficients. Finally, we provide some practical guidelines for applied researchers who are interested in using the C4 model.

Single versus multiple indicators

We generally recommend using multiple instead of single indicators for each trait method unit (TMU). Thereby, researchers are able to specify trait-specific method effects. In addition, we recommend that researchers start with the most general C4 model (i.e., assuming indicator-specific latent factors at the rater and the target level). In a second step, researchers may then evaluate whether the latent factors are unidimensional as shown in Figure 1.

Informative versus noninformative priors

Researchers should be careful when selecting prior distributions for specific model parameters. Following the general advice given in the Bayesian literature, we recommend choosing diffuse (noninformative) priors in cases of vague or uncertain prior information. In the present study, we used a mix of informative and noninformative priors (see online Appendix C for a detailed description). We used the normal distribution $N(\mu, 1/\theta)$ for selecting priors for the intercept and factor-loading parameters in the C4 model, where μ is the mean and $1/\theta$ is the inverse of the variance (so-called precision) of the parameters. Given that we centered all observed variables at the group mean, it could be expected that the intercept parameters will be close to zero. Thus, we chose $\mu_{ijk} \sim N(0, .1)$, which means that 95% of the intercept parameters fall in the range of −.63 to .63. According to previous applications of the scales (items), the factor-loading parameters belonging to the same TMU should be very similar (i.e., close to 1). Therefore, we chose $\lambda_{ij1}^{T} \sim N(1, .1)$, $\lambda_{ij2}^{CM} \sim N(1, .1)$, $\lambda_{ij2}^{R} \sim N(1, .1)$, and $\lambda_{ij2}^{INT} \sim N(1, .1)$. This prior specification corresponds to a 95% highest posterior density interval of 0.37 and 1.63. With regard to the trait factor loadings belonging to the nonreference method (here, set of nonindependent interchangeable raters), we chose $\lambda_{ij2}^{T} \sim N(2, .2)$, which corresponds to a 95% interval of -0.69 and 1.09. Despite the fact that these prior specifications are quite informative, they seem realistic with regard to previous applications of these scales as well as with respect to previous application of similar CFA-MTMM models. For example, it would be very surprising if the factor loading parameters of the same TMU were negative. However, to ensure that these prior specifications do not substantially change the model results, we recommend that researchers cross check their results by using different (noninformative) prior settings. In the present study, we scrutinized the robustness of the parameter estimates using noninformative priors (e.g., $\mu_{ijk} \sim N[0, .0001]$ or $\lambda_{ijk}^{T} \sim N[0, .0001]$; recall that the second term is the precision). For more complex parameter distributions (i.e., the residual variances and the variance-covariance structure of the latent variables), we used noninformative priors, commonly reported in the Bayesian literature (see, e.g., Muthén & Asparouhov, 2012). To be more specific, we used the inverse gamma distribution $IG(.001, .001)$ for the residual variances and the inverse Wishart distribution $IW(I, 2)$ for the variance-covariance structure of the latent variables. Due to the properties of the inverse Wishart distribution, it is often difficult to choose specific values for its parameters that yield noninformative priors for the specific covariance elements. One remedy that has been suggested by Gelman and Hill (2007) is the scaled inverse Wishart distribution.

Model assumptions and measurement design

Despite its numerous advantages, the C4 model rests on assumptions as does any other statistical model. To be more precise, the C4 model implies that raters and targets are sampled independently from each other. This assumption might be too restrictive in practice. For example, in the present study it can be questioned whether students could be selected independently from teachers. Therefore, the results of the present study should be interpreted with some caution. However, this independent sampling

assumption is implicitly made in all cross-classified multilevel regression models, which are typically applied in educational studies. In fact, one advantage of the C4 model is that this critical assumption is explicitly stated so that researchers are able to design their study accordingly. Thus, researchers may design their studies in such a way so that this assumption will likely be met, for example, by randomly assigning targets to raters.

Required sample size

To provide detailed information concerning the required sample size for empirical applications of the C4 model, extensive simulation studies are needed, which we leave to future studies. Currently, we refer to rules of thumb that have been reported based on previous simulation studies investigating the statistical performance of multilevel structural equation models (ML-SEM). According to past simulation studies, the number of level-2 observations is of particular importance for the accuracy of the parameter estimates in ML-SEM. Thus, it is often suggested to sample at least 100 level-2 units (Julian, 2001; Koch et al., 2014). In the case of the C4 model, this would mean to sample at least 100 targets and 100 raters. In addition, past simulation studies have indicated that the number of level-1 units is important and recommended to sample at least five level-1 units per cluster (Koch et al., 2014). This would mean that each interchangeable rater rates on average five targets and that each target is rated by five interchangeable raters.

Limitations and future research

Future research should be devoted to aspects that examine the robustness of the C4 model and broaden its applicability. We strongly encourage researchers to conduct extensive simulation studies that provide information on the statistical performance of the C4 model under a wide array of conditions. Future studies may also be devoted to extending the C4 model to a greater number of measurement levels (student and teacher ratings nested in classes, schools, or districts) or to longitudinal measurement designs.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgements: The authors thank the two anonymous reviewers for their valuable and constructive comments on earlier drafts of the manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

Asparouhov, T., & Muthén, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. Retrieved from http://www.statmodel.com/download/NCMErevision2.pdf

Bauer, D., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475–493. doi: 10.1037/a0032475

Beretvas, S. N. (2011). Cross-classified and multiple membership models. In J. Hox & R. J. Kyle (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). New York, NY: Routledge Taylor & Francis Group.

Bull, H. D., Schultze, M., Geiser, C., & Scheithauer, H. (2013). *The role of cognitive empathy in the association between social intelligence and relational aggression: A MTMM analysis of informant agreement in the assessment of empathy deficits in relationally aggressive adolescents*. (Manuscript submitted for publication.)

Bull, H. D., Schultze, M., & Scheithauer, H. (2009). School-based prevention of bullying and relational aggression: The fairplayer.manual. *European Journal of Developmental Science*, *3*(3), 312–317. doi: 10.1007/s11121-009-0128-y

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi: 10.1037/h0046016

Carretero-Dios, H., Eid, M., & Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: An application to the validation of the state-trait cheerfulness inventory. *Journal of Research in Personality*, *45*, 153–164. doi:10.1016/j.jrp.2010.12.007

Castro-Schilo, L., Widaman, K. F., & Grimm, K. J. (2013). Neglect the structure of multitrait-multimethod data at your peril: Implications for associations with external variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 181–207. doi: 10.1080/10705511.2013.769385

Cho, S.-J., Partchev, I., & De Boeck, P. (2012). Parameter estimation of multiple item response profile model. *British Journal of Mathematical and Statistical Psychology*, *65*, 438–466. doi:10.1111/j.2044-8317.2011.02036.x

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25. doi:10.1016/j.csda.2010.04.015

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*, 331–360. doi:10.1207/s15327043hup1004_2

Curtis, S. M. (2012). mcmcplots: Create plots from MCMC output [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=mcmcplots

Danay, E., & Ziegler, M. (2011). Is there really a single factor of personality? A multirater approach to the apex of personality. *Journal of Research in Personality*, *45*, 560–567. doi:10.1016/j.jrp.2011.07.003

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

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

Eid, M., & Diener, E. (2006). Introduction: The need for multimethod measurement in psychology. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 3–8). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., & Nussbeck, F. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 289–299). Washington, DC: American Psychological Association.

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

Eid, M., Nussbeck, F. W., 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. doi:10.1037/a0013219

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Tech. Rep.). Research Report RR791. London, UK: Department for Education and Skills.

Gaugler, T., & Akritas, M. G. (2011). Testing for interaction in two-way random and mixed effects models: The fully nonparametric approach. *Biometrics*, *67*, 1314–1320. doi:10.1111/j.1541-0420.2011.01579.x

Geiser, C., Eid, M., West, S. G., Lischetzke, T., & Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling*, *19*(3), 409–436. doi:10.1080/10705511.2012.687658

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods*, *13*, 49–57. doi:10.1037/1082-989X.13.1.49

Geiser, C., Koch, T., & Eid, M. (2014). Data-generating mechanisms versus constructively defined latent variables in multitrait-multimethod analysis: A comment on Castro-Schilo, Widaman, and Grimm (2013). *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 509–523. doi:10.1080/10705511.2014.919816

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. doi:10.1214/ss/1177011136

Ghorpade, J. (2000). Managing five paradoxes of 360-degree feedback. *The Academy of Management Executive*, *14*, 140–150. doi:10.5465/AME.2000.2909846

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*, 364–375. doi:10.1177/0049124194022003005

Goldstein, H. (2011). Cross-classified data structures. In W. A. Shewhart & S. S. Wilks (Eds.), *Multilevel statistical models* (pp. 243–254). Chichester, UK: John Wiley & Sons. doi:10.1002/9780470973394.ch12

Hox, J., & Maas, C. (2006). Multilevel models for multimethod measures. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 269–282). Washington, DC: American Psychological Association.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86. doi:10.1037/1082-989X.5.1.64

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics*, *37*, 518–542. doi:10.3102/1076998611417628

Jeon, M., & Rijmen, F. (2014). Recent developments in maximum likelihood estimation of MTMM models for categorical data. *Frontiers in Psychology*, *5*(269), 1–7. doi:10.3389/fpsyg.2014.00269

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*(3), 325–352. doi:10.1207/S15328007SEM0803_1

Kelava, A. (2009). *Multikollinearität in nicht-linearen latenten Strukturgleichungsmodellen [Multicollinearity in non-linear structural equation models]*. (Doctoral dissertation). Goethe University, Frankfurt, Germany. Retrieved from http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/6267

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, *12*(3), 247–252. doi:10.1016/0022-1031(76)90055-X

Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111–124). Hillsdale, NJ: Lawrence Erlbaum.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin 112*(1), 165–172. doi: 10.1037/0033-2909.112.1.165

Koch, T. (2013). *Multilevel structural equation modeling of multitrait-multimethod-multioccasion data*. (Doctoral dissertation). Freie Universität Berlin, Germany. Retrieved from http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000013668/Thesis_Koch_gedreht.pdf

Koch, T., Eid, M., & Lochner, K. (in press). Multitrait-multimethod-analysis: The psychometric foundation of CFA-MTMM models. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing*. London, UK: John Wiley & Sons.

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multileve CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology*, *5*(311), 1–9. doi:10.3389/fpsyg.2014.00311

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., ... Weiß, M. (2003). *Pisa 2000 - Dokumentation der Erhebungsinstrumente* [PISA 2000: Documentation of the measures].

Luo, W., & Kwok, O.-M. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research, 44*, 182–212. doi: 10.1080/00273170902794214

Maas, C. J., Lensvelt-Mulders, G. J., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*(3), 72–77. doi: http://dx.doi.org/10.1027/1614-2241.5.3.72

Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R., & Brodbeck, F. (2015). A multilevel CFA–MTMM approach for multisource feedback instruments: Presentation and application of a new statistical model. *Structural Equation Modeling: A Multidisciplinary Journal*. doi:10.1080/10705511.2014.990153

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

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187–1197. doi:10.1037/0003-066X.52.11.1187

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*, 275–300. doi:10.1037/1082-989X.9.3.275

Meiser, T., & Steinwascher, M. A. (2014). Different kinds of interchangeable methods in multitrait-multimethod analysis: A note on the multilevel CFA-MTMM model by Koch et al. (2014). *Frontiers in Psychology*, *5*(615), 1–3. doi:10.3389/fpsyg.2014.00615

Moosbrugger, H., Schermelleh-Engel, K., & Klein, A. (1997). Methodological problems of estimating latent interaction effects. *Methods of psychological research online*, *2*, 95–111.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. doi: http://dx.doi.org/10.1037/a0026802

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analysing multitrait-multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results? *British Journal of Mathematical and Statistical Psychology*, *59*, 195–213. doi:10.1348/000711005X67490

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

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using gibbs sampling [Computer software manual]. Retrieved from http://citeseer.ist.psu.edu/plummer03jags.html

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research 45*(1), 45–72. doi: 10.1080/00273170903504729

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 41–63. doi:10.1111/j.1467-985X.2007.00517.x

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research*, *106*, 64–76. doi:10.1080/00220671.2012.667010

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004, October). GLLAMM manual [Computer software manual]. Working Paper 160. Berkeley, CA: U.C. Berkeley Division of Biostatistics Working Paper Series. Retrieved from http://biostats.bepress.com/ucbbiostat/paper160

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323. doi:10.1016/j.jeconom.2004.08.017

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, *18*, 321–349. doi:10.3102/10769986018004321

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd Ed.). Thousand Oaks, CA: Sage.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205. doi:10.1037/1082-989X.8.2.185

Sahai, H., & Ojeda, M. M. (2005). *Two-way crossed classification with interaction. Vol. II: Unbalanced data theory, methods, applications, and data analysis*. New York, NY: Springer. doi:10.1007/978-0-8176-8168-5_4

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). *Skalen zur Erfassung des schulischen Selbstkonzepts (SESSKO)* [Scales for the measurement of academic self-concept.]. Göttingen, Germany: Hofgrefe.

Schultze, M. (2012). *Evaluating what the crowd says. A longitudinal structural equation model for exchangeable and structurally different methods for evaluating interventions*. Diploma thesis, Freie Universität, Berlin, Germany.

Schultze, M., Koch, T., & Eid, M. (2015). The effects of nonindependent rater sets in multilevel–multitrait–multimethod models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 439–448. doi:10.1080/10705511.2014.937675

Schumacker, R. E., & Marcoulides, G. A. (1998). *Interaction and nonlinear effects in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Searle, S. R. (2006). *Linear models for unbalanced data (Wiley series in probability and statistics)*. Hoboken, NJ: John Wiley & Sons.

Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components (Wiley series in probability and statistics)*. (Vol. 391). Hoboken, NJ: John Wiley & Sons.

Shrout, P. E. (1995). Measuring the degree of consensus in personality judgments. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods and theory: A festschrift honoring Donald W. Fiske* (pp. 79–92). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639. doi:10.1111/1467-9868.00353

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

Steyer, R., & Eid, M. (2001). *Messen und Testen* [Measurement and testing.] (2nd Ed.). Heidelberg, Germany: Springer.

Su, Y.-S., & Yajima, M. (2015). R2jags: A package for running jags from r [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=R2jags (R package version 0.05-01)

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

Yammarino, F. J. (2003). Modern data analytic techniques for multisource feedback. *Organizational Research Methods*, *6*, 6–14. doi:10.1177/1094428102239423

Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics*, *25*, 35–44. doi:10.1016/S0090-2616(97)90035-8

Ziegler, A., Dresel, M., Schober, B., & Stöger, H. (2008). *Motivationstestbatterie für Schülerinnen und Schüler der Jahrgangsstufen 5-10 (MTB 5-10)* [Motivational test battery for students pertaining to grade 5-10.]. Unpublished instrument. Ulm, Germany: Universität.

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

Appendix A: Formal definition of the C4 model

The formal definition of the C4 model is based on the following random experiment and the following assumptions.

Step 1: The random experiment

The random experiment of MTMM measurement designs with structurally different and nonindependent interchangeable methods (raters) is given by the Cartesian crossproduct $\times$ of the following sets $\Omega(\cdot)$:

$$\Omega = \Omega_T \times \Omega_R \times \Omega_{ijk}. \tag{A1}$$

Equation (A1) states that first a target t is randomly chosen out of a set of possible targets Ω_T. Then, a rater r is randomly selected out of a common set of possible raters Ω_R and the rating ω_{ijk} on item $i = \{1, \ldots, i, \ldots, I\}$, construct $j = \{1, \ldots, j, \ldots, J\}$, and method $k = \{1, \ldots, k, \ldots, K\}$ is observed. One crucial assumption regarding the random experiment is that the selection of targets is independent from the selection of raters and vice versa.

The possible outcomes are mapped into the set of possible targets p_T: $\Omega \rightarrow \Omega_T$, into the set of possible raters p_R: $\Omega \rightarrow \Omega_R$, and into the set of real numbers $(Y_{rtijk}, Y_{tijk}) : \Omega \rightarrow \mathbb{R}$. The variables ($p_T$, p_R, Y_{rtijk}, and Y_{tijk}) are random variables on the probability space; p_T is called target variable and p_R is called rater variable.

Step 2: Definition of true scores and error variables

According to this sampling experiment, it is possible to define the latent variables in the C4 model as random variables. In particular, the true scores pertaining to the structurally different methods are defined as conditional expectations of the observed variables given the target variable:

$$\tau_{tij1} := E(Y_{tij1}|p_T), \quad \text{(structurally different method)}. \tag{A2}$$

For reasons of simplicity, we choose the first method ($k = 1$, the structurally different method) as reference method. In a similar logic, the true scores pertaining to the set of nonindependent interchangeable methods can be defined in terms of conditional expectations:

$$\tau_{rtij2} := E(Y_{rtij2}|p_T, p_R), \text{ (interchangeable methods)}. \tag{A3}$$

According to Definition (A3), the true scores of the set of nonindependent interchangeable methods ($k = 2$, nonreference method) are defined as conditional expectations of the observed variables Y_{rtij2} given the target variable p_T and the rater variable p_R.

The measurement error variables are then defined as residuals with respect to their corresponding true scores:

$$\epsilon_{tij1} := Y_{tij1} - \tau_{tij1} = Y_{tij1} - E(Y_{tij1}|p_T), \tag{A4}$$

$$\epsilon_{rtij2} := Y_{rtij2} - \tau_{rtij2} = Y_{rtij2} - E(Y_{rtij2}|p_T, p_R). \tag{A5}$$

Due to this definition, the measurement error variables are necessarily uncorrelated with the true scores and have an expectation of zero.

Step 3: Decomposition of the true score of the nonindependent interchangeable methods

Next, the true scores of the interchangeable methods τ_{rtij2} are further decomposed as follows:

$$\begin{aligned} E(Y_{rtij2}|p_T, p_R) &= E(Y_{rtij2}|p_T) + E(Y_{rtij2}|p_R) \\ &\quad + \left[E(Y_{rtij2}|p_T, p_R)\right. \\ &\quad \left. - E(Y_{rtij2}|p_T) - E(Y_{rtij2}|p_R)\right], \end{aligned} \tag{A6}$$

$$\tau_{rtij2} = T_{tij2} + R_{rij2} + Int_{rtij2}. \tag{A7}$$

Equations (A6) and (A7) state that the true scores pertaining to the set of nonindependent interchangeable methods can be decomposed into a true target $T_{tij2} := E(Y_{rtij2}|p_T)$ effect, a true rater $R_{rij2} := E(Y_{rtij2}|p_R)$ effect, and a true target-rater-interaction $Int_{rtij2} := E(Y_{rtij2}|p_T, p_R) - E(Y_{rtij2}|p_T) - E(Y_{rtij2}|p_R)$ variable. The true target effect is defined as the conditional expectation of the observed variables given the target variable [i.e., $E(Y_{rtij2}|p_T)$]. The true rater effect is defined as the conditional expectation of the observed variables given the rater variable [i.e., $E(Y_{rtij2}|p_R)$], and the true target-rater-interaction effect is defined as part of the true scores τ_{rtij2} that is not due to the true target T_{tij2} and true rater R_{rij2} effect.

Step 4: Definition of common method effects

According to the previous steps, the measurement equation of the observed variables can be represented as follows:

$$Y_{tij1} = T_{tij1} + \epsilon_{tij1}, \tag{A8}$$

$$Y_{rtij2} = T_{tij2} + R_{rij2} + Int_{rtij2} + \epsilon_{rtij2}. \tag{A9}$$

In Equation (A8), the true score of the reference method τ_{tij1} has been replaced by an indicator-specific latent trait T_{tij1} variable. Common method effects can be defined on the target level by predicting the true score variables of the nonreference method (dependent variable, T_{tij2}) by the true score variables measured by the reference method (independent variable, T_{tij1}). The latent linear regression can be expressed as follows:

$$E(T_{tij2}|T_{tij1}) = \mu_{ij2} + \lambda_{ij2}T_{tij1}. \tag{A10}$$

The common method effect is then defined as a latent residual variable with regard to the true target variable as measured by the reference method (here, T_{tij1}):

$$\begin{aligned} CM_{tij2} &:= T_{tij2} - E(T_{tij2}|T_{tij1}) = T_{tij2} \\ &\quad -(\mu_{ij2} + \lambda_{ij2}T_{tij1}). \end{aligned} \tag{A11}$$

The common method variables capture the part of the true target variable as measured by the nonreference method that cannot be explained by the true target variable as measured by the reference method. Due to the definition of the common method variables as latent residuals, they are necessarily uncorrelated with the latent trait variables (T_{tij1}) as well as functions of T_{tij1} and have an expectation (mean) of zero.

Resubstituting Equations (A10) and (A11) into Equation (A9), yields

$$Y_{tij1} = T_{tij1} + \epsilon_{tij1}, \tag{A12}$$

$$\begin{aligned} Y_{rtij2} &= \mu_{ij2} + \lambda_{ij2}T_{tij1} + CM_{tij2} + R_{rij2} \\ &\quad + Int_{rtij2} + \epsilon_{rtij2}. \end{aligned} \tag{A13}$$

Step 5: Additional assumptions

In the last step, additional assumptions are imposed in order to identify and estimate the model parameters. The following assumptions must be imposed in order to identify and estimate the C4 model:

$$Int_{rtij2} = \lambda_{ij2}^{Int} Int_{rtj2}, \tag{A14}$$

$$Cov(\epsilon_{tijk}, \epsilon_{t(ijk)'}) = 0, \text{ with } (ijk) \neq (ijk)', \tag{A15}$$

$$Cov(\epsilon_{rtijk}, \epsilon_{rt(ijk)'}) = 0, \text{ with } (ijk) \neq (ijk)', \tag{A16}$$

$$Cov(\epsilon_{tijk}, \epsilon_{rt(ijk)'}) = 0, \text{ with } (ijk) \neq (ijk)'. \tag{A17}$$

Equation (A14) implies that the interaction effects (Int_{rtij2}) are homogeneous across different items of the same trait method unit (TMU). As a consequence of this assumption, latent interaction factors (Int_{rtj2}) can be defined. This Assumption (A14) is necessary for separating measurement error influences from true interaction effects. Note that the factor loading λ_{ij2}^{Int} does not vary across clusters (i.e., no random factor loadings). The Assumptions (A15) to (A17) imply that the latent measurement error variables are mutually uncorrelated with each other, which is a common and widely accepted assumption in cross-sectional latent variable models.

Additional assumptions that are useful and recommended for parsimony reasons, but which are not necessary for model identification purposes, are

$$CM_{tij2} = \lambda_{ij2}^{CM} CM_{tj2}, \tag{A18}$$

$$R_{rij2} = \lambda_{ij2}^{R} R_{rj2}. \tag{A19}$$

The Assumptions (A18) to (A19) allow specifying unidimensional latent method factors (CM_{tj2} and R_{rj2}) instead

of indicator-specific latent method factors. In most empirical applications, method effects will be highly correlated across different indicators pertaining to the same TMU.

Finally, the measurement equation of the least restrictive variant of the C4 model with two methods (one structurally different method and one set of nonindependent interchangeable methods) is given by

$$Y_{tij1} = T_{tij1} + \epsilon_{tij1}, \tag{A20}$$

$$Y_{rtij2} = \mu_{ij2} + \lambda_{ij2} T_{tij1} + CM_{tij2} + R_{rij2} + \lambda_{ij2}^{INT} INT_{rtj2} + \epsilon_{rtij2}. \tag{A21}$$

Appendix B: Prior settings

We used informative priors with regard to the intercepts and factor loadings for multiple reasons. First, due to the centering it could be assumed that the intercepts will be estimated close to zero. Second, the factor loadings pertaining to the same TMU should be close to 1 because previous studies indicated that these items are rather homogeneous. Third, the trait factor loadings pertaining to the nonreference method should be positive but smaller than the loading parameters of the same TMU. This can be derived from previous applications of the CTC(M-1) model (Koch, 2013; Koch et al., 2014; Nussbeck et al., 2006), revealing low convergence between self-reports and other reports. According to our prior specification $\lambda_{ij2}^{T} \sim N(.2, .2)$, the 95% of the values would be within the interval of -0.69 and 1.09.

$$\begin{aligned}
\mu_{ijk} &\sim N(0, .1) \\
\lambda_{ij1}^{T} &\sim N(1, .1) \\
\lambda_{ij2}^{T} &\sim N(.2, .2) \\
\lambda_{ij2}^{CM} &\sim N(1, .1) \\
\lambda_{ij2}^{R} &\sim N(1, .1) \\
\lambda_{ij2}^{INT} &\sim N(1, .1)
\end{aligned}$$

We used noninformative priors for the variances and covariances of the latent variables as commonly recommended in the literature.

$$\begin{aligned}
\epsilon_{rtij2} &\sim Gamma^{-1}(.001, .001) \\
\epsilon_{tij1} &\sim Gamma^{-1}(.001, .001)
\end{aligned}$$

$$\begin{aligned}
\mathbf{\Phi}_{T} &\sim Wishart^{-1}(I, 2) \\
\mathbf{\Phi}_{CM} &\sim Wishart^{-1}(I, 2) \\
\mathbf{\Phi}_{R} &\sim Wishart^{-1}(I, 2) \\
\mathbf{\Phi}_{INT} &\sim Wishart^{-1}(I, 2)
\end{aligned}$$