LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

WIRTSCHAFTS- UND ORGANISATIONSPSYCHOLOGIE
ECONOMIC AND ORGANISATIONAL PSYCHOLOGY
CHAIR: PROF. DR. FELIX BRODBECK

WOP Working Paper
No. 2016 / 3

A Multilevel CFA–MTMM Approach for Multisource Feedback Instruments:

Presentation and Application of a New Statistical Model.

Jana Mahlke

Martin Schultze

Tobias Koch

Michael Eid

Regina Eckert

Felix C. Brodbeck

Author Note

Jana Mahlke[1] (jana.mahlke@fu-berlin.de); Martin Schultze[1] (martin.schultze@fu-berlin.de); Tobias Koch[1] (tobias.koch@fu-berlin.de); Michael Eid[1] (michael.eid@fu-berlin.de); Regina Eckert[2] (EckertR@ccl.org); Felix C. Brodbeck[3] (brodbeck@psy.lmu.de)

[1] Freie Universitaet Berlin, Department of Psychology, Division of Methods and Evaluation, Freie Universitaet Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany; Phone: +49 30 838 55638

[2]Center for Creative Leadership, Rue Neerveld 101-103 Neerveldstraat, 1200 Brussels, Belgium; Phone: +32 2 679 09 19

[3] Ludwig-Maximilians-Universitaet Muenchen, Department of Psychology, Chair of Economic and Organizational Psychology, Leopoldstraße 13, 80802 Munich, Germany; Phone: +49 89 2180 5201

Correspondence concerning this article should be addressed to Jana Mahlke.

Please cite this paper as follows:

Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R., & Brodbeck, F. C. (2016). A Multilevel CFA–MTMM Approach for Multisource Feedback Instruments: Presentation and Application of a New Statistical Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 91-110. DOI: 10.1080/10705511.2014.990153.

Abstract

Multisource feedback instruments are a widely used tool in human resource management. However, comprehensive validation studies remain scarce and there is a lack of statistical models that account appropriately for the complex data structure. Because both peers and subordinates are nested within the target but stem from different populations the assumption of traditional multilevel structural equation models that the sample on a lower level stems from the same population is violated. We present a multilevel confirmatory factor analysis multitrait-multimethod (ML-CFA-MTMM) model that considers this peculiarity of multisource feedback instruments. The model is applied to two scales of the Benchmarks® instrument and it is demonstrated how measures of reliability and of convergent and discriminant validity can be obtained using multilevel structural equation modeling software. We discuss the results as well as some implications and guidelines for the use of the model.

Keywords: confirmatory factor analysis, convergent and discriminant validity, multisource feedback, method effects

Multisource feedback assessments have gained wide popularity in human resource management within the last decades. Today, they are applied in almost all fortune 500 companies (Ghorpade, 2000; Yammarino & Atwater, 1997). The main purpose of these instruments, often referred to as 360 degree feedback, is to provide a complete and valid picture of a target's leadership behavior by collecting self-ratings as well as ratings from subordinates, peers and the target's supervisor. Additionally, external or internal clients and customers can be asked for an evaluation. An analysis and interpretation of these multi-perspective assessments—typically with a special interest in commonalities and discrepancies between the different perspectives—helps to identify individual needs for professional development.

While in other research areas a high discrepancy of multisource ratings is undesired and interpreted as an indicator of lacking convergent validity of the instruments used, in the context of multisource feedback the setting is different. Multiple sources instead of a single one evaluate a target leader because it is believed that they will disagree and will thus provide unique information (Borman, 1974; Hoffman, Lance, Bynum, & Gentry, 2010). However, some researchers have questioned the existence of rating source effects and have argued that rater effects are entirely idiosyncratic (Lebreton, Burgess, Kaiser, Atchley, & James, 2003; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Viswesvaran, Schmidt, & Ones, 2002, 2005).

Many other issues, especially concerning the "true" (i.e., free of measurement error) degree of consistency and specificity of the different perspectives, have not been clarified conclusively. This is mostly due to the fact that there are very few statistical models that account appropriately for the complex data structure of multisource feedback assessment. In consequence, questions such as "How much do self-ratings and subordinate ratings overlap?", “To what degree do subordinates and peers agree in their perception?” and “Are there rating source effects at all and how much variance is caused by the unique view of the specific raters?” cannot be answered satisfactorily. However, these are the types of questions of most import for the psychometric robustness as well as practical utility of multisource feedback as a tool for both the development and assessment of leaders. The purpose of this article is to present a multilevel confirmatory factor analysis multitrait-multimethod (ML-CFA-MTMM) model, which takes the very special data structure of multisource performance ratings into account. The model is an extension of the ML-CFA-MTMM model presented by Eid et al. (2008).

The analysis of validity and reliability of measurement instruments has a long history in psychological research. In the case of 360 degree feedback it is especially important to know how well the instrument covers the different perspectives. In other words: Is the instrument able to reflect the view of the target leaders, the view of subordinates, of peers and of supervisors? What amount of variance is usually shared by the different perspectives and what amount is rater or perspective specific? Additionally, one wants to separate these systematic variance components from unsystematic variance to draw conclusions that are free of measurement error. Because the implementation multisource feedback is very expensive, such an assessment would only be reasonable if the convergence between different raters and different rater groups is low. In the case of high convergence it would be sufficient to assess only one perspective, e.g., the leader’s view. Moreover, it is also costly to assess many different facets of leadership behavior. If the different facets are not distinctive—that is, discriminant validity is low—one would consider reducing costs and effort by assessing only one facet.

Since Campbell and Fiske’s pioneering article in 1959, MTMM analysis is probably the most popular technique to address questions of convergent and discriminant validity (Eid & Diener, 2006). In the context of multisource performance ratings, the different competencies represent the traits, whereas each rater represents a single method. In the early stages of development, MTMM analyses were based on the correlations of manifest measures only (e.g., Conway & Huffcutt, 1997; Viswesvaran, Ones, & Schmidt, 1996), but this strategy has several limitations, which can be overcome by the application of CFA models to MTMM data: Trait, method and error components can be separated from each other, assumptions of the underlying model can be tested empirically and trait and method factors can be linked to further latent variables (Eid et al., 2008). Meanwhile, many researchers use CFA-MTMM models to evaluate psychometric properties of ratings (see Kenny & Kashy, 1992; Marsh, 1989). However, as there is a vast number of different structural equation modeling approaches, one needs to choose the appropriate model for data analysis carefully. In the case of multisource feedback data, where different methods are replaced by different raters, a special problem arises from measurement designs with a varying number of raters per target (Putka, Lance, Le, & McCloy, 2011). In this situation, researchers often select a subset of a fixed number of these raters for each target for any given source (e.g., two subordinates and two peers per target) and then fit a traditional CFA-MTMM model to their data matrix. In doing so, the fact that raters are nested within targets is ignored and a large amount of information is disregarded. Putka et al. (2011) demonstrated how this technique compromises the trustworthiness of CFA-MTMM results and recommended using a multilevel CFA-MTMM strategy as presented by Eid et al. (2008) in cases of a unique, nonoverlapping set of raters per target. With a multilevel CFA-MTMM model, it is possible to account for the hierarchical data structure with multiple raters nested within one target. Eid et al.

(2008) developed a model that allows to analyze the convergence of two different types of raters: (1) several level 1 methods, that is, a group of multiple raters (e.g., subordinates) that are nested within the target (e.g., the leader) and (2) one level 2 method, e.g., a self-report of the target. However, this model cannot be applied to 360 degree feedback assessments because there are level 1 methods (peers and subordinates) that stem from two different populations. The simultaneous inclusion of several peers and several subordinates is intricate: As subordinates stem from one population and peers stem from a different population they should not be modeled as belonging to the same level 1 sample. Instead, two sets of methods should be incorporated on level 1. Unfortunately, there are so far no multilevel models that can handle this kind of data structure—even though the need is obvious not only for multisource feedback data but also for other applications. In order to close this gap, we developed a model that can handle data structures such as 360 degree feedback data with self-reports on level 2 and multiple sets of methods stemming from different populations on level 1.

The goal of this article is to enable other researchers who deal with similar data structures to adopt this new approach. Therefore, we

- briefly review the multilevel MTMM model for different types of methods of Eid et al. (2008),
- present a new model, which allows the inclusion of multiple sets of level 1 methods,
- show how available multilevel structural equation modeling software can be applied to estimate such a model, and
- illustrate this new model analyzing real 360 degree feedback data.

The Model for one Level 2 Method and one Set of Level 1 Methods

Eid et al. (2008) presented this model for a combination of several peer reports (level 1 methods) and one self-report (level 2 method). The data are analyzed with a two-level CFA-MTMM model. A correlated trait-correlated (method-1) approach (CT-C([M-1]; Eid, Lischetzke, & Nussbeck, 2006; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid, 2000; Koch, Eid, & Lochner, 2013) is used on level 2 with the self-report selected as the reference method. Three factors for each construct are defined within this framework:

- a *trait factor*, which captures the construct as measured by the indicators of the self-report,
- a *common method factor,* which depicts the part of the peer ratings that is shared among the peers but is not shared with the reference method, and
- a *unique method factor*, which represents the deviation of a single peer rater from the common view of peer raters.

As a consequence of this decomposition of the manifest variables, the following variance components of the nonreference indicators (i.e., the peer ratings) can be estimated (Carretero-Dios, Eid, & Ruch, 2011; Eid et al., 2008):

- The *consistency* of self- and peer ratings indicates the proportion of true variance that is explained by the self-report and is thus an indicator of convergent validity.
- The *common method specificity* of peer ratings indicates the proportion of true variance that is shared among the peers but not shared with the self-report.
- The *unique method specificity* of peer ratings indicates the proportion of true variance that is not shared with the other peer(s) and is thus unique to the specific peer rater.

- The *reliability* represents the proportion of manifest variance that is not due to measurement error.

In addition to the variance components, correlations between the latent factors provide estimates of further psychometric properties of the instrument used. One could for example analyze the discriminant validity of constructs by correlating the trait factors. It is also possible to correlate the common method factors (respectively the unique method factors) of different traits to examine whether the method effects generalize across traits.

The model is restricted to one set of level 1 methods. In the next section we will show how this model can be extended to a model with two sets of level 1 methods.

The ML-CFA-MTMM Model for two Sets of Level 1 Methods

The following model was designed for data situations that have two sets of level 1 methods (e.g., a set of peers and a set of subordinates) and one level 2 method (e.g., the self-report; see Figure 1). For simplicity, we will describe the model in which there is only one level 2 method (self-report) in addition to the two sets level 1 methods. The model can easily be extended to the situation of more than one method on level 2 (e.g., the supervisor report) using the approach of Carretero-Dios et al. (2011). We will explain the model for two sets of level 1 methods and a self-report on level 2 (see Figure 2) with all of its components in detail. We chose the self-report as the reference method and we use three indicators per trait that are unidimensionally measuring a common trait. Our notation follows the notation by Eid et al. (2008). Superscripts distinguish between the different methods: “RM” refers to the reference method (self-report) and "NM$_m$" to the nonreference methods with the subscript *m* to differentiate between the nonreference method of subordinates (*m* = 1) and the nonreference method of peers (*m* = 2).

The measurement model for the self-report is given by

$$Y_{tik}^{\mathrm{RM}} = \mu_{ik}^{\mathrm{RM}} + \lambda_{\mathrm{T}ik}^{\mathrm{RM}} T_{tk}^{\mathrm{RM}} + E_{tik}^{\mathrm{RM}}. \quad (1)$$

μ_{ik}^{RM} denotes the intercepts of the observed variables Y_{tik}^{RM}, $\lambda_{\mathrm{T}ik}^{\mathrm{RM}}$ represents the factor loadings on the (reference method) trait factors T_{tk}^{RM}, and E_{tik}^{RM} depicts the measurement errors on level 2. The subscripts indicate *t* = target, *i* = indicator and *k* = trait. The nonreference method indicators can be decomposed in the following way:

$$Y_{rtik}^{\mathrm{NM}_m} = \mu_{ik}^{\mathrm{NM}_m} + T_{tik}^{\mathrm{NM}_m} + \lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m} UM_{rtk}^{\mathrm{NM}_m} + E_{rtik}^{\mathrm{NM}_m}, \quad (2)$$

where $\mu_{ik}^{\mathrm{NM}_m}$ denotes the intercepts of the observed variables $Y_{rtik}^{\mathrm{NM}_m}$, $T_{tik}^{\mathrm{NM}_m}$ denotes the indicator-specific traits of the nonreference method indicators, and $\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}$ represents the loadings on the unique method factors $UM_{rtk}^{\mathrm{NM}_m}$. $E_{rtik}^{\mathrm{NM}_m}$ depicts the measurement errors on level 1. We use the additional subscript *r* for the individual raters.

The unique method factor $UM_{rtk}^{\mathrm{NM}_m}$ represents the part of the manifest variable that is due to the single rater within the set of level 1 methods (i.e., a specific subordinate or peer) and that is not shared with the common view of this set of level 1 methods. For example, a value of $UM_{rt1}^{\mathrm{NM}_1}$ is the deviation of the true rating of subordinate *r* from the expected value of subordinate ratings for target *t*. It shows the degree to which a single rater *r* deviates from the mean of all subordinate raters belonging to the same target *t*. The indicator-specific traits $T_{tik}^{\mathrm{NM}_m}$ are the expected values of $Y_{rtik}^{\mathrm{NM}_m}$ for the targets across all raters belonging to group *m*. For example, a value of $T_{t11}^{\mathrm{NM}_1}$ is the expected value for a single target *t* on the first indicator of the first trait across all subordinates belonging to this target. In other words, it is the expected value of the target-specific distribution of subordinates' true ratings. The basic idea of the model is to contrast the view of others with the view of the target person. Therefore, the trait variables of the nonreference methods $T_{tik}^{\mathrm{NM}_m}$ are regressed on the trait variables of the reference method T_{tk}^{RM}. The residuals of this latent regression are assumed to be unidimensional for the three indicators per trait. They measure the common method factor $CM_{tk}^{\mathrm{NM}_m}$, which represents that part of the common view of a set of level 1 methods that is not shared with the reference method. This decomposition of the nonreference method trait factors is expressed by:

$$T_{tik}^{\mathrm{NM}_m} = \lambda_{\mathrm{T}ik}^{\mathrm{NM}_m} T_{tk}^{\mathrm{RM}} + \lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m} CM_{tk}^{\mathrm{NM}_m} \ . \qquad (3)$$

Here, $\lambda_{Tik}^{\mathrm{NM}_m}$ denotes the factor loadings on the reference method trait factors T_{tk}^{RM}, and $\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}$ denotes the factor loadings on the common method factors $CM_{tk}^{\mathrm{NM}_m}$.

The model equation for the nonreference method indicators is obtained by inserting Equation 3 into Equation 2:

$$Y_{rtik}^{\mathrm{NM}_m} = \mu_{ik}^{\mathrm{NM}_m} + \lambda_{\mathrm{T}ik}^{\mathrm{NM}_m} T_{tk}^{\mathrm{RM}} + \lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m} CM_{tk}^{\mathrm{NM}_m} + \lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m} UM_{rtk}^{\mathrm{NM}_m} + E_{rtik}^{\mathrm{NM}_m}. \qquad (4)$$

In this model, there are two types of method factors. On the one hand, a value of a common method factor represents the deviation of the expected value of a rater group (the "average" view of a rater group) from the value expected by the self-report. When the value is positive, a target receives higher values from his or her raters compared to all other targets having the same self-rated trait score, i.e. the target is overestimated. A negative value shows that a target's trait is underestimated by his or her rater group. Hence, a value of the common method factor depends on the target and the rater group rating this target. The variance of the common method factor indicates how large the differences are between rater groups who rate targets with the same trait values. On the other hand, a value of the unique method factor indicates to which degree a single rater deviates from his or her group of raters. When the value is positive, the single rater overestimates the target compared to the other raters belonging to the same group. When the value is negative, the single rater underestimates the target compared to the other raters belonging to the same group. The variance of the unique method factor shows how dissimilar single raters are.

For identification purposes, the means of the latent traits T_{tk}^{RM} and $T_{tik}^{\mathrm{NM}_m}$ are fixed to zero and the first factor loading of all factors is set to one. As the method factors $CM_{tk}^{\mathrm{NM}_m}$ and $UM_{rtk}^{\mathrm{NM}_m}$ are residual factors, their means are also zero. The following variables in the model are uncorrelated (see Eid et al., 2008): (a) trait variables, common method factors and unique method factors with error variables, (b) trait variables and common method factors with unique method factors, (c) trait variables with common method factors of the same trait and (d) error variables with each other. Furthermore, it is not possible for the unique method factors of peers and the unique method factors of subordinates to be correlated because they are based on two distinct sets of raters.

As the latent variables on the right side of Equation 4 are uncorrelated, the variance of an observed nonreference method variable can be decomposed in the following way:

$$Var\left(Y_{rtik}^{\mathrm{NM}_m}\right) = \left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tk}^{\mathrm{RM}}\right) + \left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right) + \left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right) + Var\left(E_{rtik}^{\mathrm{NM}_m}\right). \quad (5)$$

The coefficients of consistency and of common and unique method specificity can now be estimated in the same way as proposed by Eid et al. (2008). Table 1 shows how to obtain these components. The meaning of the coefficients is explained in Table 2.

There are also some important correlations in the model (see Figure 2). The correlation between the trait factors indicates the amount of discriminant validity. Correlations between common method factors show whether the method effects generalize across traits and/or across the two sets of nonreference methods. The unique method factors can only be correlated across traits but not across nonreference methods. The correlations between the unique method factors show to which degree the unique method effects generalize across traits. Table 2 gives an overview of the meaning and interpretation of the latent factors, the variance components, and the correlations of the model.

After explaining the measurement equations and the meaning of the latent variables in the model the question arises how it can be defined within the statistical software since there does not exist a preinstalled command or option for this type of model. We will first present the sample and measures and then explain the practical implementation of the model using Mplus6 (L. K. Muthén & Muthén, 1998-2010).

Application of the Model to 360 Degree Feedback Data

Sample and Measures

We applied the model to 360 degree feedback data collected by the Center for Creative Leadership® in a US American sample. The dataset included 6,065 targets (level 2 units) who rated themselves and who were rated by 27,418 subordinates and 24,847 peers (both level 1 units) with a varying number of subordinates (range: 0-38, *M* = 4.5) and peers (range: 0-17, *M* = 4.1) per target. All participants completed the Benchmarks® instrument (Lombardo, McCauley, McDonald-Mann, & Leslie, 1999), one of the most frequently used instruments of 360 degree feedback for leadership development. Benchmarks® comprises 16 competency and five derailment scales, of which we chose two competency scales. We analyzed only two of the 21 scales because the main purpose of the following section is to present the model not only formally but also in its application. Therefore, we focus on the illustration of the model by reducing the number of all other model parameters to a minimum. Certainly, the model can be extended to more than two scales given a sufficient sample size. The first scale *Leading Employees* consists of 13 items measuring how much the target leader attracts, motivates, and develops employees. The second scale *Participative Management* has nine items measuring how much the target leader involves others, listens, and builds commitment. All items have a possible range from 1 (*to a very little extent*) to 5 (*to a very great extent*). Preliminary analyses demonstrated that both constructs are assessed unidimensionally by the 13 respectively nine items. Factor loadings were used to allocate the items to three parcels per scale following the recommendations of Little, Cunningham, Shahar, and Widaman (2002) to achieve item-to-construct balance. Parcels were built by averaging the respective items. For Leading Employees the first parcel consists of five items and the second and third parcel consist of four items each. The three parcels of Participative Management contain three items each. The main reasons for building item parcels were to increase the reliability of the indicators (Little et al., 2002), to reduce the number of manifest variables, to ensure continuous observed variables, and again to minimize the model complexity for illustration purposes.

Practical Implementation of the Model

We conducted the ML-CFA-MTMM analysis with Mplus 6 (L. K. Muthén & Muthén, 1998-2010) using a robust maximum likelihood (MLR) estimator. The program assumes that there is only one set of level 1 units for each level 2 unit. In our application, however, we have two sets of level 1 units per level 2 unit. This is in contrast to traditional multilevel analyses. We solved this problem in the following way:

1. Formally, we consider all 27,418 subordinates and 24,847 peers as level 1 units as in a traditional multilevel analysis. This means that there are 52,265 level 1 units.
2. In contrast to traditional multilevel analysis, we consider the ratings belonging to the two different sets of raters as two different types of observed variables. For the assessment of Leading Employees, we have three indicators for subordinates and three different indicators for peers. The same is true for Participative Management.
3. As peers cannot have values on subordinate ratings, peers have missing values on all subordinate indicators. Conversely, subordinates have missing values on all peer indicators.

This makes it possible to include 52,265 level 1 units but to distinguish between the two rater groups.

This type of analysis requires data that are organized in a long format with as many lines per target as there are nonreference reports for the target and with one column for each parcel-method combination. The example in Table 3 shows the data matrix for three parcels (*par1-par3*) assessed by one reference method (*rm*) and two sets of nonreference methods (*nm1* and *nm2*). There are five individual nonreference raters–and therefore five lines–for the first target (*ID*=1). The columns 2-4 contain the parcel values of the self-rating (reference method). These parcel values are copied into all lines belonging to the same target. Columns 5-7 contain the parcel values of the nonreference method 1 (e.g., subordinates), the last three columns display the parcel values of the nonreference method 2 (e.g., peers).

The values of the three individual raters of the nonreference method 1 who rated the first target are in lines 1-3, columns 5-7. As the lines 4-5 are reserved for the raters of the nonreference method 2, there are logical missing values (NA) on the nonreference method 1 variables. The values of the two individual raters of the nonreference method 2 are in lines 4-5, columns 8-10, and the lines 1-3 contain missing values on the nonreference method 2 variables. The values for the next target follow in lines 6-9 with one individual rater of the nonreference method 1 and three individual raters of the nonreference method 2.

Results

The ML-CFA-MTMM model shown in Figure 2 fits the data well, $\chi^2(176, N = 52{,}265) = 5{,}935.141$, $p < .001$, RMSEA = .025, CFI = .98, SRMR (level 1) = .01, SRMR (level 2) = .04. The means, the loading parameters and the coefficients of consistency, common and unique method specificity, and reliability are presented in Table 4.

The consistency coefficients showed that between 1% and 2% of the error-free nonreference method variance is shared with the self-report. These results revealed a very low convergent validity of self-ratings and ratings from subordinates respectively peers.[1] The amount of true variance that is shared among the group of subordinates (or among the group of peers) but not with the self-report (common method specificity) varied between 21% and 28%. The unique method specificity was by far the major source of variance and indicated that between 70% and 77% of true variance was neither shared with the self-report nor with the other subordinates respectively peers but was specific to the single nonreference method raters. All reliabilities were acceptable to very good (between .54 and .90), especially when considering that each parcel consists of only three to five items.

To get a better understanding of the relationships between the different factors, we analyzed the correlations between trait, common method, and unique method factors (Table 5). The high correlation of *r* = .84 between the two trait factors showed that there was low discriminant validity of the two constructs. It was also striking that the common method effect generalized across constructs on the one hand (*r* = .96 for subordinates and *r* = .95 for peers) and across subordinates and peers on the other hand (*r* = .71 for Leading Employees and *r* = .68 for Participative Management). This means (a) that a group of subordinates or peers who over- vs. underestimated the target in Leading Employees also over- vs. underestimated this target in Participative Management (and vice versa) and (b) that a target who was over- vs. underestimated by subordinates was also over- vs. underestimated by peers (and vice versa). Common method factors were even strongly correlated across different traits *and* different sets of nonreference methods (*r* = .68 and *r* = .67), indicating that targets who were over- vs. underestimated by their peers on one trait were also over- vs. underestimated by their subordinates on the other trait. The generalizability also held for the unique method factors (*r* = .94 for subordinates and *r* = .91 for peers), indicating that a single rater who over- vs.

underestimates a target with respect to Leading Employees also tended to over- vs. underestimate the target with respect to Participative Management.

Discussion

The model presented in this paper is the first ML-CFA-MTMM model for data structures with two (or more) populations of level 1 methods that are both nested within the targets' self-reports. It overcomes many problems that are often associated with the analysis of MTMM data (Putka et al., 2011): The researcher no longer has to choose a fixed number of raters per target but can include all available raters and model multiple sets of raters that are nested within the targets. We used the new model for a validation of two scales from the Benchmarks® instrument (McCauley, Lombardo, & Usher, 1989) to demonstrate the estimation of the reliability of the indicators, the consistency (convergent validity) of the methods, the discriminant validity of the constructs, and the common and unique method specificity of subordinates and peers. The results reveal acceptable to very good parcel reliabilities for the two Benchmarks® scales. The discriminant validity, however, is low. The high correlation is not that astonishing as the two scales capture related facets and both refer to the focus of Leading Others within the three main focus areas of the Benchmarks® instrument (the other two focus areas being Leading Self and Leading the Organization; Center for Creative Leadership, 2010). Moreover, as many former studies found a considerable amount of overlap between the individual performance dimensions in 360 degree feedback instruments (e.g., Beehr, Ivanitskaya, Hansen, Erofeev, & Gudanowski, 2001; Hoffman et al., 2010; Kets de Vries, Vrignaud, & Florent-Treacy, 2004; van der Zee, Zaal,

& Piekstra, 2003), such weak discriminant validity among various scales suggests that raters perceive a leader’s competencies in a holistic fashion.

The consistency between self-reports and subordinates and between self-reports and peers is very low. It is often argued that a high discrepancy between self-ratings and observer ratings is an indicator of a manager’s lack of self-awareness (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Kulas & Finkelstein, 2007). However, this interpretation is highly one-sided as it assumes that the observer ratings of a manager’s leadership competencies are in some way more “truthful” than their self-ratings. Many other reasons for discrepancies that are partly not under the manager’s control such as differing definitions of “good leadership” between the rating sources, differing opportunities to observe the target leader’s behavior (Harris & Schaubroeck, 1988; Morgeson, Mumford, & Campion, 2005) and cultural influences (Atwater, Wang, Smither, & Fleenor, 2009; Eckert, Ekelund, Gentry, & Dawson, 2010; House, Hanges, Javidan, Dorfman, & Gupta, 2004) are discussed in the literature (for a review see Fleenor, Smither, Atwater, Braddy, & Sturm, 2010). Regardless, the lack of consistency is a strong argument for the benefits of multisource feedback assessments, as it proves that considering different rating sources actually results in more information.

One fundamental benefit of our new model is that the consistency of ratings can be evaluated not only between the different perspectives but also among the individual peers and among the individual subordinates. The consistency within both of these groups is considerably higher than the consistency between self-reports and others’ ratings. This indicates that subordinates (respectively peers) share a common view that differs from the target’s self-perception. Furthermore, the model allows to state that the common view generalizes across nonreference methods and across traits: On the one hand, subordinates (respectively peers) who over- or underestimate the target in their common rating of Leading Employees tend to have the same bias in their common rating of Participative Management. This association is very strong and shows that the rater groups might not distinguish between these facets. On the other hand, subordinates and peers tend to have the same common bias in rating the target. The analysis of this correlation on the latent level is enabled for the first time by our model. Our findings show that there is high agreement between subordinates and peers in the evaluation of a given manager. It raises an important question that could have essential implications for the application of multisource feedback programs: If subordinates and peers strongly agree in their perception would it not be more economical to include only one of these perspectives in the assessment? Further studies need to shed light on this issue. It is especially interesting to analyze whether the degree of agreement depends on the different competencies that are usually rated in a multisource feedback and on their observability to the different raters. Peers may, for example, have more opportunities than subordinates to observe the target's interaction with the supervisor (Morgeson et al., 2005) whereas subordinates are predestined to evaluate the target's leadership behavior. The agreement between peers and subordinates may be higher for competencies that are observable for both groups and lower for competencies that are more accessible to one of the groups.

The unique view of the different subordinates and the different peers is by far the major source of variance. This result is in agreement with previous studies (Greguras & Robie, 1998; Lance, 1994; Scullen, Mount, & Goff, 2000; Woehr, Sheehan, & Bennett Jr., 2005) and confirms Yammarino's (2003) assumption that "multisource ratings may inform us more about the rater providing the data and his or her views rather than about the focal manager who is being rated and his or her actual performance" (p. 9-10). However, as decisions in organizations are made by the very people providing ratings in multisource feedback and are based on their perceptions, gathering this information is nonetheless of high practical value.

The major contribution of this paper is that it introduces a statistical model that separates on a latent level two different kinds of method effects commonly encountered in multisource feedback (unique and common method effects). Consequently, researchers can use the model in future studies to analyze the causes of method effects. Further level 2 variables can be included to explain common method effects, for example, variables characterizing the target (e.g., gender, duration of leadership experience, personality variables) or the group of raters (e.g., team climate, team satisfaction). Additional level 1 variables characterizing the individual raters (e.g., gender, duration of employment, income, job satisfaction, personality variables) can be added to explain unique method effects.

Limitations

The model presented in this article depends on some requirements and assumptions, which can be erroneous in specific applications. First, it is assumed that all raters are fully nested with unique, nonoverlapping sets of subordinates and peers per target. Every rater is allowed to provide exactly one rating in the data set – either as a target *or* as a subordinate *or* as a peer. In an application of the model it should be made sure that this requirement is fulfilled, particularly when applying it to company-specific datasets. In their simulation study Schultze, Koch, and Eid (2013) analyzed the effect of including raters who provide ratings for up to ten targets, e.g., peers who assess two or more colleagues. This study uses a sub-model of the model presented here with only one set of methods on level 1. The results indicate that the existence of raters evaluating more than one target has no effect on the decomposition of variance that was presented above because parameter estimates were only minimally distorted. In contrast, there is an estimation bias of some standard errors on level 1 (independent of sample size) and on level 2 (partly reducible by increasing level 1 sample size). In sum, all interpretations in this paper would remain valid even if there had been subordinates or peers that assess multiple targets because we analyzed variance components and correlations but draw no inferential conclusions.

Second, multilevel modeling with latent variables requires large sample sizes. According to the simulation study by Hox and Maas (2001), the sample size on level 2 is more important than on level 1 and should comprise at least 100 level 2 units. Further simulation studies are necessary to reveal whether this recommendation holds for this type of model as well. Third, the model was defined and presented for indicators that measure a common trait. In applications where this assumption is too strict, the model can easily be extended to a model with indicator-specific trait variables (see Eid et al., 2008). Finally—as the focus here was to demonstrate the consideration of two sets of level 1 methods—we didn't include supervisor ratings in our analysis even though they are usually assessed in a 360 degree feedback. Carretero-Dios et al. (2011) demonstrated how these ratings could be easily integrated as additional level 2 methods.

Conclusions

We presented a model that offers many new opportunities for researchers who wish to analyze MTMM data with multiple sets of level 1 methods. The model can handle a varying number of ratings per target, it accounts for the multilevel structure of the data and it captures the common and the unique method factor on the latent level, thus, free of measurement error. The consistency can be obtained not only between self-reports and subordinates or peers but also among the subordinates and among the peers. Finally, the model can be supplemented with any dependent or independent variables to serve as a basis for causal analyses in the context of leadership research.

Acknowledgements

The authors would like to thank the Center for Creative Leadership® for providing the data used in this research.

References

Atwater, L., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology*, *51*, 577–598. doi:10.1111/j.1744-6570.1998.tb00252.x

Atwater, L., Wang, M., Smither, J. W., & Fleenor, J. W. (2009). Are cultural characteristics associated with the relationship between self and others' ratings of leadership? *Journal of Applied Psychology*, *94*, 876–86. doi:10.1037/a0014561

Beehr, T. A., Ivanitskaya, L., Hansen, C. P., Erofeev, D., & Gudanowski, D. M. (2001). Evaluation of 360 degree feedback ratings: Relationships with each other and with performance and selection predictors. *Journal of Organizational Behavior*, *22*, 775–788. doi:10.1002/job.113

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance*, *12*, 105–124. doi:10.1016/0030-5073(74)90040-3

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

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

Eckert, R., Ekelund, B. Z., Gentry, W. A., & Dawson, J. F. (2010). "I don't see me like you see me, but is that a problem?" Cultural influences on rating discrepancy in 360-degree feedback instruments. *European Journal of Work and Organizational Psychology, 19*, 259–278. doi:10.1080/13594320802678414

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 multimethod measurement in psychology* (pp. 3–9). Washington, DC: American Psychological Association.

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

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, *8*, 38–60. 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,* 230–53. doi:10.1037/a0013219

Fleenor, J. W., Smither, J. W., Atwater, L., Braddy, P. W., & Sturm, R. E. (2010). Self-other rating agreement in leadership: A review. *The Leadership Quarterly, 21,* 1005–1034. doi:10.1016/j.leaqua.2010.10.006

Ghorpade, J. (2000). Managing five paradoxes of 360-degree feedback. *Academy of Management Perspectives*, *14*(1), 140–150. doi:10.5465/AME.2000.2909846

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, *83*, 960–968. doi:10.1037//0021-9010.83.6.960

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*, 43–62. doi:10.1111/j.1744-6570.1988.tb00631.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, *63*, 119–151. doi:10.1111/j.1744-6570.2009.01164.x

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P., & Gupta, V. (Eds.). (2004). *Leadership, culture, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage Publications.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*, 157–174. doi:10.1207/S15328007SEM0802_1

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

Kets de Vries, M. F. R., Vrignaud, P., & Florent-Treacy, E. (2004). The Global Leadership Life Inventory: Development and psychometric properties of a 360-degree feedback instrument. *The International Journal of Human Resource Management*, *15*, 475–492. doi:10.1080/0958519042000181214

Koch, T., Eid, M., & Lochner, K. (2013). Multitrait-multimethod-analysis: The psychometric foundation of multitrait-multimethod (MTMM) models. In *Handbook of psychometric testing*. Hoboken, New Jersey: Wiley-Blackwell. Manuscript submitted for publication.

Kulas, J. T., & Finkelstein, L. M. (2007). Content and reliability of discrepancy-defined self-awareness in multisource feedback. *Organizational Research Methods*, *10*, 502–522. doi:10.1177/1094428107301100

Lance, C. E. (1994). Test of a latent structure of performance ratings derived from Wherry's (1952) theory of rating. *Journal of Management, 20*, 757–771. doi:10.1177/014920639402000404

Lebreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80–128. doi:10.1177/1094428102239427

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173. doi:10.1207/S15328007SEM0902_1

Lombardo, M. M., McCauley, C. D., McDonald-Mann, D., & Leslie, J. B. (1999). *BENCHMARKS® Developmental Reference Points.* Greensboro, NC: Center for Creative Leadership.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361. doi:10.1177/014662168901300402

McCauley, C. D., Lombardo, M. M., & Usher, C. J. (1989). Diagnosing management development needs: An instrument based on how managers develop. *Journal of Management, 15*, 389–403. doi:10.1177/014920638901500303

Morgeson, F. P., Mumford, T. V., & Campion, M. A. (2005). Coming full circle: Using research and practice to address 27 questions about 360-degree feedback programs. *Consulting Psychology Journal: Practice and Research, 57,* 196–209. doi:10.1037/1065-9293.57.3.196

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology*, *51*, 557–577. doi:10.1111/j.1744-6570.1998.tb00251.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. doi:10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A cautionary note on modeling multitrait-multirater data arising from ill-structured measurement designs. *Organizational Research Methods*, *14*, 503–529. doi:10.1177/1094428110362107

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*, 956–970. doi:10.1037/0021-9010.85.6.956

Van der Zee, K. I., Zaal, J. N., & Piekstra, J. (2003). Validation of the multicultural personality questionnaire in the context of personnel selection. *European Journal of Personality*, *17*, 77–100. doi:10.1002/per.483

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. doi:10.1037/0021-9010.81.5.557

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence. *Journal of Applied Psychology, 87*, 345–354. doi:10.1037//0021-9010.87.2.345

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131. doi:10.1037/0021-9010.90.1.108

Woehr, D. J., Sheehan, M. K., & Bennett Jr., W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology*, *90*, 592–600. doi:10.1037/0021-9010.90.3.592

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. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics*, *25*(4), 35–44. doi:10.1016/S0090-2616(97)90035-8

Table 1

Definition of the variance components for the indicators of the nonreference methods

$$CO\left(Y_{rtik}^{\mathrm{NM}_m}\right) = \frac{\left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tk}^{\mathrm{RM}}\right)}{\left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tk}^{\mathrm{RM}}\right) + \left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right) + \left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right)}$$

$$CMS\left(Y_{rtik}^{\mathrm{NM}_m}\right) = \frac{\left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right)}{\left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tk}^{\mathrm{RM}}\right) + \left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right) + \left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right)}$$

$$UMS\left(Y_{rtik}^{\mathrm{NM}_m}\right) = \frac{\left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right)}{\left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tk}^{\mathrm{RM}}\right) + \left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right) + \left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right)}$$

$$Rel\left(Y_{rtik}^{\mathrm{NM}_m}\right) = 1 - \frac{Var\left(E_{rtik}^{\mathrm{NM}_m}\right)}{Var(Y_{rtik}^{\mathrm{NM}_m})}$$

Note. CO = consistency coefficient; *CMS* = common method specificity coefficient; *UMS* = unique method specificity coefficient; *Rel* = reliability coefficient; $Y_{rtik}^{\mathrm{NM}_m}$ = observed variables of the nonreference methods; T_{tk}^{RM} = trait factors; $CM_{tk}^{\mathrm{NM}_m}$ = common method factors; $UM_{rtk}^{\mathrm{NM}_m}$ = unique method factors; $E_{rtik}^{\mathrm{NM}_m}$ = error variables; $\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}$ = trait factor loadings; $\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}$ = common method factor loadings; $\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}$ = unique method factor loadings; *r* = rater; *t* = target; *i* = indicator; *k* = trait; *m* = nonreference method; *Var* = variance.

Table 2

Overview of latent factors, variance components and factor correlations

Component	Description	Explanation
Latent factors		
T_{tk}^{RM}	Trait factor of the reference method	Unidimensional latent trait variable of the reference method, measured by the self-report indicators
$T_{tik}^{\mathrm{NM}_m}$	Trait factor of the nonreference method	Indicator-specific latent trait variable of the nonreference method, measured by the subordinate or peer indicators
$CM_{tk}^{\mathrm{NM}_m}$	Common method factor	Trait-specific common view of subordinates or peers that is not shared with the target's view
$UM_{rtk}^{\mathrm{NM}_m}$	Unique method factor	Unique deviation of a single subordinate or peer rating from the common view of subordinates or peers for a given target
E_{tik}^{RM}	Measurement error on level 2	Measurement error on level 2
$E_{rtik}^{\mathrm{NM}_m}$	Measurement error on level 1	Measurement error on level 1
Variance components		
$CO\left(Y_{rtik}^{\mathrm{NM}_m}\right)$	Consistency	Proportion of true variance that is shared with the self-report, thus an indicator of convergent validity of self-report and nonreference method (group of subordinates or peers)
$CMS\left(Y_{rtik}^{\mathrm{NM}_m}\right)$	Common method specificity	Proportion of true variance that is due to the common view of subordinates or peers not shared with the self-report
$UMS\left(Y_{rtik}^{\mathrm{NM}_m}\right)$	Unique method specificity	Proportion of true variance that is due to single views of the nonreference method raters not shared with the self-report and not shared with other members of the same rater group (subordinates or peers)
$Rel\left(Y_{rtik}^{\mathrm{NM}_m}\right)$	Reliability	Proportion of manifest variance that is not due to measurement error
Factor correlations		
$Cor(T_{t1}^{\mathrm{RM}}, T_{t2}^{\mathrm{RM}})$	Discriminant validity	Degree to which the two traits (measured by the self-report) are related
$Cor(CM_{t1}^{\mathrm{NM}_1}, CM_{t1}^{\mathrm{NM}_2})$, $Cor(CM_{t2}^{\mathrm{NM}_1}, CM_{t2}^{\mathrm{NM}_2})$	Generalizability of the common method factors across the two sets of nonreference methods	Degree to which the trait-specific over- vs. underestimation by subordinates is related to the over- vs. underestimation by peers (on the same trait)
$Cor(CM_{t1}^{\mathrm{NM}_1}, CM_{t2}^{\mathrm{NM}_1})$, $Cor(CM_{t1}^{\mathrm{NM}_2}, CM_{t2}^{\mathrm{NM}_2})$	Generalizability of the common method factors across traits	Degree to which the over- vs. underestimation by subordinates or by peers on one trait is related to the over- vs. underestimation on the other trait (measured by the same nonreference method)
$Cor(CM_{t1}^{\mathrm{NM}_1}, CM_{t2}^{\mathrm{NM}_2})$, $Cor(CM_{t2}^{\mathrm{NM}_1}, CM_{t1}^{\mathrm{NM}_2})$,	Generalizability of the common method factors across traits and nonreference methods	Degree to which the over- vs. underestimation by subordinates or peers on one trait is related to the over- vs. underestimation by the other nonreference method on the other trait
$Cor(UM_{rt1}^{\mathrm{NM}_1}, UM_{rt2}^{\mathrm{NM}_1})$, $Cor(UM_{rt1}^{\mathrm{NM}_2}, UM_{rt2}^{\mathrm{NM}_2})$	Generalizability of the unique method factors across traits	Degree to which the over- vs. underestimation by the single raters (from a group of subordinates or peers) on one trait is related to the over- vs. underestimation on the other trait (measured by the same individual rater)
$Cor(T_{t1}^{\mathrm{RM}}, CM_{t2}^{\mathrm{NM}_1})$,	Correlation between a reference method trait factor	Degree to which the common bias of subordinates or peers on one trait is related to the other trait (measured by the self-

$Cor(T_{t1}^{\mathrm{RM}}, CM_{t2}^{\mathrm{NM}_2})$, $Cor(T_{t2}^{\mathrm{RM}}, CM_{t1}^{\mathrm{NM}_1})$, $Cor(T_{t2}^{\mathrm{RM}}, CM_{t1}^{\mathrm{NM}_2})$	and a common method factor of the other trait	report)

Note. Y_{tik}^{RM} =observed variables of the reference methods; $Y_{rtik}^{\mathrm{NM}_m}$ = observed variables of the nonreference methods; *r* = rater; *t* = target; *i* = indicator; *k* = trait; *m* = nonreference method.

Table 3

Data format for the application of the model for one level 2 method and two sets of level 1 methods

	Columns									
	1	2	3	4	5	6	7	8	9	10
Rows	*ID*	*par1_ rm*	*par2_ rm*	*par3_ rm*	*par1_ nm1*	*par2_ nm1*	*par3_ nm1*	*par1_ nm2*	*par2_ nm2*	*par3_ nm2*
1	1	3.5	4	5	5	4	3.5	NA	NA	NA
2	1	3.5	4	5	4.5	5	4	NA	NA	NA
3	1	3.5	4	5	2	3.5	5	NA	NA	NA
4	1	3.5	4	5	NA	NA	NA	4.5	4	3
5	1	3.5	4	5	NA	NA	NA	3.5	4	4.5
6	2	4.5	3	2.5	4	3.5	4	NA	NA	NA
7	2	4.5	3	2.5	NA	NA	NA	4	3	3
8	2	4.5	3	2.5	NA	NA	NA	3	3.5	2
9	2	4.5	3	2.5	NA	NA	NA	4.5	2.5	2.5

Note. ID = identification variable for the target; *par1-par3* = three parcels of one construct; *rm* = reference method; *nm1* = nonreference method 1 (e.g., subordinates); *nm2* = nonreference method 2 (e.g., peers), NA = (logical) missing value.

Table 4

Means, Factor Loadings, and Coefficients of Consistency, Common Method Specificity, Unique Method Specificity, and Reliability

	Means	Trait factor loading	Common method factor loading	Unique method factor loading	Consistency	Common method specificity	Unique method specificity	Reliability
Indicator	$\mu_{ik}^{\text{RM}}/\mu_{ik}^{\text{NM}m}$	$\lambda_{\text{T}ik}^{\text{RM}}/\lambda_{\text{T}ik}^{\text{NM}m}$	$\lambda_{\text{CM}ik}^{\text{NM}m}$	$\lambda_{\text{UM}ik}^{\text{NM}m}$	$CO(Y_{rtik}^{\text{NM}m})$	$CMS(Y_{rtik}^{\text{NM}m})$	$UMS(Y_{rtik}^{\text{NM}m})$	$Rel(Y_{tik}^{\text{RM}})/$ $Rel(Y_{rtik}^{\text{NM}m})$
				Leading Employees				
Self								
Y_{t11}^{RM}	3.92	1.00						0.78
Y_{t21}^{RM}	4.02	0.95						0.54
Y_{t31}^{RM}	3.81	1.03						0.67
Subordinates								
Y_{rt11}^{NM}	4.00	0.25	1.00	1.00	0.02	0.24	0.74	0.90
Y_{rt21}^{NM}	4.14	0.23	0.82	0.81	0.02	0.24	0.73	0.70
Y_{rt31}^{NM}	4.00	0.27	0.97	0.96	0.02	0.25	0.73	0.84
Peers								
Y_{rt11}^{NM}	3.95	0.18	1.00	1.00	0.01	0.24	0.75	0.87
Y_{rt21}^{NM}	3.96	0.21	0.99	0.88	0.02	0.28	0.70	0.67
Y_{rt31}^{NM}	3.94	0.21	0.99	0.94	0.02	0.26	0.72	0.78
				Participative Management				
Self								
Y_{t12}^{RM}	4.04	1.00						0.73
Y_{t22}^{RM}	4.02	0.92						0.63
Y_{t32}^{RM}	3.92	0.94						0.63
Subordinates								
Y_{rt12}^{NM}	4.10	0.22	1.00	1.00	0.02	0.23	0.76	0.88
Y_{rt22}^{NM}	4.09	0.22	1.05	0.96	0.02	0.26	0.72	0.83
Y_{rt32}^{NM}	4.01	0.21	0.93	0.96	0.02	0.21	0.77	0.80
Peers								
Y_{rt12}^{NM}	4.03	0.19	1.00	1.00	0.02	0.22	0.77	0.86
Y_{rt22}^{NM}	4.03	0.19	1.05	0.96	0.02	0.25	0.74	0.82
Y_{rt32}^{NM}	3.94	0.18	0.95	0.96	0.02	0.22	0.77	0.78

Note. Y_{tik}^{RM} = observed variables of the reference method; Y_{rt11}^{NM} = observed variables of the nonreference method; *r* = rater; *t* = target; *i* = indicator; *k* = trait; *m* = nonreference method. For identification purposes the first factor loading of all factors is set to one.

Table 5

Factor Variances and Factor Correlations

		1	2	3	4	5	6	7	8	9	10
1.	T_{t1}^{RM}	*.17*									
2.	T_{t2}^{RM}	.84*	*.20*								
3.	$CM_{t1}^{\mathrm{NM}_1}$	X	.00	*.14*							
4.	$CM_{t1}^{\mathrm{NM}_2}$	X	.03	.71*	*.10*						
5.	$CM_{t2}^{\mathrm{NM}_1}$	.01	X	.96*	.67*	*.12*					
6.	$CM_{t2}^{\mathrm{NM}_2}$	-.07*	X	.68*	.95*	.68*	*.10*				
7.	$UM_{rt1}^{\mathrm{NM}_1}$	X	X	X	X	X	X	*.42*			
8.	$UM_{rt1}^{\mathrm{NM}_2}$	X	X	X	X	X	X	X	*.31*		
9.	$UM_{rt2}^{\mathrm{NM}_1}$	X	X	X	X	X	X	.94*	X	*.41*	
10.	$UM_{rt2}^{\mathrm{NM}_2}$	X	X	X	X	X	X	X	.91*	X	*.34*

Note: Estimated variances are in the main diagonal (italicized). Estimated correlations are in the subdiagnoal.

X = nonadmissible correlations. T_{tk}^{RM} = trait factors; $CM_{tk}^{\mathrm{NM}_m}$ = common method factors; $UM_{rtk}^{\mathrm{NM}_m}$ = unique method factors; *r* = rater; *t* = target; *i* = indicator; *k* = trait (1 = Leading Employees, 2 = Participative Management); *m* = nonreference method (1 = subordinates, 2 = peers).

*$p < .001$

Figure Captions

Figure 1. Example of two sets of level 1 methods (peers and subordinates) and one level 2 method (self-report).

Figure 2. Multilevel confirmatory factor analysis multitrait-multimethod model for one level 2 method and two sets of level 1 methods. Y_{tik}^{RM}/ $Y_{rtik}^{\mathrm{NM}_m}$ = observed variables; T_{tk}^{RM}/ $T_{tik}^{\mathrm{NM}_m}$ = trait factors; $CM_{tk}^{\mathrm{NM}_m}$ = common method factors; $UM_{rtk}^{\mathrm{NM}_m}$ = unique method factors; E_{tik}^{RM}/ $E_{rtik}^{\mathrm{NM}_m}$ = error variables; $\lambda_{\mathrm{T}ik}^{\mathrm{RM}}$/ $\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}$ = trait factor loadings; $\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}$ = common method factor loadings; $\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}$ = unique method factor loadings; *r* = rater; *t* = target; *i* = indicator; *k* = trait; *m* = nonreference method. For simplicity reasons, only one loading parameter per trait-method unit is depicted for the first construct.

Footnotes

[1] Please note that the indicator used here to quantify convergent validity is a variance component and not—as it is more common—a correlation coefficient.