Fachbereich Erziehungswissenschaft und Psychologie der Freien Universität Berlin

Validation of 360-Degree Feedback Assessments –

Development, Evaluation, and Application of a Multilevel Structural Equation Model

Dissertation

zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. phil.)

vorgelegt von

Dipl.-Psych.

Jana Mahlke

Berlin, Dezember 2018

Erstgutachter:

Prof. Dr. Michael Eid (Freie Universität Berlin)

Zweitgutachter:

Prof. Dr. Stefan Krumm (Freie Universität Berlin)

Disputation am 12. Februar 2019

Table of Contents

	Acknowledgements	I
	Summary	II
	Zusammenfassung	V
	List of Figures	IX
	List of Tables	X
Chapter 1	Introduction	1
Chapter 2	A Multilevel CFA–MTMM Approach for Multisource Feedback Instruments: Presentation and Application of a New Statistical Model	27
Chapter 3	Analysing Multisource Feedback with Multilevel Structural Equation Models: Pitfalls and Recommendations from a Simulation Study	79
Chapter 4	Disagreement in Self-Other-Ratings on Leadership Competencies: Are Managers Lacking Self-Awareness?	133
Chapter 5	General Discussion	185
	References	209
	Eigenständigkeitserklärung	XI

Acknowledgements

Just like most precious things in life, the work on this thesis entailed ups and downs. I am grateful to have so many supporting people around me to enjoy the ups and help me through the downs. First of all, I want to thank my first supervisor, Michael Eid. Hearing his laughter down the corridor always was contagious and his warm-heartedness made me like my job. Michael knew when it was time to let me work autonomously and when to give me helpful guidance. This work would have not been finished without Michael's support and his admirable expertise. I also want to thank Stefan Krumm who was immediately willing to be the second supervisor of this thesis and whose expertise in psychological diagnostics is highly appreciated. I was very lucky over the years with my colleagues and with the tradition of mutual support in Michael's team. Student assistants helped me in the process of preparing the manuscripts, thank you!

Finally, I want to thank my family: My parents who support and trust me whatever I do, my sister who has helped me through so much more than this, and Moritz, Emil, and Paula, who make me remember every day that there are things in life that are more important than a dissertation.

Abstract

The aim of this thesis is to develop, evaluate, and apply a multilevel structural equation model for the validation of 360-degree feedback instruments. The concept of these instruments is that a manager evaluates his or her own leadership competencies and is additionally evaluated by peers, subordinates, and the supervisor to gain a multi-perspective understanding of the managers' strengths and development potentials. Even though 360-degree feedback instruments are very popular, "probably no more than one in four has been professionally developed and adequately tested for validity and reliability" (Lepsinger & Lucia, 2009, p. 55). However, the validation is an important prerequisite for the acceptance of the assessment by all stakeholders in a company as well as for the appropriateness of inferred statements based on the instrument.

360-degree feedback assessments are a special case of multisource ratings and can be conceptualized in the multitrait-multimethod (MTMM) framework. Structural equation models that have been developed to define reliability, discriminant validity of the traits, and convergent validity of the different methods (i.e., the different raters) differentiate between structurally different methods and interchangeable methods (Eid et al., 2008). In the following chapters, I will thoroughly elucidate the complex data structure of 360-degree feedback assessment and demonstrate that peers and subordinates are two different sets of interchangeable methods that are both nested within the respective target manager. They require a multilevel model that is able to include two distinct level-1 populations.

As such a model does not yet exist, it is developed and presented in this thesis. It uses a planned missing data structure to incorporate peers and subordinates as two level-1 populations. Chapter 2 shows how this model is defined and how it is implemented in statistical software. The chapter includes an empirical application to Benchmarks®, a

widespread 360-degree feedback instrument. The validation reveals acceptable to good reliabilities of the item parcels, low discriminant validity of the subscales, and low convergent validity of the managers self-ratings and ratings of peers and subordinates. About one quarter of variance in peer and subordinate ratings is not shared with the manager's ratings but with the other group members. The lion's share of variance in the single peer and subordinate ratings, however, is neither shared with the manager nor with the other peers or subordinates but is idiosyncratic. The convergent validity between the group of peers and the group of subordinates rating one manager is high indicating that peers and subordinates share a common view that diverges from the managers' self-perception.

The Monte Carlo simulation in Chapter 3 demonstrates that sample sizes of 100 target managers and two peers and subordinates are sufficient to achieve unbiased parameter estimation. Precise estimation of standard errors necessitates 400 target managers or at least four peers and subordinates. If this sample size is not reached, mainly standard errors of common method factors are affected by bias. The simulation study uncovers a strong leniency bias of the χ^2-test statistic. Inferential decisions based on this test would result in too many falsely accepted models. This bias can partly be reduced by a correction of the degrees of freedom that is necessary due to the special planned missing data structure. However, the bias remains substantial and it is not recommended to trust the χ^2-test for this type of model.

Discrepancies between self- and others' ratings that are typically encountered in 360-degree feedback assessments are often interpreted as a lack of the manager's self-awareness. However, there are many other possible reasons for diverging ratings such as personality and individual characteristics of the raters, contextual and cultural variables, and cognitive and motivational aspects. Therefore Chapter 4 reviews existing literature on the association between self-other-agreement (SOA) on leadership competencies and self-awareness. As the results of previous studies are inconsistent, used different statistical approaches, and included

different rating perspectives, a systematic analysis on the correlation between SOA and self-awareness is conducted. Three different multilevel structural equation models reveal that the correlations found between the two constructs are almost completely attributable to shared method variance. When this method variance is explicitly modeled by using the multilevel structural equation model that is presented in the course of this thesis, there remains no substantial association between SOA and self-awareness. In light of this result, it is not advisable to interpret disagreement in ratings as an indicator of lacking self-awareness.

In the following general discussion the scientific contribution of the thesis is stressed. Possible adoptions of the model to other contexts, approaches how to deal with variations of the data structure, and limitations of the present thesis are discussed.

Zusammenfassung

Ziel der vorliegenden Dissertation ist die Entwicklung, Evaluierung und Anwendung eines Mehrebenen-Stukturgleichungsmodells zur Validierung von 360-Grad-Feedbackinstrumenten. Das Konzept dieser Instrumente besteht darin, dass eine Führungskraft ihre Führungskompetenzen selbst einschätzt und außerdem von Kolleg*innen, Mitarbeiter*innen und der oder dem Vorgesetzten beurteilt wird, um ein multiperspektivisches Verständnis der Stärken und Entwicklungspotenziale der Führungskraft zu erhalten. Obwohl 360-Grad-Feedbackinstrumente weit verbreitet sind, wurde vermutlich nicht mehr als eines von vieren professionell entwickelt und adäquat auf Validität und Reliabilität getestet (Lepsinger & Lucia, 2009, S. 55). Allerdings ist die Validierung eine wichtige Voraussetzung sowohl für die Akzeptanz des Verfahrens durch alle im Unternehmen Beteiligten als auch für die Zulässigkeit der Schlussfolgerungen, die auf diesem Instrument basierend formuliert werden.

360-Grad-Feedbackverfahren sind ein Spezialfall von Multisource-Ratings und können als Multitrait-Multimethod (MTMM)-Daten konzeptualisiert werden. Strukturgleichungsmodelle, die für die Bestimmung der Reliabilität, der diskriminanten Validität der Konstrukte und der konvergenten Validität der verschiedenen Methoden (d.h. der verschiedenen Rater) entwickelt wurden, unterscheiden zwischen strukturell verschiedenen und austauschbaren Methoden (Eid et al., 2008). In den folgenden Kapiteln werde ich die komplexe Datenstruktur von 360-Grad-Feedbackverfahren ausführlich beleuchten und zeigen, dass Kolleg*innen und Mitarbeiter*innen zwei verschiedene Sets austauschbarer Methoden sind, die beide in der jeweiligen Führungskraft geschachtelt sind. Sie erfordern daher ein Mehrebenenmodell, das in der Lage ist, zwei distinkte Level-1-Populationen abzubilden.

Da solch ein Modell bisher nicht existiert, wird es in der vorliegenden Dissertation entwickelt und vorgestellt. Es verwendet eine Planned-Missing-Datenstruktur, um Kolleg*innen und Mitarbeiter*innen als zwei Level-1-Populationen zu berücksichtigen. Kapitel 2 zeigt, wie dieses Modell definiert und in statistischer Software implementiert wird. Das Kapitel beinhaltet eine empirische Anwendung anhand von Benchmarks®, einem weitverbreiteten 360-Grad-Feedbackinstrument. Die Validierung zeigt akzeptable bis gute Reliabilitäten der Itemparcel, geringe diskriminante Validität der Subskalen und geringe konvergente Validität zwischen den Selbstbeurteilungen der Führungskräfte und den Beurteilungen durch Kolleg*innen und Mitarbeiter*innen. Ungefähr ein Viertel der Varianz in den Beurteilungen durch Kolleg*innen und Mitarbeiter*innen wird zwar nicht mit der Führungskraft, aber mit den anderen Gruppenmitgliedern geteilt. Der Großteil der Varianz der Beurteilungen der einzelnen Kolleg*innen oder Mitarbeiter*innen wird allerdings weder mit der Führungskraft noch mit den anderen Kolleg*innen oder Mitarbeiter*innen geteilt, sondern ist idiosynkratisch. Es besteht hohe konvergente Validität zwischen der Gruppe der Kolleg*innen und der Gruppe der Mitarbeiter*innen in der Beurteilung einer Führungskraft. Dies zeigt, dass Kolleg*innen und Mitarbeiter*innen eine gemeinsame Sicht teilen, die von der Selbstwahrnehmung der Führungskraft abweicht.

Die Monte-Carlo-Simulationsstudie in Kapitel 3 demonstriert, dass Stichprobengrößen von 100 Führungskräften und zwei Kolleg*innen und Mitarbeiter*innen ausreichend sind, um unverzerrte Parameterschätzungen zu erhalten. Die präzise Schätzung von Standardfehlern benötigt 400 Führungskräfte oder wenigstens vier Kolleg*innen und Mitarbeiter*innen. Wenn diese Stichprobengröße nicht erreicht wird, so sind hauptsächlich die Standardfehler der Common-Method-Faktoren von Verzerrung betroffen. Die Simulationsstudie deckt eine starke Leniency-Verzerrung der χ^2-Teststatistik auf. Inferenzstatistische Schlüsse, die auf diesem Test basieren, würden zu viele falsch akzeptierte

Modelle zur Folge haben. Dieser Bias lässt sich teilweise durch eine Korrektur der Freiheitsgrade reduzieren, die aufgrund der speziellen Planned-Missing-Datenstruktur notwendig ist. Trotzdem bleibt eine substanzielle Verzerrung bestehen und es wird empfohlen, dem χ^2-Test für dieses Modell nicht zu vertrauen.

Diskrepanzen zwischen Selbst- und Fremdeinschätzungen, die typischerweise in 360-Grad-Feedbackverfahren beobachtet werden, werden oft als mangelnde Selbstaufmerksamkeit der Führungskraft interpretiert. Allerdings gibt es viele weitere mögliche Ursachen für divergierende Ratings, wie z. B. Persönlichkeit und individuelle Eigenschaften der Rater, kontextuelle und kulturelle Einflussgrößen und kognitive und motivationale Aspekte. Daher wird in Kapitel 4 bestehende Literatur zum Zusammenhang zwischen der Übereinstimmung von Selbst- und Fremdratings von Führungskompetenzen und Selbstaufmerksamkeit gesichtet. Da die Ergebnisse bisheriger Studien inkonsistent sind, unterschiedliche statistische Ansätze verwendet wurden und unterschiedliche Ratingperspektiven Berücksichtigung fanden, wird eine systematische Analyse der Korrelation zwischen der Übereinstimmung von Selbst- und Fremdratings und der Selbstaufmerksamkeit durchgeführt. Mithilfe dreier verschiedener Multilevel-Strukturgleichungsmodelle wird gezeigt, dass die zwischen den Konstrukten gefundenen Korrelationen beinahe vollständig auf gemeinsame Methodenvarianz zurückzuführen sind. Wenn diese Methodenvarianz im Multilevel-Strukturgleichungsmodell, das in dieser Dissertation vorgestellt wird, expliziert modelliert wird, so bleibt kein substanzieller Zusammenhang zwischen der Übereinstimmung von Selbst- und Fremdratings und der Selbstaufmerksamkeit erhalten. Angesichts dieses Ergebnisses wird davon abgeraten, Diskrepanzen zwischen Selbst- und Fremdeinschätzungen als Indikator für mangelnde Selbstaufmerksamkeit zu verwenden.

In der folgenden Diskussion wird der wissenschaftliche Beitrag der vorliegenden Dissertation verdeutlicht. Mögliche Anwendungen des Modells auf andere Kontexte, Ansätze, wie mit Variationen der Datenstruktur umgegangen werden kann, und Limitationen der Dissertation werden diskutiert.

List of Figures

Chapter 1

Figure 1. Levels of measurement and sampling process of 360-degree feedback data 8

Figure 2. Schematic illustration of a multilevel structural equation model needed for the validation of 360-degree feedback data 23

Figure 3. Studies of this thesis 26

Chapter 2

Figure 1. Sampling procedure for two sets of interchangeable methods (peers and subordinates) and one structurally different method (self-report) 76

Figure 2. Multilevel confirmatory factor analysis multitrait-multimethod model for two sets of interchangeable methods and one structurally different method with indicator-specific trait variables 77

Figure 3. Multilevel confirmatory factor analysis multitrait-multimethod model for two sets of interchangeable methods and one structurally different method with general trait variables 78

Chapter 3

Figure 1. Multilevel confirmatory factor analysis multitrait-multimethod model for two sets of level-1 methods with general trait variables 128

Figure 2. Level-1 variance-covariance matrix for the model with two traits 129

Figure 3. Density curve of 500 observed χ^2-values and expected theoretical curves 130

Figure 4. Boxplots of the parameter estimation bias (PEB) 131

Figure 5. Boxplots of the standard error bias (SEB) 132

Chapter 5

Figure 1. Examples of possible multisource feedback applications 202

List of Tables

Chapter 1

Table 1. Structurally different and interchangeable methods in 360-degree feedback data 17

Chapter 2

Table 1. Definition of the variance components for the indicators of the nonreference methods 71

Table 2. Overview of latent factors, variance components, and factor correlations 72

Table 3. Data format for the application of the model for two sets of interchangeable methods and one structurally different method 73

Table 4. Means, factor loadings, and coefficients of consistency, common method specificity, unique method specificity, and reliability 74

Table 5. Factor variances and factor correlations 75

Chapter 3

Table 1. Definition of the variance components for the indicators of the nonreference methods 124

Table 2. Data format for the application of the model 125

Table 3. Simulation design 126

Table 4. Factor correlations 127

Chapter 5

Table 1. Statistical approaches to typical 360-degree feedback research questions 196–197

Chapter 1

Introduction

1. Multisource Feedback Ratings

Multisource feedback ratings have gained wide popularity in the field of leadership development and evaluation. Their rise in the 1990s culminated in an estimate of 90% of Fortune 1000 companies applying some form of multisource assessment (Edwards & Ewen, 1996). Multisource feedback processes are not only implemented in industrial settings but in many other contexts. They all have in common that the attributes of interest of the target persons are evaluated from multiple perspectives. For example, teachers' performance can be assessed by the teachers' self-ratings and by students' and colleagues' ratings (see Berk, 2009). Physicians can deliver self-ratings on their competencies and be evaluated by colleagues and patients (for a review see Donnon, Al Ansari, Al Alawi, & Violato, 2014). Psychological attributes can be assessed with a multi-perspective approach, e.g., by using self, teacher, and parent versions of questionnaires of children's or adolescents' strengths and difficulties (Stone, Otten, Engels, Vermulst, & Janssens, 2010). Multisource feedback also has a long history in military (e.g., Wherry & Fryer, 1949; Williams & Leavitt, 1947) and is still applied when it comes to assessing leadership in the army (for a discussion see Lee, 2015).

The most prominent application of multisource feedback, however, is the rating of a manager's leadership skills, styles, or knowledge (Lepsinger & Lucia, 2009, p.11). The main purpose of these assessments typically is either to support managers' professional development by detecting individual strengths and weaknesses or to evaluate managers in order to make personnel decisions (Antonioni, 1996). The process is often referred to as *360-degree feedback* because it considers the full circle of possible informants: The manager evaluates him- or herself and is evaluated by his or her subordinates, peers, and supervisors.

Additional evaluations can (but do not have to) be collected from clients and/or customers (Church & Bracken, 1997).

Including so many participants in the assessment process costs a lot of time and money. So the question arises why companies are willing to invest these resources. The answer is twofold (Morgeson, Mumford, & Campion, 2005): On the one hand, this procedure corresponds to most companies' philosophy of commitment and involvement of employees. By including all stakeholders in the feedback process, companies hope to increase the acceptance of the assessment. On the other hand, it is well known, that the reliability and validity of an assessment can profit by adding informants (Lawler, 1967). Feedback from the boss, subordinates, and peers is highly valued in addition to the manager's self-ratings as it is assumed to offer information that otherwise would not be available (London & Smither, 1995). While "downward feedback", that is, feedback from a supervisor to a subordinate, has a long history in traditional employee performance evaluation (Atwater, Roush, & Fischthal, 1995), "upward feedback" from subordinates to their supervisors originates in the ambition to rely not only on one source of information when evaluating a supervisor (that is, the supervisor's self-report) but to provide a broader picture of his or her performance (Lepsinger & Lucia, 2009, pp. 7–8). As subordinates are most directly affected by the target manager's behavior, their ratings can be a valuable feedback for him or her (Atwater et al., 1995). Finally, teamwork becomes an increasingly important success factor for companies, so that peer ratings are considered indispensable to gain an insight into a manager's behavior (Lepsinger & Lucia, 2009, p. 10).

1.1 Self-Other-Agreement in 360-Degree Feedback Ratings

Once the assessment is completed, managers usually receive detailed, yet anonymized, feedback on their rating results. One main focus is the identification of gaps between self-ratings and others' ratings as they are assumed to hint at "blind spots" or lack of self-awareness and therefore at a need of personnel development (Eckert, Ekelund, Gentry, & Dawson, 2010). But is it really reasonable to call the manager to account for such gaps? In research there is a long-lasting and still ongoing debate about the meaning of discrepancies between ratings of different sources. Authors discovered already in the 1950s that supervisors and subordinates do not agree in their ratings on a manager's behavior and stated that "there may be real differences in what is measured or perceived from above and below" (Besco & Lawshe, 1959, p. 579). Campbell, Dunnette, Lawler, and Weick (1970) and Borman (1974) argued that different raters will not necessarily agree in their ratings and that this is not a sign of lacking reliability but expected due to different rating perspectives. Other authors, however, followed the traditional idea of reliability and used interrater agreement as a measure of rating accuracy (Yammarino & Atwater, 1993, 1997).

Summing up, the majority of studies on rater discrepancies in 360-degree feedback can be separated in two camps: One large set of studies used the agreement of self- and others' ratings as an indicator of self-awareness and thus expected, in the ideal case, the ratings to agree (e.g., Atwater & Yammarino, 1992; Berson & Sosik, 2007; Bratton, Dodd, & Brown, 2011; Church, 1997; Fletcher & Baldry, 2000; Van Velsor, Taylor, & Leslie, 1993; Wohlers & London, 1989). The other set of studies argued that rating discrepancies are expected because raters differ in their perceptions and that only these discrepancies justify the effort of a 360-degree feedback assessment. In these studies, it was often tried to detect factors that explain the degree of rating discrepancies (e.g., Brutus, Fleenor, & McCauley, 1999; Eckert et al., 2010; Fleenor, Smither, Atwater, Braddy, & Sturm, 2010; Gentry, Hannum, Ekelund, & de Jong, 2007; Ostroff, Atwater, & Feinberg, 2004; Vecchio & Anderson, 2009).

In order to analyze whether discrepancies between self- and others' ratings in 360-degree feedback assessments indicate a lack of the manager's self-awareness, studies have been conducted that correlate the two constructs of self-other-agreement (SOA) and self-awareness (e.g., Church, 1997; Mersman & Donaldson, 2000; Van Velsor et al., 1993; Wohlers & London, 1989). Self-awareness comprises the processes of "anticipating how others perceive you, evaluating yourself and your actions according to collective beliefs and values, and caring about how others evaluate you" (Baumeister, 2005, p. 7). Accordingly, it is assumed that the higher the manager's self-awareness, the lower the discrepancies between self- and others' ratings, regardless of the direction of discrepancies. However, the results of these studies were mostly inconsistent with this hypothesis and also inconsistent with each other. The relationship between the two constructs therefore remains unclear. Yet, shedding light on this topic is not only of research interest but also highly relevant for those who participate in a 360-degree feedback. Feedback givers as well as receivers need to know if ratings discrepancies should be interpreted as a lack of the manager's self-awareness.

1.2 Validation of 360-Degree Feedback Instruments

Ensuring reliability and validity of ratings in a 360-degree feedback assessment is important for research purposes as well as for a company that wants to apply the instrument successfully. Even though there are many empirical studies that work with data gained in 360-degree feedback assessments, this prerequisite – using a well-validated instrument –is often ignored. Consequently, analyses based on this instrument and inferred statements, e.g. on the relationship of leadership competencies to other variables, may be inappropriate or meaningless if the instrument is not measuring what it claims to measure or if the measurement is not reliable (Zumbo, 2007).

The use of such an instrument can also cause severe application problems in a company. Participants might refuse to trust in the results or in the recommended change in their managerial behavior if it has not been shown that the behavior measured by the instrument is indeed related to job success (Lepsinger & Lucia, 2009, p. 55). For the communication of the feedback results it would also be helpful to know the degree to which the different rating sources typically agree or disagree for the given dimensions of leadership. It makes a difference for the interpretation of a gap between self- and others' ratings if the manager is told that convergence of raters is typically low or high on that dimension. However, according to Lepsinger and Lucia (2009), of the many instruments available, "probably no more than one in four has been professionally developed and adequately tested for validity and reliability" (p. 55).

One possible reason for the small number of thorough validations of 360-degree feedback instruments is the challenging data structure and associated statistical issues. The next sections discuss this data structure and statistical approaches used in this context. Ratings of clients and customers will not be included here as they are not of focal interest in most applications of 360-degree feedback. However, they could of course be added as a further rating source in the statistical approach that will be presented in the following parts of this thesis.

1.3 Data Structure of 360-Degree Feedback Assessments

Ratings from 360-degree feedback assessments can be conceptualized as a special case of multitrait-multimethod (MTMM) data. Each instrument covers multiple dimensions – that is the traits –, and each of these dimensions is assessed by multiple raters – that is the methods. Some authors therefore also use the term "multitrait-multirater" (MTMR) data in this context (e.g., Conway, 1996; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Woehr, Sheehan, & Bennett, 2005). The groundbreaking works in this field go back to Campbell and Fiske (1959) who introduced the concept of measuring convergent and discriminant validity of MTMM data and to Lawler (1967) who was the first to apply this idea to multirater data of managerial job performance.

Another characteristic of data from a 360-degree feedback assessment is that it has a multilevel structure. In order to understand this two-level structure one has to take a closer look at the underlying sampling process. When a 360-degree assessment is conducted, the first step of the sampling process is to choose which target managers will be evaluated. Once these managers are selected, they deliver a self-rating and their direct boss delivers a rating. The second step is to choose some of the managers' peers and some of their subordinates for an evaluation of the target manager. (Sometimes, clients or customers of the target managers are additionally chosen in this step.) The data therefore has two levels (see Figure 1): The first sampling step constitutes the target managers' self-ratings on level 2. The second sampling step constitutes peer and subordinate ratings (nested within the respective target manager) on level 1. As there is only one direct boss, no additional sampling is necessary to define the boss who delivers the ratings, but the boss is set as soon as the target manager is chosen. Therefore, just as the self-ratings, the boss' ratings are located on level 2. Even though this two-level structure is present in most 360-degree feedback data sets, it has, to my knowledge, not been considered in previous analyses. In the rare cases of multilevel analyses in this field, raters were nested within cultures (e.g., Atwater, Wang, Smither, & Fleenor, 2009; Eckert et al., 2010) in order to analyze cultural influences on rating discrepancies.

In the following section, the most prominent approaches that have so far been used in the analysis of 360-degree feedback data are discussed.

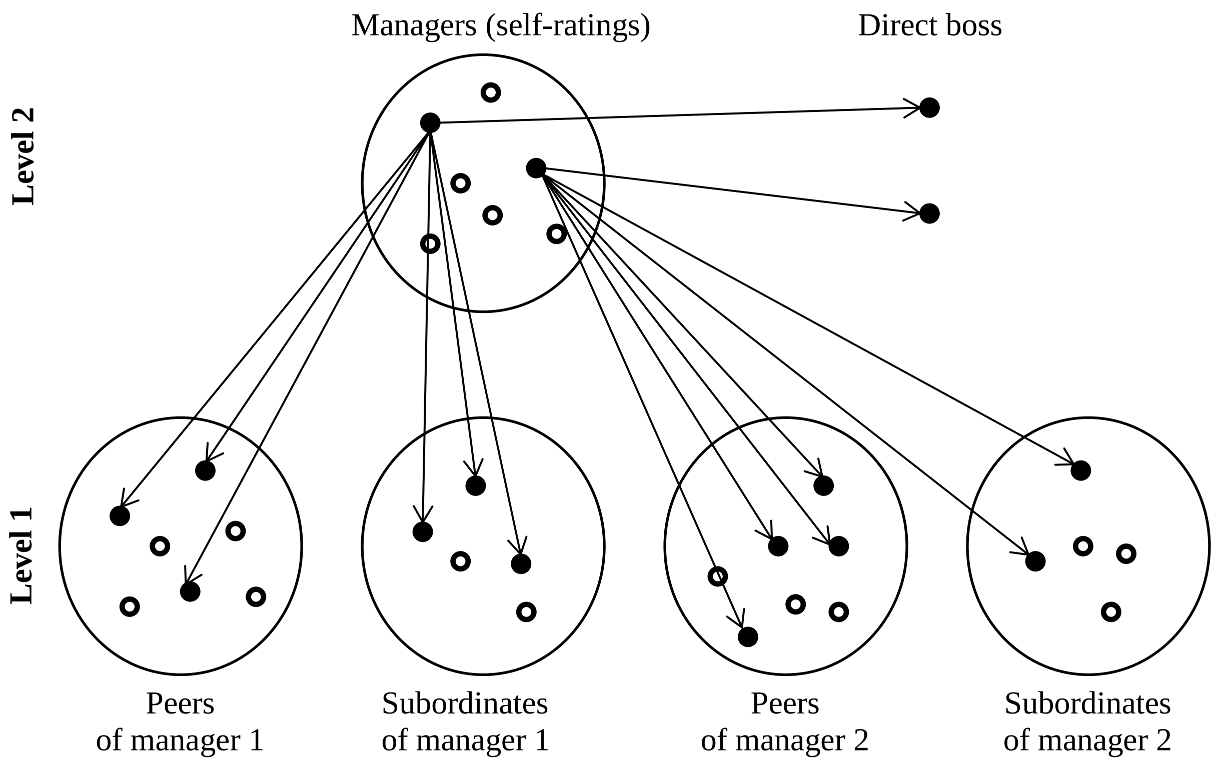

Figure 1. Levels of measurement and sampling process of 360-degree feedback data. The figure shows the sampling process exemplarily for two managers. Large circles refer to populations. Filled small circles refer to chosen subjects (as indicated by the arrows), empty small circles refer to non-chosen subjects.

2. Manifest Approaches in the Analysis of 360-Degree Feedback

2.1 Correlations in MTMM Analysis

For validation purposes, data from 360-degree feedback assessments are usually considered in a MTMM framework to define the convergent and discriminant validity of the instrument or its subscales. In the first decades of MTMM analysis, manifest correlations between different methods (i.e., different raters) rating the same trait were used to assess the degree of convergent validity, whereas manifest correlations between the different traits rated by the same method (or rater) indicated the degree of discriminant validity (e.g., Lawler, 1967; Viswesvaran, Ones, & Schmidt, 1996; for meta-analyses see Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988). These analyses yielded important insides as they highlighted that each rating contains both – information on the trait that is rated and information due to the specific rating source.

2.2 Within and Between Analysis (WABA)

Within-and-between-analysis (WABA; Yammarino & Markham, 1992; Yammarino, 1998) uses analysis of variance (ANOVA) and analysis of covariance (ANCOVA) procedures to define variation and covariation of variables within and between groups. It is thereby possible to assess the agreement within rating sources, e.g., within the group of peers of a target manager, and between rating sources, e.g., between self-ratings and peer ratings or between peer ratings and subordinate ratings. Yammarino (2003) recommended applying this technique before feedback is given to the participants of a 360-degree feedback assessment. If, for example, it is found that peers do not agree in their evaluation of a manager, Yammarino (2003) stresses that it is not advisable to build an average or aggregated peer score as it is often done. Instead, individual ratings of the peers should be presented in the feedback to the target manager.

2.3 Difference Scores of Self- and Others' Ratings

When the agreement between self- and other's ratings is of focal interest in the analysis of 360-degree feedback data, the computation of difference scores has been a widespread approach (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Edwards, 2002). The algebraic difference between self- and others' ratings, that is, the difference including the plus or minus sign, as well as the absolute or squared difference have been used depending on the specific hypothesis (Edwards, 1994). When the direction of disagreement is important, i.e., when it is of interest whether the manager is over- or underestimated, the algebraic difference should be applied. Contrariwise, absolute and squared differences are appropriate when one attempts to depict the degree of disagreement but the direction of this disagreement is not relevant.

The popularity of difference scores decreased as they have been criticized for several reasons (Edwards, 1993, 1994, 2002). On the one hand, it is known that difference scores have a lower reliability than the original components (Cronbach & Furby, 1970). On the other hand, when a third variable is included in the analysis, e.g., an explaining variable or an outcome variable, the contribution of the single components (self-ratings and others' ratings) is confounded. It is therefore not possible to determine the relationship of the single components of self-ratings and others' ratings to the third variable and the three-dimensional relationship is reduced to a two-dimensional one.

2.4 Categories of Agreement

Based on the difference scores, categories of agreement have been built (e.g., Atwater & Yammarino, 1992; Fleenor, McCauley, & Brutus, 1996; Mersman & Donaldson, 2000). These categories typically include overraters, underraters, and in-agreement raters. Managers with difference scores that do not fall within the range of one-half standard deviation around the mean difference are categorized as over- or underraters, all managers within this range are in-agreement raters. Sometimes more than three categories are used, e.g., when in-agreement raters are split into in-agreement/poor and in-agreement/good (Atwater & Yammarino, 1997). Differences between these groups on further variables of interest, e.g., personality and cognitive ability (Fletcher & Baldry, 2000), can then be analyzed. The advantage of this categorization is its intuitiveness. Results based on these results are straightforward to communicate. However, a substantive portion of variance is lost when the values on a continuous variable are collapsed into categories and results can be misleading (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002). Additionally, as the categories are based on the calculation of difference scores, all shortcomings associated with difference scores also apply here.

2.5 Regressions With Self- and Others' Ratings as Outcomes or Predictors

Edwards (1994, 1995, 2002) proposed different regression analytic approaches that circumvent the need for difference score calculation and keep the components of self- and others' ratings separate. These techniques can be used when correlates of ratings in 360-degree feedback assessments are analyzed. In the case where self- and others' ratings are outcome variables and are to be explained by a third variable, multivariate regression can be applied (Edwards, 1995). For example, Gentry et al. (2007) analyzed the effect of managerial level on the discrepancies between self- and others' ratings when these ratings were jointly analyzed as outcome variables in a multivariate regression. In the case of using self- and others' ratings as explaining variables of a third (outcome) variable, Edwards (1994, 2002) presented a polynomial regression strategy. It was for example applied to simultaneously determine the effect of self- and others' ratings on effectiveness (Atwater et al., 1998).

2.6 Aggregating Ratings of Peers or Subordinates

In the analysis of 360-degree feedback data, researchers so far circumvented the need for a multilevel design by using different strategies. Often, aggregated ratings were used (e.g., Atwater et al., 2009; Church, 1997; Flechter & Baldry, 2000). All peer ratings of a target manager were aggregated by calculating a mean peer rating and all subordinate ratings were aggregated and these mean peer and subordinate ratings were used in the subsequent analyses. While this strategy can be helpful to increase the reliability of the ratings, it is associated with some methodological shortcomings such as reduction of sample size, larger standard errors, loss of power, and loss of information (Koch, Schultze, Eid, & Geiser, 2014). Additionally, Hensel, Meijers, van der Leeden, and Kessels (2010) have exemplarily shown that six to ten peer raters were necessary to reach an acceptable reliability (.7) of the ratings. This result is in agreement with the meta-analysis of Conway and Huffcutt (1997) who report that ten subordinates are needed to reach a reliability of .8. However, this number of peers and subordinates is rarely present in 360-degree feedback assessments. According to London and Smither (1995) the typical number of peers and subordinates varies between four and six. Hensel et al. (2010) have demonstrated that for these typical data situations, not only the reliability of the aggregated ratings is low but that also the correlation between these aggregated ratings and a third variable (i.e., a supervisor rating) is underestimated. These results indicate that, in many cases, aggregating the single ratings of peers and subordinates is not sufficient to ensure reliability. Instead, the single peer and subordinate ratings should be taken into account instead of being collapsed in order to maintain this source of variation.

Other strategies that have been applied to reduce the data structure to a single level of measurement are the random selection of one supervisor, one peer, and one subordinate out of a larger set of available raters (e.g., Hoffman, Lance, Bynum, & Gentry, 2010) or the restriction of the number of supervisors, peers, and subordinates to an arbitrary number (e.g., two supervisors, two peers, and two subordinates in Mount et al., 1998) even if some managers might have more than two raters from one source. Just like the aggregation approach, both strategies disregard a lot of available information. When choosing only one rating per source it is furthermore not possible to determine the degree of convergence within the group of supervisors, within the group of peers, or within the group of subordinates. Putka, Lance, Le, and McCloy (2011) warned against these strategies and demonstrated for an MTMR design, that results of confirmatory factor analysis (CFA) models differ depending on which raters from a larger set of raters were selected. One of their suggestions is to adopt multilevel CFA models. CFA models belong to the class of latent modeling approaches and will be discussed in the next section.

3. Latent Approaches

Manifest approaches that do not take measurement error into account can cause severe bias in parameter estimates (Rigdon, 1994). Latent approaches overcome this issue as they allow estimating latent variables that are free of measurement error while this error is explicitly modeled. Item response theory (IRT) and structural equation modeling have been used to analyze 360-degree feedback data.

3.1 Item Response Theory (IRT)

Different IRT approaches have been applied to assess measurement equivalence across rating sources. For example, Barr and Raju (2003) argue that measurement equivalence is a prerequisite for calculating and interpreting self-other rating discrepancies. If it is not assured that the measures have a common psychological metric across sources, one cannot unambiguously attribute differences in ratings to different perceptions of the target manager’s behavior. The authors use and compare three different IRT-based approaches to assess measurement equivalence of a 360-degree feedback instrument on the item level as well on the scale level. They conclude that – although the results differ depending on the specific IRT model used – overall, practical effects of rater severity or leniency due to rating source were negligible. Comparable to the manifest approaches discussed above, Barr and Raju (2003) reduced the number of available raters by using only one boss, one peer, and one subordinate report per target manager.

In another study on measurement equivalence of the different rating sources, a multiple-group CFA model and an IRT model were used (Facteau & Craig, 2001). Both techniques revealed no substantial differences between the rater groups so that the instrument is regarded as invariant across the ratings of self, the boss, peers, and subordinates. IRT is a very helpful tool if the focus of the study is the analysis of measurement equivalence. If, however, the ratings from the 360-degree feedback assessment are not only to be tested for measurement equivalence but are also to be analyzed in a framework that includes their relationship to other variables, structural equation modeling offers a very flexible approach.

3.2 Structural Equation Modeling

In a structural equation model of 360-degree feedback data, the variance in ratings can be split into the components of the latent trait, of the method, and of the error. Additionally, model assumptions can be tested empirically and all latent factors can be linked to further variables (Eid et al., 2008). Within the structural equation framework, confirmatory factor analysis (CFA) is used to define the measurement model. In traditional CFA models, there was only one indicator per trait-method-unit (TMU; e.g., Jöreskog, 1971; Kenny, 1976). This strategy, however, restrains method effects to be unidimensional across different traits. Applications of less restrictive MTMM models have shown though that method effects differ between traits (e.g., Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). One trait might be overestimated by a specific method and another trait might be underestimated by this method. In this regard, models with multiple indicators bear a large advantage over models with only one indicator per TMU and will be used in this thesis.

The first CFA models that were proposed used a correlated trait-correlated method (CT-CM) approach (Jöreskog, 1971). In this model, every indicator has a loading on one trait and on one method factor. All trait factors are allowed to correlate and all method factors are allowed to correlate whereas there are no correlations between trait and method factors. This model is appealing as it separates the variance of each indicator into a trait, a method, and an error component and allows estimating convergent validity (indicated by the trait-factor loadings), discriminant validity (indicated by the correlations between traits), and method specificity (indicated by the method-factor loadings). However, the model has encountered some serious problems (Kenny & Kashy, 1992; Marsh, 1989). First, it has been shown that the model tends to have convergence problems or yields improper solutions, i.e., parameter estimates out of the permissible range of values (for a simulation study see Castro-Schilo, Widaman, & Grimm, 2013). Second, it may produce estimates difficult to interpret (Marsh, 1989; Marsh & Grayson, 1995) and the interpretation of the trait and method factors is challenging as it is not clear what the difference between these two types of factors actually is (Eid, 2000). Third, the model restricts trait and method factors to be uncorrelated, an assumption that has been questioned or even relaxed in some applications (Schmitt & Stults, 1986).

Therefore, models that overcome these issues have been developed (e.g., Eid, 2000; Eid et al., 2003, Eid et al., 2008). They employ a correlated trait-correlated method minus one (CT-C[*M*-1]) approach instead of a CT-CM structure. The CT-C(*M*-1) modeling approach to CFA-MTMM data has been introduced by Eid (2000). Just as in the CT-CM model, the observed variance in ratings can be decomposed into trait, method, and error components. However, the model overcomes several of the issues encountered with the CT-CM model. Trait and method factors are well-defined and therefore have a clear meaning so that interpretation difficulties are avoided. Additionally, the model is identified also in situations where the CT-CM model is not. Eid (2000) demonstrated that the model can be fitted to data from a study on managerial performance measured by self-ratings, subordinate, and peer ratings (Mount, 1984), even when other CFA-MTMM models (such as the CT-CM model) encountered difficulties (Kenny & Kashy, 1992).

In CT-C(*M*-1) models there is one fewer method factor than there are methods. One of the methods has to be chosen as the *reference method*. In applications where a “gold standard” exists, this method should be selected as the reference. Otherwise, the reference method should be chosen in a way that eases the interpretation of results (Geiser, Eid, &

Nussbeck, 2008). In 360-degree feedback assessments it usually is advisable to choose the self-report as the reference method. This reference method does not have a method factor. Consequently, the trait factors in the model represent the trait measured by the reference method. All other methods, so-called *nonreference methods*, are contrasted against the reference method by regressing them on the trait factor. The residuals of this latent regression are captured by method factors. Thus, for every trait in the model there are as many method factors as there are nonreference methods. They capture deviations of the specific method from what is expected due to the reference method. In these models, method effects are not restricted to be unidimensional but are trait-specific. The model can be defined with indicator-specific trait variables if the indicators measure slightly different facets of one construct or with common trait factors if the indicators of one construct are homogeneous (Eid et al., 2008). As the method factors are residuals from the regression on the trait they are not allowed to correlate with the trait.

3.2.1 Interchangeable vs. Structurally Different Methods

If researchers consider both – the use of structural equation modeling as well as a potential multilevel data structure – the range of possible models is tremendous and many researchers struggle to find “the right one” for their MTMM data. In the end, their decision often is data-driven and arbitrary (Eid et al., 2008).

Therefore, Eid et al. (2008) argue that researchers should carefully classify their MTMM data structure at hand to choose an appropriate structural equation model. In order to provide decision guidance, they differentiate between *structurally different* and *interchangeable methods*. In their introductory example, they explain that multiple students who rate their teacher’s teaching quality are interchangeable methods “because they are drawn from a larger group of students who have more or less the same access to the teacher’s behavior” (Eid et al., 2008, p. 232). Contrariwise, when children’s personality traits are evaluated this could be realized via self-ratings, the teacher's ratings, and parents' ratings. These three methods are structurally different "because they are not randomly chosen from a common set of equivalent raters" (Eid et al., 2008, p. 232) but explicitly selected.

Applying this concept to 360-degree feedback data leads to the following classification (see Table 1): Self-ratings and direct boss ratings are structurally different methods. Once the target managers are chosen, these methods cannot be randomly drawn from a larger group of "selfs" or of direct bosses but it is fixed who delivers the rating. It is the target manager him- or herself and his or her direct boss. However, there are multiple peers and multiple subordinates for a given target manager, and some of them are chosen to evaluate the manager. Therefore, the peers are interchangeable among each other and the subordinates are interchangeable among each other. Contrariwise, the group of peers and the group of subordinates are structurally different from one another and also from the boss ratings and from the target manager's self-ratings.

Table 1

Structurally Different and Interchangeable Methods in 360-Degree Feedback Data

Four structurally different methods	Two sets of interchangeable methods
Self	
Direct boss	
Group of peers	Peer 1 – Peer 2 – Peer 3 – ...
Group of subordinates	Subordinate 1 – Subordinate 2 – Subordinate 3 – ...

In some 360-degree feedback assessments, not only the direct boss but multiple supervisors are included. If these supervisors have the same access to the target manager's behavior, they define a third set of interchangeable methods.

It is important to notice that the interchangeability of methods is a theoretical concept that results from the sampling process. In an application of a 360-degree feedback, two peers for example might not appear to be interchangeable against each other in a colloquial meaning of the word because each one of them has his or her own opinion and one of them might for example be more well-disposed to the manager than the other. Nevertheless, from a statistical point of view, all peers stem from one population and all subordinates stem from one population. The distinction between interchangeable and structurally different methods is comparable to the one between random and fixed factors in the analysis of variance (e.g., Eisenhart, 1947).

The distinction between structurally different and interchangeable methods has important consequences for the selection of an appropriate structural equation model (Eid et al., 2016; Eid et al., 2008). If there are interchangeable methods they imply a multilevel data structure with the ratings of the interchangeable methods located on level 1 and being nested within level-2 clusters. In the example of Eid et al. (2008) students (level-1 units) were nested within their teacher (level 2-cluster). Multiple structurally different methods can be most adequately captured with a CT-C(*M*-1) structure. One of the structurally different methods has to be chosen as the reference method, the remaining structurally different methods are nonreference methods that are regressed on the reference method and have method factors associated that capture the deviations from this latent regression.

If a data set contains structurally different and interchangeable methods, a combination of these two approaches should be adopted. Eid et al. (2008) presented this model, a multilevel CFA-MTMM model, for the combination of one target delivering a self-report and two peer reports. The two peers are interchangeable while the self-report is a structurally different method. The nested structure of the two peers within the targets requires a multilevel structure with peer ratings on level 1. The self-ratings are located on level 2 and serve as the reference method. They predict the target-specific level-2 variables, or random intercepts, of the peer ratings, that is, the “average” ratings of the two peers of every target. The residuals of this latent regression are captured by two method factors for each trait. One method factor is located on level 2. It depicts the variance in peer ratings that is common to the two peers (and therefore captured on level 2) but not shared with the self-report. It is the deviation of the “average” peer rating from its expected value given the self-report. This factor is called *common method factor*. On level 1, there is a method factor that captures the variance in peer ratings that is neither shared with the self-report nor with the other peer that evaluates the same target. This variance is thus unique to the single peers and the factor is called *unique method factor*.

In this model it is possible to determine (1) the convergent validity of self-ratings and peer ratings, (2) the consistency, or common method specificity, of peer ratings, (3) the uniqueness, or unique method specificity, of peer ratings, and (4) the reliability. The first three coefficients can be expressed as proportions of true – that is, error-free – variance. Additionally, the correlations between traits are indicators of discriminant validity. The correlation of two method factors indicates the degree to which method effects generalize across traits or across methods. Trait factors and common method factors belonging to the same TMU are not allowed to correlate because the common method factors are residuals of the latent regression on the trait. The model can be defined with indicator-specific or with common trait factors (Eid et al., 2008).

Even though there are exactly two peers for every target in the example of Eid et al. (2008), the model could also be applied if there were a varying number of peers per target. It is a traditional multilevel model in which each cluster can have a different level-1 sample size. As data from 360-degree feedback assessments usually do not have a fixed number of ratings from others but some target manager’s might have for example two peer ratings and others might have three or four peer ratings, this flexibility makes the model superior to models that neglect the multilevel structure.

In their validation study of the State-Trait Cheerfulness Inventory (STCI-T; Ruch, Köhler, & van Thriel, 1996), Carretero-Dios, Eid, and Ruch (2011) demonstrated how a further structurally different method can be added to the multilevel CFA-MTMM model for interchangeable and structurally different methods. The extension is straightforward because the additional structurally different method is measured on level 2. It is a further nonreference method that is predicted by the reference method and has a corresponding method factor on level 2.

Following the distinction between structurally different and interchangeable methods, 360-degree feedback data contain (at least) two sets of interchangeable methods – one being the set of peers, the other being the set of subordinates. As discussed above, both rating sources – peers and subordinates – are nested within their target managers and located on level 1. In traditional multilevel models, it is assumed that all level-1 units stem from one cluster-specific population, e.g. students are nested within teachers or companies are nested within cultures or countries. However, the situation is more complex here. Peers stem from one population that is nested within the target managers and subordinates stem from a different population that is as well nested within the target managers. In other words: The population of peers and the population of subordinates are both nested within the same target managers but they are distinct from each other. A model that comprises two sets of interchangeable methods on level 1 has not yet been presented.

Therefore it is not yet possible to adequately include peer and subordinate ratings in one multilevel structural equation model. This extension is intricate as it requires a modeling approach that considers two distinct sets of level-1 populations. So far, there are no models that are able to handle this type of data structure. This restriction does not only apply for the multilevel CFA-MTMM models presented above but for multilevel models in general. It is therefore necessary to develop a multilevel structural equation model that can handle multiple distinct sets of interchangeable methods to adequately cover all rating sources of a 360-degree feedback assessment within one comprehensive model.

4. Outlook to the Following Chapters

The aim of this thesis is to develop, evaluate, and apply a multilevel structural equation model for 360-degree feedback data. A full model that considers all ratings of the 360-degree feedback should incorporate the two sets of interchangeable methods (peers and subordinates) and the additional structurally different methods (self-ratings and boss ratings).

The nested structure of peers and subordinates within the target managers requires a multilevel model with peer and subordinate ratings on level 1, while self-ratings, boss ratings, and “average” peer and “average” subordinate ratings should be captured via a CT-C(*M*-1) structure on level 2. In the interpretation of the results of a 360-degree feedback assessment, the target manager’s self-ratings are usually contrasted with the ratings made by others. Therefore, the self-report will be chosen as the reference method in all models of this thesis.

In Chapter 2, the new model will be presented. Its starting point is the multilevel CFA-MTMM model of Eid et al. (2008) for the combination of interchangeable and structurally different methods. This model includes one set of interchangeable methods (the two peers) and one structurally different method (the self-report). In order to incorporate all rating perspectives that are typically considered in 360-degree feedback assessments the model needs to be extended to one additional structurally different method and one additional set of interchangeable methods.

The extension to a further structurally different model that was presented in Carretero-Dios et al. (2011) will be adopted to include the boss ratings: They stem from a structurally different method that is not nested within the targets but measured on level 2 and will serve as a nonreference method in the model.

Figure 2 schematically shows the model that needs to be developed. The depicted model has indicator-specific trait variables. However, it can also be defined with homogeneous trait variables. Different colors indicate its starting point, the multilevel CFA-MTMM model for structurally different and interchangeable methods (Eid et al., 2008), its extension to one additional structurally different method (Carretero-Dios et al., 2011), and the extension to an additional set of interchangeable methods that has not yet been realized.

In Chapter 2, my co-authors and I will thoroughly present how this new model is defined and how it is implemented in statistical software. In this thesis, Mplus will be used as it is widespread and well-accepted among applied researchers (L. K. Muthén & Muthén, 1998–2017). Mplus has many multilevel options. However, there is no preinstalled command or model that uses multiple sets of distinct level-1 units that are nested within the same level-2 clusters. Therefore, Chapter 2 will present a modeling solution that makes Mplus treat peer and subordinate ratings to stem from two different populations. Data from Benchmarks®, a well-known 360-degree feedback instrument (Lombardo, McCauley, McDonald-Mann, & Leslie, 1999) will serve as an illustrative example in Chapter 2 to show how reliability and discriminant and convergent validity can be estimated.

The next step in the development of a new modeling approach is the evaluation of the model. This is necessary to check if the model estimation in the software works as it is expected, to define the conditions under which appropriate parameter and standard error estimates are derived, and to evaluate whether the indices of model fit provided by the

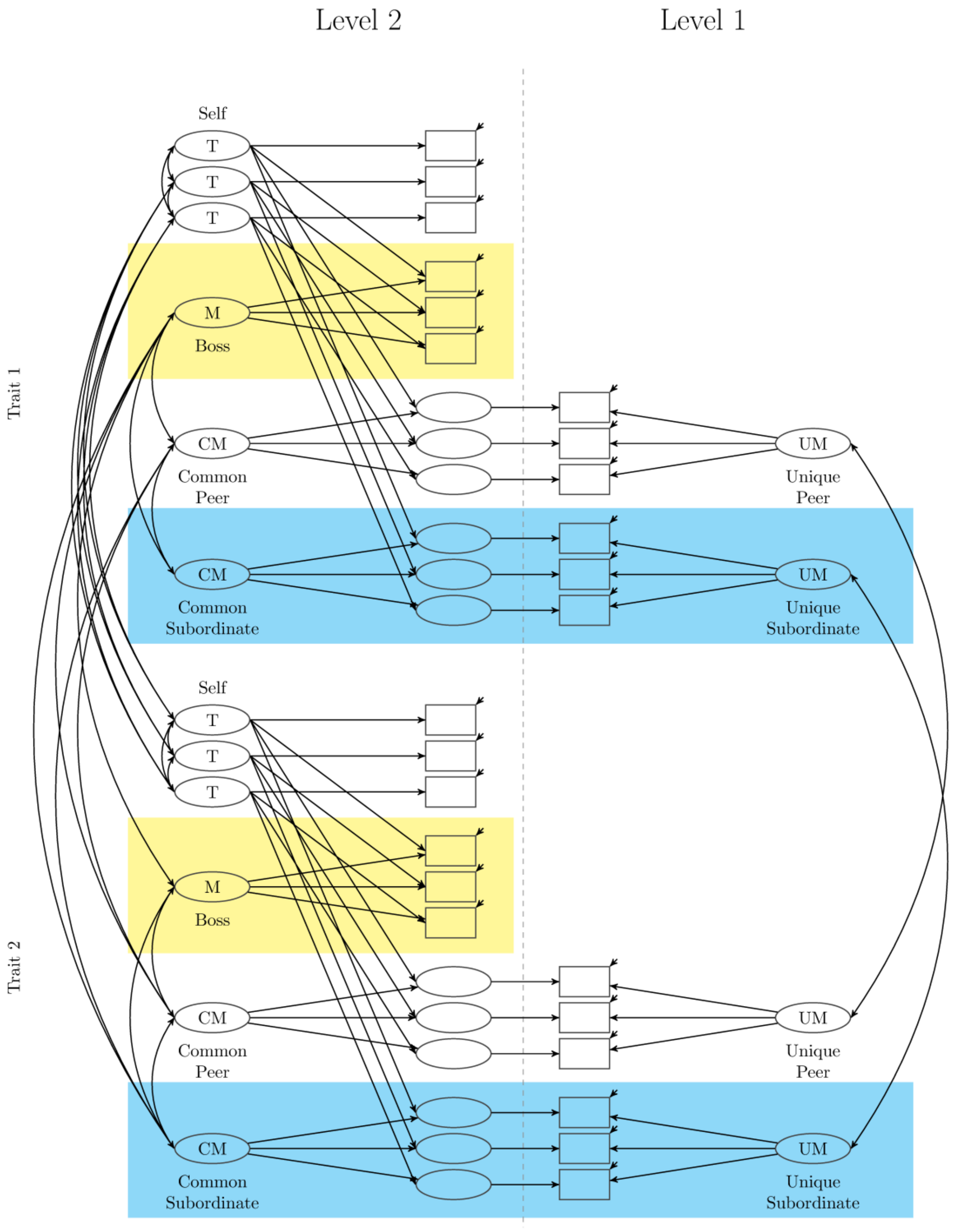

Figure 2. Schematic illustration of a multilevel structural equation model needed for the validation of 360-degree feedback data. T = trait factors; M = method factors; CM = common method factors; UM = unique method factors. Non-highlighted background: direct adoption of the multilevel CFA-MTMM model for structurally different and interchangeable methods with indicator-specific trait variables (Eid et al., 2008). Yellow background: extension of the model to include one further structurally different method (Carretero-Dios et al., 2011). Blue background: second set of interchangeable methods that has not yet been included in the

model. Empty boxes: observed variables (indicators). Empty nodes: latent level-2 variables. Small arrows to the observed variables indicate error variables. Trait variables and common method factors belonging to different traits are allowed to correlate (for simplicity reasons not shown in the figure).

software are trustworthy. Chapter 3 presents an extensive Monte Carlo simulation study of the new model. The simulation design will reflect typical data situations that are encountered in 360-degree feedback assessments: The sample sizes on level 2, i.e. the number of target managers, will vary between 100 and 400; the sample sizes on level 1, i.e. the number of peers and subordinates, will vary between two and eight. A small model with two traits and a larger model with three traits will be defined and the amount of convergent and discriminant validity will vary in order to define the effects of these design factors on model estimation. The chapter will conclude with recommendations for the practical application of the model.

In Chapter 4, the new model will be applied to contribute to the discussion on the meaning of SOA on leadership competencies and self-awareness. Previous studies on this topic have neither used modern structural equation modeling techniques nor have they included all rating perspectives in one comprehensive model. The study in Chapter 4 therefore has two purposes: On the one hand, it is an effort to gain a deeper understanding of the relationship between SOA on leadership competencies and self-awareness. On the other hand, it exemplarily demonstrates how the new model can be used to help answering questions in the context of multisource feedback ratings. Again, Benchmarks® data will be used.

The assumed relationship between SOA and self-awareness typically is that lower disagreement between self- and others' ratings (regardless of the direction of disagreement) is associated with higher self-awareness. This hypothesis will be analyzed with a multilevel structural equation model that uses the absolute values of latent difference variables between self- and others' ratings and correlates them to latent self-awareness (measured by the managers' self-reports as well as by all other rating perspectives). In previous analyses, the hypothesized relationship was rarely found. Instead, there were correlations between self-awareness and the algebraic difference variable of self- and others' ratings, that is, the difference variable used with the plus or minus sign. My co-authors and I will test whether this finding can be replicated with a multilevel structural equation model that uses the latent algebraic difference between self- and others' ratings and latent self-awareness. Again, all rating perspectives will be considered for the measurement of SOA and of self-awareness. As difference variables have been criticized for different reasons, a third multilevel structural equation model that keeps the components of self- and of others' ratings separate will be defined. Self- and others' ratings of leadership competencies will be modeled following the new approach that is presented in this thesis. Correlations between method factors of leadership competencies and latent self-awareness will be analyzed. They indicate whether there remain substantial associations between the method-specific components of leadership ratings, that is, the variance in others' ratings after controlling for the self-ratings, and self-awareness. This modeling strategy allows controlling for the effect of shared method variance on the correlational pattern between the two constructs. The results will be interpreted in light of the question whether low SOA should be used as an indicator of lacking self-awareness.

Figure 3 shows the workflow of this thesis and gives an overview of the following chapters 2–4. Chapter 5 provides a general discussion of the three studies including a short summary on the main results, applicability of the model to other contexts, as well as limitations and suggestions for further research.

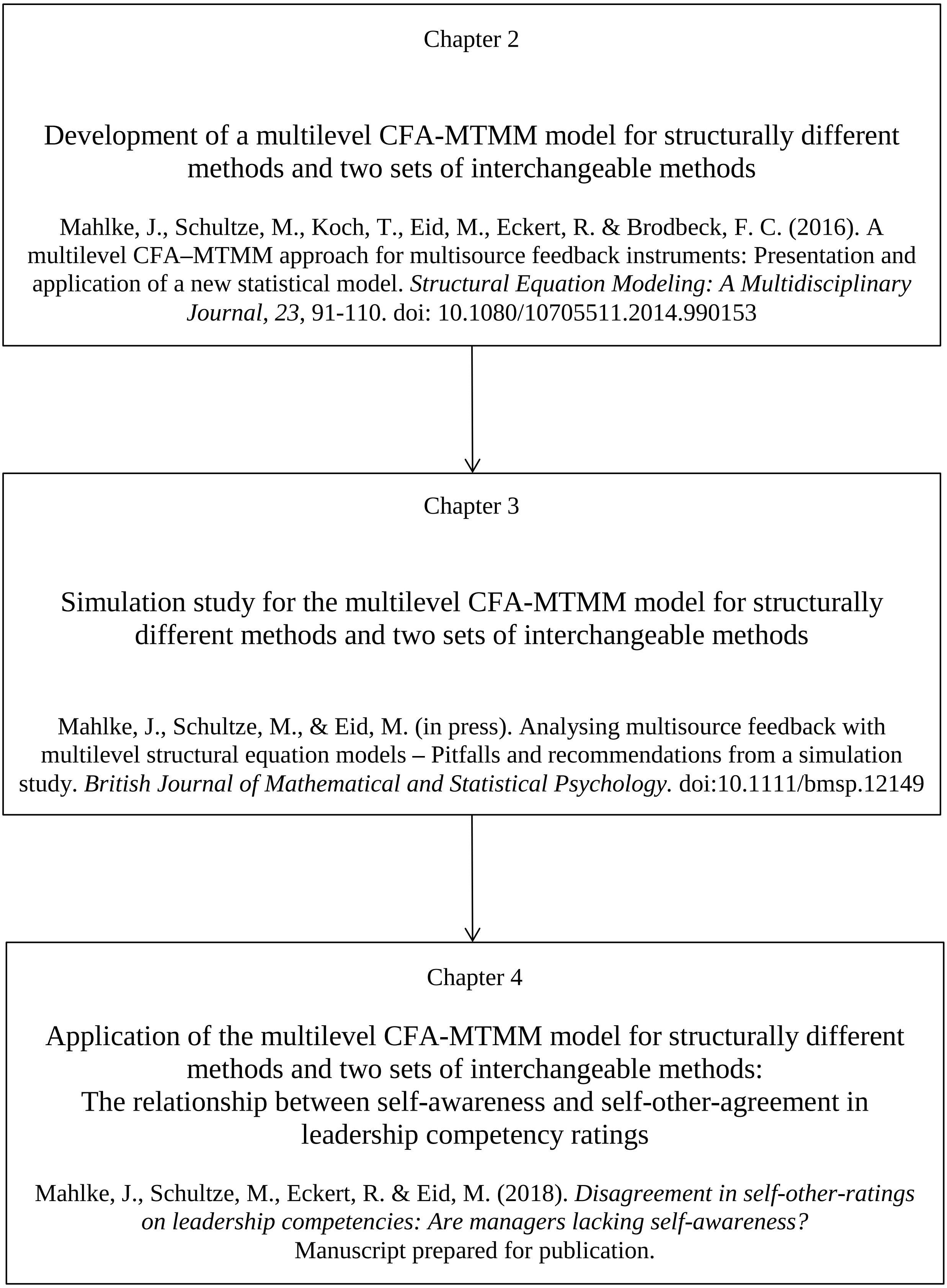
Chapter 2
Development of a multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods
Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R. & Brodbeck, F. C. (2016). A multilevel CFA–MTMM approach for multisource feedback instruments: Presentation and application of a new statistical model. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 91-110. doi: 10.1080/10705511.2014.990153
Chapter 3
Simulation study for the multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods
Mahlke, J., Schultze, M., & Eid, M. (in press). Analysing multisource feedback with multilevel structural equation models – Pitfalls and recommendations from a simulation study. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12149
Chapter 4
Application of the multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods: The relationship between self-awareness and self-other-agreement in leadership competency ratings
Mahlke, J., Schultze, M., Eckert, R. & Eid, M. (2018). *Disagreement in self-other-ratings on leadership competencies: Are managers lacking self-awareness?* Manuscript prepared for publication.

Figure 3. Studies of this thesis. CFA = confirmatory factor analysis, MTMM = multitrait-multimethod.

Chapter 2

A Multilevel CFA–MTMM Approach for Multisource Feedback Instruments: Presentation and Application of a New Statistical Model

This article is published:

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*, 91–110. https://doi.org/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). Some researchers have questioned the existence of rating source effects or have even stated that rater effects are entirely idiosyncratic. They argue that raters from the same organizational level do not agree stronger in their ratings than raters from different levels and that method variance is primarily associated with the individual rater instead of the rater's organizational level (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 of 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, so far there are 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 (e.g., teachers who provide self-ratings and are evaluated by the school principal, by multiple colleagues, and multiple students). 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

- 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.

1. Starting Point: The Model by Eid et al. (2008) for a Combination of Structurally Different and Interchangeable Methods

Eid et al. (2008) presented this model for a combination of several peer reports (level 1 methods) and one teacher self-report (level 2 method). The authors introduce the differentiation between *interchangeable* and *structurally different methods*. Interchangeable methods are found in multilevel structures where level 1 methods, e.g., students, are nested within level 2 methods, e.g., teachers. In a school class, all students have the same access to observe the teacher's behavior. When some of these students are randomly chosen to evaluate their teacher, it is irrelevant which particular students are selected—their ratings are theoretically interchangeable. Thus, the students are interchangeable methods in this example and are located on level 1.

If one additionally considers a self-report of the teacher, the person who delivers the self-report is directly identified. In this case, it is the teacher her- or himself. Thus, the self-report is a structurally different method and is located on level 2, it does not stem from the same population as the students.

The data in the example of Eid et al. (2008) with two interchangeable methods (student reports) and one structurally different method (teacher's self-report) 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. For this reference method, there is no method factor specified. All nonreference methods (student reports) are contrasted with this reference method via trait-specific method factors. We present the model equations for heterogeneous, so-called indicator-specific, trait variables. That is, we do not assume indicators belonging to one trait measure this trait unidimensionally, instead they capture slightly different facets of it. Of course, the model can also be defined more restrictively with a general trait factor in the case of homogeneous, unidimensional indicators (Eid et al., 2008).

In the model of Eid et al. (2008) superscripts distinguish between the different methods: "RM" refers to the reference method (teacher self-report) and "NM" to the nonreference methods (student reports). The subscripts indicate *r* = rater, *t* = target, *i* = indicator and *k* = trait.

The measurement model for the self-report on level 2 is given by

$$Y_{tik}^{\mathrm{RM}} = \mu_{ik}^{\mathrm{RM}} + T_{tik}^{\mathrm{RM}} + E_{tik}^{\mathrm{RM}}. \tag{1}$$

μ_{ik}^{RM} denotes the intercepts of the observed variables Y_{tik}^{RM}, T_{tik}^{RM} are the latent trait factors, and E_{tik}^{RM} depicts the measurement errors on level 2. The nonreference-method indicators can be decomposed in the following way:

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

where μ_{ik}^{NM} denotes the intercepts of the observed variables Y_{rtik}^{NM}, T_{tik}^{NM} denotes the indicator-specific trait variables of the nonreference-method indicators, and $\lambda_{\mathrm{UM}ik}^{\mathrm{NM}}$ represents the loadings on the unique method factors UM_{rtk}^{NM}. Unique means that the factor captures that part of the true student report that is neither shared with the self-report nor with the other student report, but that is unique to the single nonreference method. E_{rtik}^{NM} depicts the measurement errors on level 1. We use the additional subscript *r* for the individual raters.

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}^{NM} are regressed on the trait variables of the reference method T_{tik}^{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}^{NM}, which represents that part of the common view of nonreference methods that is not shared with the reference method. This decomposition of the nonreference-method trait factors is expressed by:

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

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

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

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

For identification purposes, the means of the latent traits T_{tik}^{RM} and T_{tik}^{NM} are fixed to zero and the first factor loading of all factors is set to one. As the method factors CM_{tk}^{NM} and UM_{rtk}^{NM} are residual factors, their means are also zero. The following variables in the model are uncorrelated: (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.

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 student ratings that is shared among the students but is not shared with the self-report, and
- a *unique method factor*, which represents the deviation of a single student rater from the common view of student raters.

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

- The *consistency* of self- and student 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 student ratings indicates the proportion of true variance that is shared among the student but not shared with the self-report.
- The *unique method specificity* of student ratings indicates the proportion of true variance that is not shared with the other student(s) and is thus unique to the specific student 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.

2. The New Model: The ML-CFA-MTMM Model for two Sets of Interchangeable Methods and one Structurally Different Method

For the analysis of 360-degree feedback or in other situations with multiple sets of interchangeable methods the model of Eid et al. (2008) cannot be applied because it is only possible to include the leader’s self-report and *either* the subordinate reports *or* the peer reports but never both of the latter perspectives at the same time. To the best of our knowledge there are so far no models that can appropriately handle such complex data structures. Therefore, the model presented in this paper was specifically designed for data situations that have two sets of interchangeable methods (e.g., a set of subordinates and a set of peers) and one level 2 method (e.g., the self-report; see Figure 1 for an illustration of the sampling procedure and Appendix A for the underlying random experiment). Hence, our extension of the model of Eid et al. (2008) provides a comprehensive framework to quantify

- the amount of convergent validity between self-reports and subordinates and self-reports and peers,
- the degree to which subordinates and peers agree in their perception of a leader—among their respective group and between the two groups—, and
- the degree to which the perception of a leader is unique to the single subordinates and peers.

For simplicity, we describe the model in which there is only one level 2 method (self-report) in addition to the two sets of 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). In their validation of the State-Trait Cheerfulness Inventory (STCI-T; Ruch, Köhler, & van Triel, 1996) the authors use the target’s self-report as the reference method and peer reports as well as aggregated state ratings as nonreference methods. Peer reports are interchangeable methods and located on level 1, the self-report and aggregated state ratings are structurally different methods and located on level 2. A unique method factor on level 1 and a common method factor on level 2 contrast peer reports with the self-report. A second method factor on level 2 is defined to capture the deviation of the aggregated state ratings from the expected value given the self-report.

We 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 choose the self-report as the reference method and we use three indicators per trait. Again, the superscript “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 on level 2 is given by the same equation as in the model of Eid et al. (2008):

$$Y_{tik}^{\mathrm{RM}} = \mu_{ik}^{\mathrm{RM}} + T_{tik}^{\mathrm{RM}} + E_{tik}^{\mathrm{RM}}. \tag{1}$$

The following equation decomposes the nonreference-method indicators of both groups of interchangeable methods (subordinates and peers) on level 1 by adding the index *m* to all manifest and latent variables and factor loadings.

$$Y_{rtik}^{\mathrm{NM}_m} = \mu_{ik}^{\mathrm{NM}_m} + T_{tik}^{\mathrm{NM}_m} + \lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m} UM_{rtk}^{\mathrm{NM}_m} + E_{rtik}^{\mathrm{NM}_m} . \tag{5}$$

The meaning of all components is the same as in Equation 2. 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. To contrast the view of subordinates and peers with the view of the target person, the trait variables of the nonreference methods $T_{tik}^{\mathrm{NM}_m}$ are regressed on the trait variables of the reference method T_{tik}^{RM}. Again, the residuals of this latent regression are assumed to be unidimensionally measuring a common method factor $CM_{tk}^{\mathrm{NM}_m}$. This factor captures that part of the nonreference-method trait factors that is shared among the particular group of subordinates or peers but that is not shared with the reference method (i.e., the self-report). The following equation expresses this decomposition of the nonreference-method trait factors:

$$T_{tik}^{\mathrm{NM}_m} = \lambda_{\mathrm{T}ik}^{\mathrm{NM}_m} T_{tik}^{\mathrm{RM}} + \lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m} CM_{tk}^{\mathrm{NM}_m} \ . \tag{6}$$

The meaning of all summands can be adopted from Equation 3, while all nonreference-method components receive the additional index *m* to distinguish between the two groups of interchangeable methods.

Inserting Equation 6 into Equation 5 results in the model equation for the nonreference-method indicators:

$$Y_{rtik}^{\mathrm{NM}_m} = \mu_{ik}^{\mathrm{NM}_m} + \lambda_{\mathrm{T}ik}^{\mathrm{NM}_m} T_{tik}^{\mathrm{RM}} + \lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m} CM_{tk}^{\mathrm{NM}_m} + \lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m} UM_{rtk}^{\mathrm{NM}_m} + E_{rtik}^{\mathrm{NM}_m}. \tag{7}$$

In this model, the trait variables T_{tik}^{RM} measure the constructs from the perspective of the reference method, that is the self-report in our application. Additionally, 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.

Another advantageous feature of the model is that the user is flexible in choosing the reference method. Each of the interchangeable level 1 methods can serve as the reference method on level 2 by using the trait variables of the peer reports or the trait variables of the subordinate reports (Geiser, Eid, & Nussbeck, 2008). If, for example, the subordinate reports were considered as gold standard in a particular application because they were assumed to measure the construct more appropriately than the self-report one could regress the self-ratings and the trait variables of the peer ratings on the trait variables of the subordinate ratings on level 2. The reference factor would then capture the construct from the subordinates’ point of view. The targets’ self-ratings as well as the peer reports would be expressed as residuals from the reference method, i.e., the subordinate reports, via method factors.

The model is identified by fixing the means of the latent traits T_{tik}^{RM} and $T_{tik}^{\mathrm{NM}_m}$ to zero and setting the first factor loading of all factors to one. The means of the method factors $CM_{tk}^{\mathrm{NM}_m}$ and $UM_{rtk}^{\mathrm{NM}_m}$ are zero because these factors depict residuals. The list of model variables that are uncorrelated can be taken from the model of Eid et al. (2008) presented above. Additionally, 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. Appendix A shows the definitions of all latent variables.

As the latent variables on the right side of Equation 7 are uncorrelated, the variance of an observed nonreference-method variable can be decomposed in the following way:

$$Var\left(Y_{rtik}^{\mathrm{NM}_m}\right) = \left(\lambda_{\mathrm{T}ik}^{\mathrm{NM}_m}\right)^2 Var\left(T_{tik}^{\mathrm{RM}}\right) + \left(\lambda_{\mathrm{CM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(CM_{tk}^{\mathrm{NM}_m}\right) + \left(\lambda_{\mathrm{UM}ik}^{\mathrm{NM}_m}\right)^2 Var\left(UM_{rtk}^{\mathrm{NM}_m}\right) + Var\left(E_{rtik}^{\mathrm{NM}_m}\right). \qquad (8)$$

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 with respect to the reference method. 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.

The formal extension of the model from one to two sets of interchangeable methods on level 1 is straightforward. However, the question of how the model can be defined within the statistical software is intricate 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).

3. Application of the Model to 360-Degree Feedback Data

3.1 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 and nine items, respectively.

As we do not intend to analyze the internal structure of the Benchmarks® instrument but to separate method specific influences from measurement error we built item parcels. 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. We parceled across items, not across methods. Thus, the parceling procedure was applied to all trait-method-units (TMUs) separately so that parcels do not contain cross-method information. Our approach resulted in three indicators per trait-method unit. 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.

3.2 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. Annotated Mplus code is shown in Appendix B. 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. However, this approach introduces quite a lot of missing data into the analyses, which can be conceived of as planned missing data. The resulting data structure can be viewed as a special case of the two-method measurement approach to planned missing data handling (Graham, Taylor, Olchowski, & Cumsille, 2006). Much like the model introduced in this study, the two-method measurement design makes use of the CT-C(M-1) modeling approach—i.e., setting a reference method against which other methods are contrasted—and uses full information likelihood estimation to incorporate planned missing data. This approach to planned missing data handling has been shown to perform well and result in negligible parameters and standard error biases when using full information maximum likelihood estimation (Graham et al., 2006, Garnier-Villareal, Rhemtulla, & Little, 2014). While the data analyzed here does not contain missing values beyond the planned missing values introduced by this approach, it should be noted that non-planned missingness might not be optimally represented in this approach.

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 (i.e., subordinates), the last three columns display the parcel values of the nonreference method 2 (i.e., 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 planned 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.

4. Results

The ML-CFA-MTMM model shown in Figure 2 fit the data well, $\chi^2(154, N = 52{,}265) = 2{,}531.93$, $p < .001$, RMSEA = .017, CFI = .99, SRMR (level 1) = .01, SRMR (level 2) = .10. We also specified a model with common (instead of indicator-specific) trait variables (Figure 3). This model showed good fit as well but was worse than the less restrictive model with indicator-specific trait variables (AIC = 404,960.45 and BIC = 405,962.10 for the model with indicator-specific traits vs. AIC = 407,668.91 and BIC = 408,475.55 for the model with general traits). The means, the loading parameters and the coefficients of consistency, common and unique method specificity, and reliability of the model with indicator-specific trait variables are presented in Table 4.

All reliabilities were acceptable to very good (between .63 and .91), especially when considering that each parcel consists of only three to five items. The consistency coefficients showed that between 5% and 9% of the error-free nonreference-method variance was shared with the self-report. These results revealed a very low convergent validity of self-ratings and ratings from subordinates and peers[1]. The amount of true variance that was shared among the group of subordinates (or among the group of peers) but not with the self-report (common method specificity) varied between 22% and 26%. The unique method specificity was by far the major source of variance and indicated that between 64% and 73% of true variance was neither shared with the self-report nor with the other subordinates or peers but was specific to the single nonreference-method raters. The coefficients in Table 1 represent to which degree interindividual rater differences are due to the different sources of variance. They give an answer to the question why individual raters differ.

It might also be interesting to analyze rater differences on level 2, that means, rater differences that are free of unique method effects and measurement error. These differences are related to differences between "average" ratings of a target. If one divides the consistency coefficient by the sum of the consistency and common method specificity coefficient one obtains a variance component that indicates the convergent validity of the reference method and the nonreference methods on level 2. These level 2 consistency coefficients are much larger than the consistency coefficients in Table 4. The values are between 0.18 and 0.26 indicating that between 18% and 26% of the trait variables of the nonreference methods can be determined by the reference method. If one takes the square root of these coefficients one obtains the correlations between the latent trait variables of the reference method and the nonreference methods. These correlations are between $r = 0.42$ and $r = 0.51$ indicating high convergent validity. The fact that these convergent validities are larger than the ones in Table 4 can be explained by the aggregation effect. It shows that it is more appropriate to consider average subordinate and average peer ratings than single subordinate and peer ratings if one is interested in the assessment of leadership behavior.

[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.

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). There are two different kinds of trait factor correlations. First, correlations between trait factors belonging to the same construct (*r* = .84 to .91 for Leading Employees and *r* = .88 to .95 for Participative Management) revealed that the indicator-specific trait variables were nearly homogeneous. However, in the model comparison presented above, the model with general trait variables showed a worse fit than the model with indicator-specific trait variables. Second, the high correlations of *r* = .65 to .85 between trait variables belonging to different constructs showed that leaders tended to rate themselves in the same manner on both constructs, indicating low discriminant validity of the two constructs as measured by the reference method, i.e., the self-report. It was also striking that the common method effect generalized across constructs (*r* = .95 for subordinates and *r* = .94 for peers) indicating low discriminant validity with respect to the nonreference methods. Common method factors were even strongly correlated across different traits *and* different sets of nonreference methods (*r* = .69 and *r* = .68), 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. These correlations are measures of low discriminant validity corrected for method-specific influences.

The generalizability also held for the unique method factors (*r* = .93 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. This shows that there is also low discriminant validity on the level of the single raters belonging to a nonreference method.

There are strong correlations between the common method factors of the subordinates and peers (*r* = .73 for Leading Employees and *r* = .70 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). These correlations indicate high convergent validity on the level of the nonreference methods. It is important to note that these correlations are correlations of variables that are free of measurement error and unique method effects. This is in line with the aggregation effect mentioned above. It shows again that it is more appropriate to consider average subordinate and average peer ratings to assess leadership behavior with higher validity.

5. Discussion

The model presented in this paper is the first ML-CFA-MTMM model for data structures with two (or more) populations of interchangeable 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 high correlation between the two constructs measured by the self-report indicates low discriminant validity on the level of the reference method. Not only the two trait variables measured by the self-report showed high correlations but also common method effects and unique method effects correlated strongly across the two constructs. This 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 was 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. Moreover, our analyses have shown that the consistency is much higher if it is not analyzed on the level of single raters but on the level of "average" ratings. This shows that one should not trust single subordinate and peer ratings if one is interested in the valid assessment of leadership behavior. Instead it is important to consider several ratings as it is usually done.

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 was considerably higher than the consistency between self-reports and others' ratings. This indicates that subordinates (and peers) shared a common view that differs from the target's self-perception. Furthermore, the model allows to state that the common view generalized across nonreference methods and across traits: On the one hand, subordinates (and peers) who over- or underestimated the target in their common rating of Leading Employees tended to have the same bias in their common rating of Participative Management. This association was very strong and showed that the rater groups might not distinguish between these facets. On the other hand, subordinates and peers tended 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 was high agreement between subordinates and peers in the evaluation of a given manager. This indicates high convergent validity on the level of the nonreference methods.

Further studies need to reveal if this pattern is replicated. 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 was by far the major source of variance if one considers the differences between single raters. This result is in agreement with previous studies (Greguras & Robie, 1998; Lance, 1994; Scullen, Mount, & Goff, 2000; Woehr, Sheehan, & Bennett Jr., 2005) and supports 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, in particular, if one considers the ratings on the aggregate level.

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. Finally, our approach of how to deal with multiple populations of level 1 units can be applied not only to the multilevel CT-C(M-1) model that we presented here but also to other multilevel MTMM models.

5.1 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. According to our experience the incidence of raters rating more than one target is below 1% in our data set because it was collected in different organizations and at different points in time.

In their simulation study Schultze, Koch, and Eid (in press) 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 used a sub-model of the model presented here with only one set of methods on level 1. The results indicated that the existence of raters evaluating more than one target has no effect on the decomposition of variance (reliability, convergent validity of the different types of raters, and the degree of agreement between interchangeable raters) that was presented above because parameter estimates were only minimally distorted. Standard errors, on the other hand, are biased to such a degree that the authors recommended not trusting the inferences made regarding the latent covariance structure on level 1, i.e., the correlation between unique method factors, and trusting those concerning the covariance between the common method factors only in situations with sufficiently many raters per target (5 or more). However, because the sample size is extremely large in our application, and the parameters are therefore estimated with very high precision, and because the number of raters is very high and the number of overlapping raters very low, also the inferential conclusions can be trusted in our analysis.

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. This requirement is fulfilled in our application. We are currently running simulation studies to reveal whether this recommendation holds for this type of model as well. 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.

5.2 Conclusions

We presented a model that offers many new opportunities for researchers who wish to analyze MTMM data with multiple sets of interchangeable 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-multimethod 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

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. D. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development, 38*, 411-422. doi:10.1177/0165025414542711

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 & Coffman (2006). *Psychological Methods, 13,* 49-57. doi:10.1037/1082-989X.13.1.49

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

Graham J. W., Taylor B. J., Olchowski A. E., & Cumsille P.E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323-343. doi:10.1037/1082-989x.11.4.323

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

Ruch, W., Köhler, G., & van Thriel, C. (1996). Assessing the temperamental basis of the sense of humor: Construction of the facet and standard trait forms of the State-Trait-Cheerfulness-Inventory – STCI. *International Journal of Humor Research, 9, 303–339*. doi: 10.1515/humr.1996.9.3-4.303

Schultze, M., Koch, T., & Eid, M. (in press). The effects of non-independent rater sets in Multilevel-Multitrait-Multimethod models. *Structural Equation Modeling*.

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

Appendix A

Random Experiment

The ML-CFA-MTMM model for a combination of structurally different methods and two sets of interchangeable methods as described in the text can be defined based on the following random experiment:

$$\Omega = \Omega_T^{\mathrm{RM}} \times \Omega_O^{\mathrm{RM}} \times \Omega_R^{\mathrm{NM}_1} \times \Omega_O^{\mathrm{NM}_1} \times \Omega_R^{\mathrm{NM}_2} \times \Omega_O^{\mathrm{NM}_2}, \qquad \text{(A1)}$$

where Ω_T^{RM} is the set of possible targets (e.g., supervisors), Ω_O^{RM} is the set of possible behaviors or observations of the particular target *t*, $\Omega_R^{\mathrm{NM}_1}$ is the first set of interchangeable raters (e.g., peers) for a particular target *t*, $\Omega_O^{\mathrm{NM}_1}$ is the set of possible observations (ratings) for target *t* rated by rater *r* stemming from the first set of interchangeable raters $\Omega_R^{\mathrm{NM}_1}$, $\Omega_R^{\mathrm{NM}_2}$ is the second set of interchangeable raters (e.g., subordinates) for the particular target *t*, $\Omega_O^{\mathrm{NM}_2}$ is the set of possible observations (ratings) for target *t* rated by rater *r* stemming from the second set of interchangeable raters $\Omega_R^{\mathrm{NM}_2}$, and × is the set product operator. The superscript RM denotes the reference (gold standard) method and the superscript NM$_m$ refers to the nonreference method.

In sum, the above random experiment describes the following sampling procedure:

1. A target person (ratee, observation unit, *t*) is randomly sampled from a set of all possible targets Ω_T^{RM}.
2. Then, the self-reported behavior of target *t* is recorded. Note that the set Ω_O^{RM} of possible observations is itself a set product $\Omega_O^{\mathrm{RM}} = \Omega_{O_{11}}^{\mathrm{RM}} \times \ldots \times \Omega_{O_{ik}}^{\mathrm{RM}} \times \ldots \times \Omega_{O_{ln}}^{\mathrm{RM}}$ of possible observations for indicator *i* and construct *k.*

3. With respect to the particular target *t*, a rater r_{NM_1} is randomly sampled from the *first* set of possible interchangeable raters (e.g., peers, $\Omega_R^{\mathrm{NM}_1}$).
4. The possible observations of rater *r* from the *first* set of interchangeable raters for target *t* are recorded $\Omega_O^{\mathrm{NM}_1}$. Again, $\Omega_O^{\mathrm{NM}_1}$ is itself a set product $\Omega_O^{\mathrm{NM}_1} = \Omega_{O_{11}}^{\mathrm{NM}_1} \times \ldots \times \Omega_{O_{ik}}^{\mathrm{NM}_1} \times \ldots \times \Omega_{O_{ln}}^{\mathrm{NM}_1}$ of possible observations of indicator *i* and construct *k*.
5. In addition, rater r_{NM_2} is randomly sampled from the *second* set of possible interchangeable raters (e.g., subordinates, $\Omega_R^{\mathrm{NM}_2}$) for the particular target *t*. Note that the two sets of interchangeable raters $\Omega_R^{\mathrm{NM}_1}$ and $\Omega_R^{\mathrm{NM}_2}$ are structurally different from one another, meaning that each set of interchangeable raters represents a different population of possible raters (e.g., peers, subordinates, etc.) for a particular target *t*.
6. Finally, the possible observations of rater *r* from the *second* set of interchangeable raters for target *t* are recorded $\Omega_O^{\mathrm{NM}_2}$, with $\Omega_O^{\mathrm{NM}_2}$ being a set product following the similar logic as described above.

Two aspects of the random experiment are important and shall be emphasized. First, the self-reports (observations, Ω_O^{RM}) are fix for each target *t*. That means, there is only one (self-) rating for each target. In contrast to that, the observations (ratings) of the interchangeable raters require to sample the target person and the particular rater from the m^{th} set (here: 1 or 2) of possible interchangeable raters. The second important aspect is that the interchangeable ratings of different sets of interchangeable raters m and m' are mutually exclusive from another, meaning that the ratings of rater r_{NM_1}rating target *t* exists only for the *first* set of interchangeable raters, not for the *second* set of interchangeable raters. In other words, ratings for the second set of interchangeable raters are missing by design for target *t* and rater r_{NM_1}.

In order to define the latent variables as random variables, we consider the following mappings:

1. The mapping of the possible outcomes Ω into the sets of possible targets Ω_T^{RM}; i.e., $\mathrm{p}_T^{\mathrm{RM}}: \Omega \rightarrow \Omega_T^{\mathrm{RM}}$.
2. The mapping of the possible outcomes Ω into the m^{th} set of possible interchangeable raters $\Omega_R^{\mathrm{NM}_m}$; i.e., $\mathrm{p}_R^{\mathrm{NM}_m}: \Omega \rightarrow \Omega_R^{\mathrm{NM}_m}$.
3. The mapping of the possible outcomes Ω into the sets of real numbers; i.e., $Y_{tik}^{\mathrm{RM}}: \Omega \rightarrow \mathbb{R}$ and $Y_{rtik}^{\mathrm{NM}_m}: \Omega \rightarrow \mathbb{R}$,

where $\mathrm{p}_T^{\mathrm{RM}}$ is the target-variable, $p_R^{\mathrm{NM}_m}$ is the rater-variable pertaining to the m^{th} set of possible interchangeable raters, Y_{tik}^{RM} is the observed variable of the targets' self-reports, and $Y_{rtik}^{\mathrm{NM}_m}$ is the observed variable of the m^{th} set of possible interchangeable raters.

Definition of True Scores and Measurement Error Variables

Based on the random experiment described above, the true score and corresponding measurement error variables can be defined as follows:

$$\tau_{tik}^{\mathrm{RM}} \coloneqq E(Y_{tik}^{\mathrm{RM}} | \mathrm{p}_T^{\mathrm{RM}}) \qquad \text{(reference-method true scores)} \qquad \text{(A2)}$$

$$\tau_{rtik}^{\mathrm{NM}_m} \coloneqq E(Y_{rtik}^{\mathrm{NM}_m} | \mathrm{p}_T^{\mathrm{RM}}, \mathrm{p}_R^{\mathrm{NM}_m}) \qquad \text{(nonreference-method true scores)} \qquad \text{(A3)}$$

$$E_{tik}^{\mathrm{RM}} \coloneqq Y_{tik}^{\mathrm{RM}} - E(Y_{tik}^{\mathrm{RM}} | \mathrm{p}_T^{\mathrm{RM}}) \qquad \text{(reference-method error variables)} \qquad \text{(A4)}$$

$$E_{rtik}^{\mathrm{NM}_m} \coloneqq Y_{rtik}^{\mathrm{NM}_m} - E(Y_{rtik}^{\mathrm{NM}_m} | \mathrm{p}_T^{\mathrm{RM}}, \mathrm{p}_R^{\mathrm{NM}_m}) \qquad \text{(nonreference-method error variables)} \qquad \text{(A5)}$$

It is important to note that the true scores $\tau_{rtik}^{\mathrm{NM}_m}$ and error variables $E_{rtik}^{\mathrm{NM}_m}$ are measured at level 1 (i.e., the rater level) and the true scores τ_{tik}^{RM} and the corresponding error variables E_{tik}^{RM} are measured at level 2 (i.e., target level). To relate both true scores to one another, the conditional expectations of the level 1 true scores given the target variable are considered:

$$\tau_{tik}^{\mathrm{NM}_m} \coloneqq E(\tau_{rtik}^{\mathrm{NM}_m}|\mathrm{p}_T^{\mathrm{RM}}) = E\left[E\left(Y_{rtik}^{\mathrm{NM}_m}|\mathrm{p}_T^{\mathrm{RM}}, p_R^{\mathrm{NM}_m}\right)|\mathrm{p}_T^{\mathrm{RM}}\right] \quad \text{(A6)}$$

The level 2 true scores $\tau_{tik}^{\mathrm{NM}_m}$ can be conceived as expectations of the true ratings for a particular target across all interchangeable raters of the m^{th} set for that specific target. Note also that $\tau_{tik}^{\mathrm{NM}_m}$ represents the true scores at level 2 (target level) for both sets of interchangeable raters (e.g., peers and subordinates), hence $\tau_{tik}^{\mathrm{NM}_1}$ and $\tau_{tik}^{\mathrm{NM}_2}$.

The residuals of these latent regressions are defined as unique method variables at level 1:

$$UM_{rtik}^{\mathrm{NM}_m} := \tau_{rtik}^{\mathrm{NM}_m} - \tau_{tik}^{\mathrm{NM}_m} \quad \text{(A7)}$$

Given that these variables are defined as latent residual variables, they have an expectation (mean) of zero and are uncorrelated with $\tau_{tik}^{\mathrm{NM}_m}$. In addition, it is assumed that these variables are independent and identically distributed across target persons.

Next, the true scores of the nonreference methods are regressed on the true scores of the reference method at level 2:

$$E\left(\tau_{tik}^{\mathrm{NM}_m}\middle|\tau_{tik}^{\mathrm{RM}}\right) = \mu_{tik} + \lambda_{tik}\tau_{tik}^{\mathrm{RM}} \quad \text{(A8)}$$

Again, the residuals of these latent regressions are defined as common method variables at level 2:

$$CM_{tik}^{\mathrm{NM}_m} \coloneqq \tau_{tik}^{\mathrm{NM}_m} - E\left(\tau_{tik}^{\mathrm{NM}_m}\middle|\tau_{tik}^{\mathrm{RM}}\right) = \tau_{tik}^{\mathrm{NM}_m} - (\mu_{tik} + \lambda_{tik}\tau_{tik}^{\mathrm{RM}}) \quad \text{(A9)}$$

Again, the common method variables are defined as zero-mean normally distributed residual variables that are uncorrelated with τ_{tik}^{RM}..

Appendix B

Annotated Mplus Code

```
! Mplus input for the model with indicator-specific traits

TITLE: Multilevel confirmatory factor analysis multitrait-multimethod (ML-
       CFA-MTMM) model for two sets of interchangeable methods and one
       structurally different method with indicator-specific traits

! Name of the .dat-file with the data
DATA:

  FILE = lead_part.dat;

VARIABLE:

  ! Definition of the variable names
  NAMES =   ID  par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm
            par1_nm1 par2_nm1 par3_nm1 lead1_nm1 lead2_nm1 lead3_nm1
            par1_nm2 par2_nm2 par3_nm2 lead1_nm2 lead2_nm2 lead3_nm2;

            ! par1-par3 -> Indicators of Participative Management
            ! lead1-lead3 -> Indicators of Leading Employees
            ! _rm -> Reference-method reports (self-reports)
            ! _nm1 -> Nonreference-method 1 reports (subordinates)
            ! _nm2 -> Nonreference-method 2 reports (peers)

  ! Definition of the variables to be used in the analysis
  USEVAR =  par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm
            par1_nm1 par2_nm1 par3_nm1 lead1_nm1 lead2_nm1 lead3_nm1
            par1_nm2 par2_nm2 par3_nm2 lead1_nm2 lead2_nm2 lead3_nm2;

  ! Definition of the indicators that are used on level 2 only
  !(between level)
  BETWEEN = par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm;

  ! Definition of missing value flag (The only missing values in the data
  ! are those that we generated by ourselves in order to create the data
  ! format shown in Table 3.)
  MISSING = ALL(-9999);

  ! Definition of the cluster variable (indicating the target)
  CLUSTER = ID;

ANALYSIS:

  ! Request a multilevel CFA
  TYPE = TWOLEVEL;
  H1ITERATIONS = 10000;

MODEL:

  ! Specification of the model on level 1 (Within level)
  ! A unique method factor is defined for each trait-method-unit (TMU).
  ! The first loading of each factor is fixed to one for identification
  ! purposes.
  %WITHIN%

  UMrt1_nm1 by lead1_nm1@1 lead2_nm1 lead3_nm1;
  UMrt1_nm2 by lead1_nm2@1 lead2_nm2 lead3_nm2;
  UMrt2_nm1 by par1_nm1@1 par2_nm1 par3_nm1;
```

```
UMrt2_nm2 by par1_nm2@1 par2_nm2 par3_nm2;

! Correlations between unique method factors belonging to different
! nonreference methods are set to zero.
UMrt1_nm1 with UMrt1_nm2@0 UMrt2_nm2@0;
UMrt2_nm1 with UMrt1_nm2@0 UMrt2_nm2@0;

! Specification of the model on level 2 (Between level)
%BETWEEN%

! Definition of the indicator-specific trait variables Ttik_nm1 and
! Ttik_nm2 for the nonreference methodsTt11_nm1 by lead1_nm1@1;
Tt21_nm1 by lead2_nm1@1;
Tt31_nm1 by lead3_nm1@1;
Tt11_nm2 by lead1_nm2@1;
Tt21_nm2 by lead2_nm2@1;
Tt31_nm2 by lead3_nm2@1;
Tt12_nm1 by par1_nm1@1;
Tt22_nm1 by par2_nm1@1;
Tt32_nm1 by par3_nm1@1;
Tt12_nm2 by par1_nm2@1;
Tt22_nm2 by par2_nm2@1;
Tt32_nm2 by par3_nm2@1;

! Definition of the indicator-specific trait factors Ttik_rm for the
! reference method and regression on the trait variables of the
! nonreference methods
Tt11_rm by lead1_rm@1 Tt11_nm1 Tt11_nm2;
Tt21_rm by lead2_rm@1 Tt21_nm1 Tt21_nm2;
Tt31_rm by lead3_rm@1 Tt31_nm1 Tt31_nm2;
Tt12_rm by par1_rm@1 Tt12_nm1 Tt12_nm2;
Tt22_rm by par2_rm@1 Tt22_nm1 Tt22_nm2;
Tt32_rm by par3_rm@1 Tt32_nm1 Tt32_nm2;

! Definition of the trait-specific method factors
CMt1_nm1 by Tt11_nm1@1 Tt21_nm1 Tt31_nm1;
CMt1_nm2 by Tt11_nm2@1 Tt21_nm2 Tt31_nm2;
CMt2_nm1 by Tt12_nm1@1 Tt22_nm1 Tt32_nm1;
CMt2_nm2 by Tt12_nm2@1 Tt22_nm2 Tt32_nm2;

! Residual variances of the nonreference-method trait variables are set
! to zero as the trait variables are assumed to be homogeneous
Tt11_nm1@0;
Tt21_nm1@0;
Tt31_nm1@0;
Tt11_nm2@0;
Tt21_nm2@0;
Tt31_nm2@0;

! Correlations between reference-method trait variables and common
! method factors belonging to the same trait-method-unit (TMU) are set
! to zero.
Tt11_rm with CMt1_nm1@0 CMt1_nm2@0;
Tt21_rm with CMt1_nm1@0 CMt1_nm2@0;
Tt31_rm with CMt1_nm1@0 CMt1_nm2@0;
Tt12_rm with CMt2_nm1@0 CMt2_nm2@0;
Tt22_rm with CMt2_nm1@0 CMt2_nm2@0;
Tt32_rm with CMt2_nm1@0 CMt2_nm2@0;

! Residual variances of the nonreference-method indicators are set to
! zero as they are estimated on the within level.
```

```
  lead1_nm1@0;
  lead2_nm1@0;
  lead3_nm1@0;
  lead1_nm2@0;
  lead2_nm2@0;
  lead3_nm2@0;
  par1_nm1@0;
  par2_nm1@0;
  par3_nm1@0;
  par1_nm2@0;
  par2_nm2@0;
  par3_nm2@0;

! Request sample statistics and standardized solution
OUTPUT: SAMPSTAT STDYX;
```

```
! Mplus input for the model with general trait factors

TITLE: Multilevel confirmatory factor analysis multitrait-multimethod (ML-
       CFA-MTMM) model for two sets of interchangeable methods and one
       structurally different method with general trait factors

! Name of the .dat-file with the data
DATA:

   FILE = lead_part.dat;

VARIABLE:
   ! Definition of the variable names
   NAMES =   ID  par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm
             par1_nm1 par2_nm1 par3_nm1 lead1_nm1 lead2_nm1 lead3_nm1
             par1_nm2 par2_nm2 par3_nm2 lead1_nm2 lead2_nm2 lead3_nm2;

             ! par1-par3 -> Indicators of Participative Management
             ! lead1-lead3 -> Indicators of Leading Employees
             ! _rm -> Reference-method reports (self-reports)
             ! _nm1 -> Nonreference-method 1 reports (subordinates)
             ! _nm2 -> Nonreference-method 2 reports (peers)

   ! Definition of the variables to be used in the analysis
   USEVAR =  par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm
              par1_nm1 par2_nm1 par3_nm1 lead1_nm1 lead2_nm1 lead3_nm1
              par1_nm2 par2_nm2 par3_nm2 lead1_nm2 lead2_nm2 lead3_nm2;

   ! Definition of the indicators that are used on level 2 only
   ! (between level)
   BETWEEN = par1_rm, par2_rm par3_rm lead1_rm lead2_rm lead3_rm;

   ! Definition of missing value flag (The only missing values in the data
   ! are those that we generated by ourselves in order to create the data
   ! format shown in Table 3.)
   MISSING=ALL(-9999);

   ! Definition of the cluster variable (indicating the target)
   CLUSTER = ID;

ANALYSIS:

   ! Request a multilevel CFA
   TYPE = TWOLEVEL;
   H1ITERATIONS = 10000;

MODEL:

   ! Specification of the model on level 1 (Within level)
   ! A unique method factor is defined for each trait-method-unit (TMU).
   ! The first loading of each factor is fixed to one for identification
   ! purposes.
   %WITHIN%

   UMrt1_nm1 by lead1_nm1@1 lead2_nm1 lead3_nm1;
   UMrt1_nm2 by lead1_nm2@1 lead2_nm2 lead3_nm2;
   UMrt2_nm1 by par1_nm1@1 par2_nm1 par3_nm1;
   UMrt2_nm2 by par1_nm2@1 par2_nm2 par3_nm2;

   ! Correlations between unique method factors belonging to different
   ! nonreference methods are set to zero.
```

```
UMrt1_nm1 with UMrt1_nm2@0 UMrt2_nm2@0;
UMrt2_nm1 with UMrt1_nm2@0 UMrt2_nm2@0;

! Specification of the model on level 2 (Between level)
%BETWEEN%

! Definition of the indicator-specific trait variables Ttik_nm1 and
! Ttik_nm2 for the nonreference methods
Tt11_nm1 by lead1_nm1@1;
Tt21_nm1 by lead2_nm1@1;
Tt31_nm1 by lead3_nm1@1;
Tt11_nm2 by lead1_nm2@1;
Tt21_nm2 by lead2_nm2@1;
Tt31_nm2 by lead3_nm2@1;
Tt12_nm1 by par1_nm1@1;
Tt22_nm1 by par2_nm1@1;
Tt32_nm1 by par3_nm1@1;
Tt12_nm2 by par1_nm2@1;
Tt22_nm2 by par2_nm2@1;
Tt32_nm2 by par3_nm2@1;

! Definition of the general trait factors Ttk_rm for the reference
! method and regression on the trait variables of the nonreference
! methods
Tt1_rm by lead1_rm@1 lead2_rm lead3_rm
Tt11_nm1 Tt21_nm1 Tt31_nm1
Tt11_nm2 Tt21_nm2 Tt31_nm2;
Tt2_rm by par1_rm@1 par2_rm par3_rm
Tt12_nm1 Tt22_nm1 Tt32_nm1
Tt12_nm2 Tt22_nm2 Tt32_nm2;

! Definition of the trait-specific method factors
CMt1_nm1 by Tt11_nm1@1 Tt21_nm1 Tt31_nm1;
CMt1_nm2 by Tt11_nm2@1 Tt21_nm2 Tt31_nm2;
CMt2_nm1 by Tt12_nm1@1 Tt22_nm1 Tt32_nm1;
CMt2_nm2 by Tt12_nm2@1 Tt22_nm2 Tt32_nm2;

! Residual variances of the nonreference-method trait variables are set
! to zero as the trait variables are assumed to be homogeneous
Tt11_nm1@0;
Tt21_nm1@0;
Tt31_nm1@0;
Tt11_nm2@0;
Tt21_nm2@0;
Tt31_nm2@0;

! Correlations between reference-method trait variables and common
! method factors belonging to the same trait-method-unit (TMU) are set
! to zero.
Tt1_rm with CMt1_nm1@0 CMt1_nm2@0;
Tt2_rm with CMt2_nm1@0 CMt2_nm2@0;

! Residual variances of the nonreference-method indicators are set to
! zero as they are estimated on the within level.
lead1_nm1@0;
lead2_nm1@0;
lead3_nm1@0;
lead1_nm2@0;
lead2_nm2@0;
lead3_nm2@0;
par1_nm1@0;
```

```
   par2_nm1@0;
   par3_nm1@0;
   par1_nm2@0;
   par2_nm2@0;
   par3_nm2@0;

! Request sample statistics and standardized solution
OUTPUT: SAMPSTAT STDYX;
```

Table 1

Definition of the Variance Components for the Indicators of the Nonreference Methods

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

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

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

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

Note. CO = consistency coefficient; *CMS* = common method specificity coefficient; *UMS* = unique method specificity coefficient; *Rel* = reliability coefficient; $\tau_{rtik}^{\mathrm{NM}_m}$ = true score variables of the nonreference methods; T_{tik}^{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_{tik}^{RM}	Trait factor of the reference method	Indicator-specific 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		
$CO(\tau_{rtik}^{\mathrm{NM}_m})$	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)
$CMS(\tau_{rtik}^{\mathrm{NM}_m})$	Common method specificity	Proportion of true variance that is due to the common view of subordinates or peers not shared with the self-report
$UMS(\tau_{rtik}^{\mathrm{NM}_m})$	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
$Rel(\tau_{rtik}^{\mathrm{NM}_m})$	Reliability	Proportion of manifest variance that is not due to measurement error
Factor correlations		
$Cor(T_{ti1}^{\mathrm{RM}}, T_{ti1}^{\mathrm{RM}})$, $Cor(T_{ti2}^{\mathrm{RM}}, T_{ti2}^{\mathrm{RM}})$	Homogeneity of the indicator-specific trait variables	Degree to which the three indicators per trait measure the trait homogeneously
$Cor(T_{ti1}^{\mathrm{RM}}, T_{ti2}^{\mathrm{RM}})$	Discriminant validity with respect to the reference method	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); degree of convergent validity on the level of nonreference methods
$Cor(CM_{t1}^{\mathrm{NM}_1}, CM_{t2}^{\mathrm{NM}_1})$, $Cor(CM_{t1}^{\mathrm{NM}_2}, CM_{t2}^{\mathrm{NM}_2})$	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; degree of discriminant validity on the level of nonreference methods
$Cor(UM_{rt1}^{\mathrm{NM}_1}, UM_{rt2}^{\mathrm{NM}_1})$, $Cor(UM_{rt1}^{\mathrm{NM}_2}, UM_{rt2}^{\mathrm{NM}_2})$	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); degree of discriminant validity on level 1
$Cor(T_{ti1}^{\mathrm{RM}}, CM_{t2}^{\mathrm{NM}_1})$, $Cor(T_{ti1}^{\mathrm{RM}}, CM_{t2}^{\mathrm{NM}_2})$, $Cor(T_{ti2}^{\mathrm{RM}}, CM_{t1}^{\mathrm{NM}_1})$, $Cor(T_{ti2}^{\mathrm{RM}}, CM_{t1}^{\mathrm{NM}_2})$	Correlation between a reference-method trait factor and a common method factor of the other trait	Degree to which the common bias of subordinates or peers on one trait is related to the other trait (measured by the self-report)

Note. $\tau_{rtik}^{\mathrm{NM}_m}$ = true score 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 two Sets of Interchangeable Methods and one Structurally Different Method

	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	$\mu_{ik}^{\mathrm{RM}}/\mu_{ik}^{\mathrm{NM}m}$	$\lambda_{\mathrm{T}ik}^{\mathrm{NM}m}$	$\lambda_{\mathrm{CM}ik}^{\mathrm{NM}m}$	$\lambda_{\mathrm{UM}ik}^{\mathrm{NM}m}$	$CO(\tau_{rtik}^{\mathrm{NM}m})$	$CMS(\tau_{rtik}^{\mathrm{NM}m})$	$UMS(\tau_{rtik}^{\mathrm{NM}m})$	$Rel(\tau_{tik}^{\mathrm{RM}})$/ $Rel(\tau_{rtik}^{\mathrm{NM}m})$
Leading Employees								
Self								
Y_{t11}^{RM}	3.92							0.81
Y_{t21}^{RM}	4.02							0.63
Y_{t31}^{RM}	3.81							0.74
Subordinates								
Y_{rt11}^{NM}	4.00	0.45	1.00	1.00	0.06	0.24	0.70	0.91
Y_{rt21}^{NM}	4.14	0.43	0.82	0.82	0.08	0.23	0.69	0.73
Y_{rt31}^{NM}	4.00	0.45	0.97	0.96	0.07	0.24	0.69	0.86
Peers								
Y_{rt11}^{NM}	3.95	0.38	1.00	1.00	0.06	0.24	0.70	0.88
Y_{rt21}^{NM}	3.96	0.44	0.97	0.88	0.09	0.26	0.64	0.71
Y_{rt31}^{NM}	3.94	0.39	0.99	0.95	0.08	0.25	0.67	0.80
Participative Management								
Self								
Y_{t12}^{RM}	4.04							0.75
Y_{t22}^{RM}	4.02							0.70
Y_{t32}^{RM}	3.92							0.66
Subordinates								
Y_{rt12}^{NM}	4.10	0.38	1.00	1.00	0.05	0.23	0.72	0.88
Y_{rt22}^{NM}	4.09	0.41	1.03	0.96	0.06	0.25	0.69	0.84
Y_{rt32}^{NM}	4.01	0.39	0.93	0.96	0.05	0.22	0.73	0.81
Peers								
Y_{rt12}^{NM}	4.03	0.35	1.00	1.00	0.05	0.22	0.73	0.86
Y_{rt22}^{NM}	4.03	0.38	1.04	0.97	0.06	0.24	0.70	0.83
Y_{rt32}^{NM}	3.94	0.37	0.97	0.96	0.06	0.22	0.72	0.79

Note. τ_{tik}^{RM} = true score variables of the reference method; $\tau_{rtik}^{\mathrm{NM}m}$ = true score variables of the nonreference methods; *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	11	12	13	14
1. T_{t11}^{RM}	*.18*													
2. T_{t21}^{RM}	.84**	*.18*												
3. T_{t31}^{RM}	.91**	.87**	*.20*											
4. T_{t12}^{RM}	.83**	.67**	.73**	*.21*										
5. T_{t22}^{RM}	.81**	.65**	.73**	.94**	*.19*									
6. T_{t32}^{RM}	.85**	.77**	.81**	.95**	.88**	*.18*								
7. $CM_{t1}^{\text{NM}_1}$	X	X	X	.02	.02	.01	*.14*							
8. $CM_{t1}^{\text{NM}_2}$	X	X	X	.05	.04*	.04*	.73**	*.10*						
9. $CM_{t2}^{\text{NM}_1}$	.08**	.05**	.06**	X	X	X	.95**	.69**	*.13*					
10. $CM_{t2}^{\text{NM}_2}$	.03	-.03	-.01	X	X	X	.68**	.94**	.70**	*.10*				
11. $UM_{rt1}^{\text{NM}_1}$	X	X	X	X	X	X	X	X	X	X	*.42*			
12. $UM_{rt1}^{\text{NM}_2}$	X	X	X	X	X	X	X	X	X	X	X	*.31*		
13. $UM_{rt2}^{\text{NM}_1}$	X	X	X	X	X	X	X	X	X	X	.93**	X	*.41*	
14. $UM_{rt2}^{\text{NM}_2}$	X	X	X	X	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_{tik}^{RM} = trait factors; $CM_{tk}^{\text{NM}_m}$ = common method factors; $UM_{rtk}^{\text{NM}_m}$ = 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 < .05, **p < .001

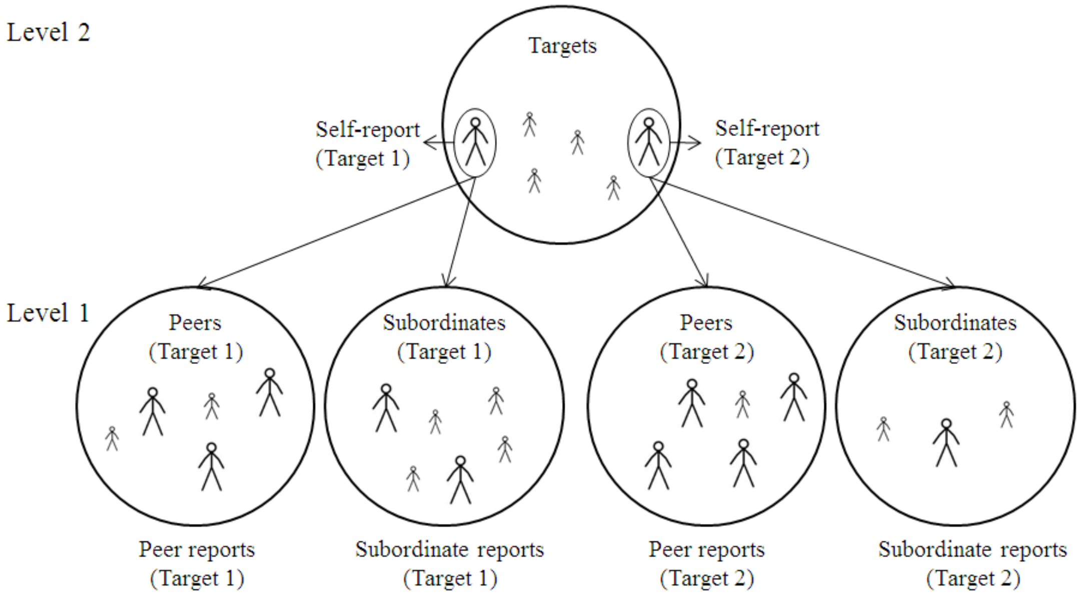

Figure 1. Sampling procedure for two sets of interchangeable methods (peers and subordinates) and one structurally different method (self-report).

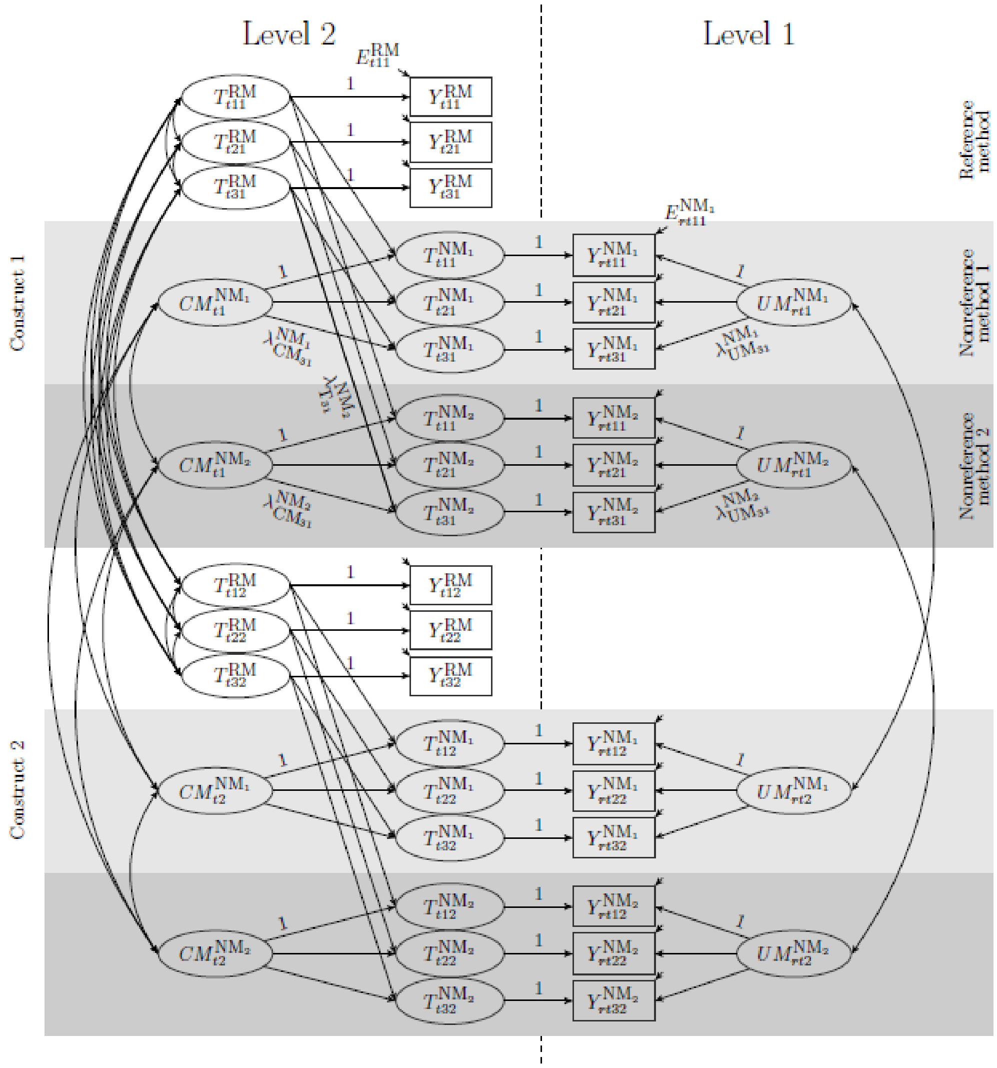

Figure 2. Multilevel confirmatory factor analysis multitrait-multimethod model for two sets of interchangeable methods and one structurally different method with indicator-specific trait variables. Y^{RM}_{tik}/ $Y^{\mathrm{NM}_m}_{rtik}$ = observed variables; T^{RM}_{tik}/ $T^{\mathrm{NM}_m}_{tik}$ = trait factors; $CM^{\mathrm{NM}_m}_{tk}$ = common method factors; $UM^{\mathrm{NM}_m}_{rtk}$ = unique method factors; E^{RM}_{tik}/ $E^{\mathrm{NM}_m}_{rtik}$ = error variables; $\lambda^{\mathrm{NM}_m}_{\mathrm{T}ik}$ = trait factor loadings; $\lambda^{\mathrm{NM}_m}_{\mathrm{CM}ik}$ = common method factor loadings; $\lambda^{\mathrm{NM}_m}_{\mathrm{UM}ik}$ = 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. Trait variables and common method factors belonging to different traits are allowed to correlate (not shown in the figure).

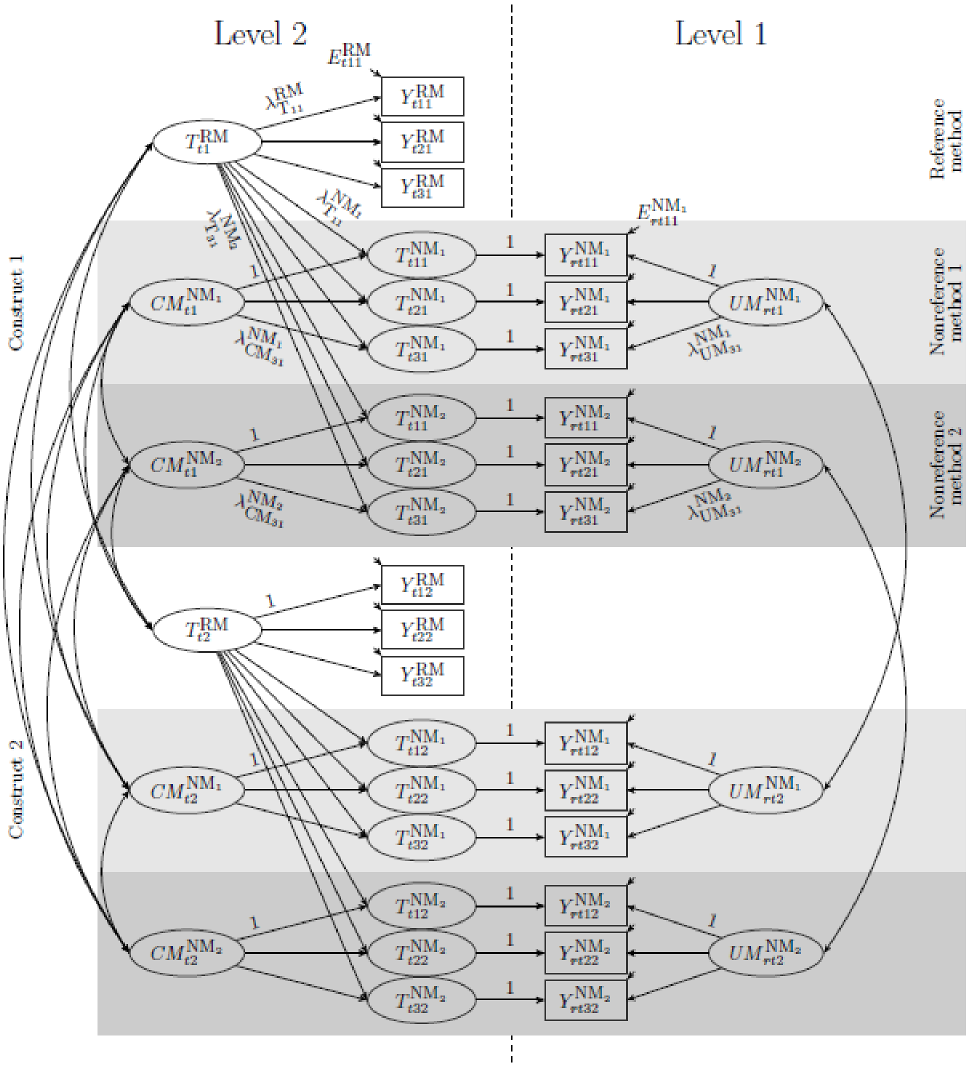

Figure 3. Multilevel confirmatory factor analysis multitrait-multimethod model for two sets of interchangeable methods and one structurally different method with general trait variables. Y_{tik}^{RM}/ $Y_{rtik}^{\text{NM}_m}$ = observed variables; T_{tk}^{RM}/ $T_{tik}^{\text{NM}_m}$ = trait factors; $CM_{tk}^{\text{NM}_m}$ = common method factors; $UM_{rtk}^{\text{NM}_m}$ = unique method factors; E_{tik}^{RM}/ $E_{rtik}^{\text{NM}_m}$ = error variables; $\lambda_{\text{T}ik}^{\text{RM}}$/$\lambda_{\text{T}ik}^{\text{NM}_m}$ = trait factor loadings; $\lambda_{\text{CM}ik}^{\text{NM}_m}$ = common method factor loadings; $\lambda_{\text{UM}ik}^{\text{NM}_m}$ = 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.

Chapter 3

Analysing Multisource Feedback With Multilevel Structural Equation Models: Pitfalls and Recommendations From a Simulation Study

This article is published:

Mahlke, J., Schultze, M., & Eid, M. (2019). Analysing multisource feedback with multilevel structural equation models – Pitfalls and recommendations from a simulation study. *British Journal of Mathematical and Statistical.* https://doi.org/10.1111/bmsp.12149

Chapter 4

Disagreement in Self-Other-Ratings on Leadership Competencies: Are Managers Lacking Self-Awareness?

This article is under review:

Mahlke, J., Schultze, M., Eckert, R., & Eid, M. (2018). *Disagreement in self-other-ratings on leadership competencies: Are managers lacking self-awareness?* Manuscript under review.

Abstract

When managers evaluate their own leadership performance and are evaluated by peers, subordinates, and supervisors in a 360-degree feedback process, the different rating perspectives usually do not completely agree in their perception. The degree of self-other-agreement is often considered as an indicator of the manager's self-awareness. However, research findings on the association between self-awareness and self-other-agreement are highly inconsistent and have not systematically shown the hypothesized relationship between the two constructs. We show that these inconsistencies may partly be explained by statistical issues. We use data from Benchmarks®, a widespread 360-degree feedback instrument, and define three different multilevel structural equation models to analyze the relationship between self-other-agreement in leadership competencies and the subscale of self-awareness. In all three models, self-ratings and ratings of peers, subordinates, and a boss are included while self-other-agreement is captured by three different approaches. An integrative discussion of the results reveals that the correlations between self-other-agreement and self-awareness are almost completely due to shared method variance. When this method variance is explicitly modeled, there remains no substantial correlation between the two constructs. We therefore advise against the common routine of interpreting disagreement in ratings as an indicator of lacking self-awareness.

Keywords: multisource feedback, self-other-agreement, self-awareness, method effects, multilevel structural equation modeling

Self-awareness has been a central psychological concept for decades. The classical theories of self-awareness (Carver & Scheier, 1982, 1998; Duval & Wicklund, 1972; Fenigstein, Scheier, & Buss, 1975; Snyder & Gangestad, 1986) share the definition that self-awareness comprises the process of making evaluations about oneself and the process of making evaluations about how one is seen by others. According to Baumeister (2005), self-awareness consists of “anticipating how others perceive you, evaluating yourself and your actions according to collective beliefs and values, and caring about how others evaluate you” (p. 7).

In the field of leadership research, self-awareness has been identified as one key element for manager performance and effectiveness (see Avolio & Gardner, 2005; Taylor, Wang, & Zhan, 2012). Despite the broad agreement on the definition and the importance of self-awareness, there is “no consensus among researchers in this field as how to best represent self-awareness conceptually or statistically” (Fletcher & Bailey, 2003, p. 397).

Morin (2011) gives an overview of self-awareness measurement tools, e.g. the Self-Consciousness-Scale (SCS; Fenigstein et al., 1975) and the Situational-Self-Awareness-Scale (Govern & Marsch, 2001) and 360-degree feedback instruments such as Benchmarks® (Center for Creative Leadership, 2010) and the Authentic Leadership Questionnaire (ALQ; Walumbwa, Avolio, Gardner, Wernsing, & Peterson, 2008) often include a self-awareness subscale. However, when it comes to operationalizing self-awareness in the context of leadership assessment, organizations usually fall back on using the agreement of self- and others’ ratings in a 360-degree feedback as a measure of the manager’s self-awareness (e.g., Atwater & Yammarino, 1992; Berson & Sosik, 2007; Bratton, Dodd, & Brown, 2011; Church, 1997; Fletcher & Baldry, 2000; Van Velsor, Taylor, & Leslie, 1993; Wohlers & London, 1989).

360-degree feedback instruments assess the target managers' leadership performance by considering multiple facets of leadership and multiple raters from different levels in the organization (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998). The managers deliver self-ratings and are typically evaluated by their peers, their subordinates, and their boss. Additional evaluations from customers or other external sources can (but do not have to) be included. The facets that are investigated strongly differ between instruments and are often customized for the company that wishes to assess leadership. They capture for example strategic and decisive behavior, communication, motivation and interpersonal skills, and the ability of involving peers or subordinates in work processes (e.g., Lombardo, McCauley, McDonald-Mann, & Leslie, 1999; Viswesvaran, Ones, & Schmidt, 1996).

In the feedback process of the 360-degree assessment, the agreement of self-ratings and ratings by peers, subordinates, or the boss is a crucial element. It is often found that the different raters do not agree in their appraisal of leadership competencies. But what does this mean? Is it an appropriate indicator of a manager's lacking self-awareness? The underlying assumption of studies using self-other-agreement (SOA) as an indicator of self-awareness is that managers who are highly self-aware will deliver a self-evaluation that is in line with ratings of others. This idea originates in the definition of self-awareness to be "the degree to which individuals understand their own strengths and weaknesses" (Wohlers & London, 1989, p. 236) and to stem from "the individual's ability to assess others' evaluations of the self and to incorporate those assessments into one's self-evaluation" (Atwater & Yammarino, 1992, p. 143).

Contrary to these conceptions that require self- and others' ratings to agree, there is a competing view that SOA is neither expectable nor desirable (Bozeman, 1997). As peers, subordinates, the boss, and the manager him- or herself have different perspectives and different opportunities to observe the manager's behavior their ratings will necessarily vary

(Murphy & Cleveland, 1995). These raters probably rate different aspects of the manager's performance, namely those, that are most relevant to the specific rater (Borman, 1991). Therefore, Taylor et al. (2012) criticize the popular routine of interpreting self-other-disagreement as a lack of the manager's self-awareness and point out that this discrepancy might as well be explained by differing perceptions of the different raters.

Actually, disagreement in ratings supports the idea of gathering information from multiple perspectives as it is done in 360-degree feedback assessment: If there was no expected gap between self- and others' ratings, collecting these multi-perspective data would offer redundant information and thus be a waste of time and money (Mersman & Donaldson, 2000).

Following this line of argument, it is of great interest to further analyze the factors that affect the degree of SOA instead of trying to reduce self-other-disagreement and many studies do so (e.g., Brutus, Fleenor, & McCauley, 1999; Gentry, Ekelund, Hannum, & de Jong, 2007; Ostroff, Atwater, & Feinberg, 2004; Vecchio & Anderson, 2009). In their review, Fleenor, Smither, Atwater, Braddy, and Sturm (2010) differentiate between factors affecting self-ratings, factors affecting others' ratings and factors affecting the congruence between self- and others' ratings. To name just a few of these factors there are biographical characteristics, personality and individual characteristics, contextual variables, as well as cognitive and motivational aspects. Within this framework of "expected disagreement", it is not plausible to use SOA as a measure of self-awareness. Instead, self-awareness could be one of many factors that affect SOA. These different conceptions of SOA and its linkage to self-awareness raise the question of the empirical relationship between the two constructs.

1. Empirical Relationship Between Self-Awareness and Self-Other-Agreement

When it is assumed that self-awareness can explain SOA the hypothesized direction usually is the following (Church, 1997; Mersman & Donaldson, 2000; Van Velsor et al., 1993): Those managers with a low degree of SOA have low levels of self-awareness. Those with high SOA also score high in self-awareness. This means that the highest degree of self-awareness is expected for those managers whose' self-ratings of their managerial competencies are in agreement with the ratings of their subordinates, peers, or boss. Contrariwise, managers with low levels of self-awareness should deviate from others' ratings in their evaluation of competencies. This deviation could take both directions: Either the manager overestimates his or her competencies or the manager underestimates his or her competencies compared to others' ratings.

Previous studies on self-awareness and SOA often detected unexpected patterns. When self-awareness is self-rated, the highest self-awareness was found for those managers who overrated themselves compared to peers (Wohlers & London, 1989) or subordinates (Van Velsor et al., 1993). Managers who underrated themselves compared to peers (Wohlers & London, 1989) or subordinates (Van Velsor et al., 1993) had the lowest self-awareness. Conversely, when self-awareness is rated by subordinates, the highest self-awareness was found for those managers who underrated themselves compared to subordinates (Van Velsor et al., 1993). Van Velsor et al. (1993) came to the conclusion that SOA is an adequate indicator of self-awareness rated by others in the group of overraters because in this group SOA as well as self-awareness were low. However, in the group of underraters SOA and self-awareness are not the same phenomenon because SOA was low while self-awareness rated by others was high.

These results only partially support the hypothesis of higher agreement for those managers with higher self-awareness: When using the self-rated self-awareness, the hypothesized relationship to SOA was found for underraters: They were not in agreement and rated themselves as having low self-awareness. When using self-awareness rated by others, the hypothesized relationship to SOA was found for overraters: They were not in agreement and were rated as having low self-awareness. Thus, in these specific cases it seems admissible to use self-other-(dis)agreement as a measure of self-awareness. For all other situations the results on self-awareness and SOA were inconsistent with the hypotheses and SOA was not a useful indicator of self-awareness.

Other authors investigated self-monitoring instead of self-awareness and its relationship to SOA in leadership assessment. Just as self-awareness, self-monitoring (Snyder, 1974) is supposed to be higher in individuals whose self-ratings are in agreement with others’ ratings because the construct is defined as the extent to which individuals monitor, regulate, and control their behaviors vis-à-vis others. Therefore, Church (1997) and Mersman and Donaldson (2000) proposed that finding the hypothesized relationship between self-monitoring and SOA supports the theory of SOA being an indicator of self-awareness. The hypothesis was corroborated by a negative correlation between self-rated self-monitoring and the absolute difference variable of leadership behavior ratings (Church, 1997). That means, the lower the self-monitoring, the higher the difference between self- and direct report ratings, regardless of direction. The author also correlated self-monitoring and a difference score that considers the direction of disagreement. This correlation was not significant, so there was no linear relationship between self-monitoring and the difference between self- and direct report ratings, when this difference variable reflects the direction of disagreement. Finding such a correlation would not confirm the hypothesized relationship as it would indicate that the self-awareness rises the higher the overestimation of managers (in the case of a positive correlation) or that the self-awareness rises the higher the underestimation of managers (in the case of a negative correlation). The results of Church (1997) therefore supported the hypothesis of lower self-awareness for those managers who have less agreement with others (regardless of the direction of disagreement) and it was postulated by the author that assessing SOA is an adequate method of operationalizing self-awareness. However, the result is not in line with the study by Wohlers and London (1989) who found a strong linear relationship between self-rated self-awareness (measured by one item) and a difference variable of managerial characteristics that considered the direction of disagreement in the way that those managers with the highest self-awareness were those who overrated themselves the most compared to their peers. Mersman and Donaldson (2000) analyzed differences in self-rated self-monitoring between the three groups of overraters, in-agreement raters, and underraters based on ratings of the manager’s task performance. Contrary to their hypotheses they found no mean differences between the groups.

Taken together, these findings are highly inconsistent and do not satisfactorily answer the question on the relationship between the two constructs of SOA and self-awareness. One possible reason for the inconsistency of studies are statistical issues.

2. Statistical Issues

2.1 Measures of Self-Other-Agreement

One main issue concerns the question how to depict SOA. In many approaches, the first step is to calculate the difference between the self-rating of the given item or scale and the others’ ratings. In cases where the others’ ratings stem from multiple peers or multiple subordinates, their ratings are averaged before calculating the difference. In cases where the boss rating is used, no averaging is necessary.

A widespread strategy that was often applied to depict SOA is to categorize the difference between self- and others' ratings to construct classes of agreement (e.g., Atwater & Yammarino, 1992; Fleenor, McCauley, & Brutus, 1996; Mersman & Donaldson, 2000; Van Velsor et al., 1993; Yammarino & Atwater, 1997). Based on the mean difference between self-ratings and others' ratings, managers are assigned to one of typically three, four, or even more classes that take into account the nature as well as the degree of agreement (e.g., overraters, underraters, in-agreement raters). Authors who analyzed the relationship between self-awareness (or self-monitoring) and SOA based on these categories, calculated the average self-awareness (or self-monitoring) in every category of agreement and tested differences via an ANOVA (Mersman & Donaldson, 2000; Van Velsor et al., 1993). This categorization is intuitive and its advocates claim it to be superior to other methods discussed in the literature for practical utility reasons (Mersman & Donaldson, 2000). However, artificial categorization of a continuous variable leads to a substantive loss of variance and can yield misleading results (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002). Consequently, most researchers switched to other modeling techniques (Fleenor et al., 2010).

For a long period of time the probable most common index of congruence was the difference score (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Edwards, 2002) which can be used algebraic, absolute, or squared (Edwards, 1994). While the algebraic difference (i.e., the difference value with its plus or minus sign) should be applied when the direction of disagreement is important, the absolute and squared differences are appropriate when one attempts to depict the degree of disagreement but wants to disregard the direction of disagreement. Authors who used a continuous difference variable and analyzed its relationship to self-awareness (or self-monitoring) calculated the Pearson correlation (Church, 1997; Wohlers & London, 1989). Algebraic difference scores that consider the direction of disagreement (Church, 1997; Wohlers & London, 1989) as well as squared or absolute difference scores (Church, 1997) have been used.

Despite its plausibility at first sight, difference scores were criticized for several statistical reasons (Edwards, 1993; 1994; 2002), e.g. the contributions of the single components in their effect on a third variable are confounded as the three-dimensional relationship between the two components and a third variable is reduced to a two-dimensional scenario and the difference scores potentially have lower reliability than the original two components (Cronbach & Furby, 1970).

Therefore, the use of polynomial regression when SOA is investigated as an independent variable (Edwards, 1993; 1994; 2002) and multivariate regression when SOA is investigated as a dependent variable (Edwards, 1995) is recommended. Both techniques keep the two components (self-ratings and others' ratings) separate instead of collapsing them into a single index and model them simultaneously.

Another strategy to circumvent the problem of reduced reliability of difference scores is to model the difference as a latent variable in structural equation modeling (SEM). In this framework, multiple manifest indicators of the difference measure a latent difference variable that is free of measurement error. When managers deliver self-ratings and are rated by multiple peers and subordinates, it is recommended to use multilevel models to take into account the hierarchical data structure (Putka, Lance, Le, & McCloy, 2011). Combining these two modeling approaches leads to multilevel SEM that is a powerful tool in the analysis of complex data structures. In organizational research, multilevel SEM has so far only occasionally been used and in these applications the between-level (level 2) of the data usually consists of different countries or societies (e.g., Dyer, Hanges, & Hall, 2005; House, Hanges, Javidan, Dorfman, & Gupta, 2004) or teams (e.g., Seibert, Silver, & Randolph, 2004). However, Mahlke et al. (2016) recently demonstrated how multilevel SEM can be used when managers are level-2 clusters and peers and subordinates are nested within their respective manager and located on level 1.

2.2 Method Effects

One possible statistical explanation for results that are inconsistent with the hypothesis of congruence between SOA and self-awareness are method effects. It is known that each rating is influenced by at least two sources: One source is the trait itself and the other source is the method used to measure the trait (Campbell & Fiske, 1959; Eid, Geiser, & Koch, 2016). Thus, ratings that stem from the same method share a relevant amount of variance attributable to that method. Consequently, the correlation between two traits measured by the same method cannot fully be explained by overlap of the two traits but some part of shared variance between the two measures will be caused by the specific method. In the context of 360-degree feedback each rater is a different method. The correlation between SOA and self-awareness therefore depicts not only commonalities of the two constructs but also method effects. When self-awareness is self-rated, it is the manager who delivers two ratings that share method variance: the self-awareness rating and the self-rating of the leadership competencies that is part of the SOA. When the self-awareness of the manager is rated by the boss, peers, or subordinates, it is the respective method of others that delivers these two ratings that share method variance: the others' rating of self-awareness and the others' rating of leadership competencies that is part of the self-other agreement.

Previous studies of multisource feedback data have shown strong method effects (e.g., Mahlke et al., 2016; Woehr, Sheehan, & Bennett, 2005). In 360-degree feedback assessments there are two different kinds of method effects: idiosyncratic effects and rater source effects (Mount et al., 1998; Scullen, Mount, & Goff, 2000). Idiosyncratic method effects are also referred to as the halo effect (Thorndike, 1920) or rater leniency/elevation (Saal, Downey, & Lahey, 1980). They all describe the variance associated with the individual rater. Rater source effects refer to the variance that is shared among raters stemming from the same source, e.g., among peers or among subordinates, due to having the same organizational perspective. Modern data analytic techniques allow differentiating between idiosyncratic effects and rater source effects. Idiosyncratic effects are by far the major source of variance in others' ratings accounting for 56% to 70% in the study by Hoffman, Lance, Bynum, and Gentry (2010) and for 64% to 73% in the study by Mahlke et al. (2016). Rater source effects are smaller but still substantial with a variance proportion between 17% to 21 % (Hoffman et al., 2010) or 22% to 26% (Mahlke et al., 2016).

3. Research Questions

As discussed above the results of previous studies on the relationship between self-awareness and SOA on leadership competencies are mostly inconsistent with each other and with what was hypothesized. Given the statistical limitations of the methods that were used, i.e., artificially categorizing a continuous variable and/or not considering measurement error and method effects, we will investigate the relationship between SOA and self-awareness with statistical methods that overcome these issues.

This relationship might depend on the specific leadership competency that is investigated. The diverse facets of leadership might differ in their difficulty to rate (Wohlers & London, 1989) and therefore also in SOA and in the strength of method effects. However, even though most instruments of leadership assessment are constructed based on the assumption of multiple distinct and only partially related dimensions, plenty of research has questioned this assumption and has found low discriminant validity of the subscales (Beehr, Ivanitskaya, Hansen, Erofeev, & Gudanowski, 2001; Hoffman et al., 2010; Kets de Vries, Vrignaud, & Florent-Treacy, 2004; Mahlke et al., 2016; van der Zee, Zaal, & Piekstra, 2003). Furthermore, researchers that used SOA as an indicator of self-awareness usually calculated this SOA based on all ratings of the instrument used (instead of considering subscales). We therefore decided to not investigate specific dimensions or subscales of leadership but to build an overall index of leadership competency by using all available items.

We will analyze the following research questions:

(1) Is there a linear correlation between self-awareness and the absolute value of SOA on leadership competencies after controlling for measurement error? This is the relationship that is usually hypothesized as it indicates that those with low self-awareness are in disagreement with others, regardless of the direction (Church, 1997; Mersman & Donaldson, 2000; Van Velsor et al., 1993).

(2) Is there a linear correlation between self-awareness and SOA when the direction of disagreement is considered and after controlling for measurement error? This relationship was often found in previous studies even though there is no reasonable explanation for such a finding (Van Velsor et al., 1993; Wohlers & London, 1989).

(3) Is there a linear correlation between self-awareness and the method factor of others' ratings on leadership competencies after controlling for measurement error and the method effect of the self-ratings? In other words: Can the view of others that diverges from the manager's self-perception be explained by (lack of) self-awareness? This analysis keeps the two components, i.e., self-ratings and others' ratings, separate and allows investigating whether self-awareness is an adequate indicator of SOA when method effects and measurement error are controlled for.

We will include all others' ratings that are usually collected in a 360-degree feedback setting: self-ratings, boss ratings, peer ratings, and subordinate ratings. Previous studies often focused on one or at most two of these perspectives, so it is not yet possible to draw a complete picture.

In the models that we use to analyze the three research questions we will consider the hierarchical data structure and the measurement error by using multilevel SEM. We will not artificially construct classes of agreement because this would result in a substantial loss of variance. In the first two models, we will use absolute (research question 1) and algebraic (research question 2) difference scores in order to compare the results with existing literature but taking into account measurement error and including all rating perspectives. In the model referring to research question 3, we will follow the recommendation of authors who advise keeping the two components of self- and others' ratings separate (Cronbach & Furby, 1970; Edwards, 1993; 1994; 2002) and define a multilevel SEM model with method effects. All of the results will be interpreted with regard to the question whether SOA on leadership competencies is an adequate indicator of self-awareness.

4. Method

4.1 Sample and Measures

We used 360-degree feedback data that was collected by the Center for Creative Leadership in companies worldwide. The sample consists of 210,369 raters from at least 117 different countries (6,894 participants with missing data on their country) of which United States (*n* = 90,274), United Kingdom (*n* = 13,658), and Canada (*n* = 27,448) contributed the largest sample sizes. Approximately 31% of the participants were female (1,185 participants with missing data on gender). The data contain self-ratings of 20,158 managers and others' ratings from 20,158 bosses, 83,347 peers, and 86,704 subordinates. Every manager received others' ratings of one boss and on average 4.13 peers (range: 3-24) and 4.30 subordinates (range: 3-42).

All participants completed Benchmarks® (Lombardo, McCauley, McDonald-Mann, & Leslie, 1999), a well validated and widespread 360-degree feedback instrument (e.g., McCauley, Lombardo, & Usher, 1989; for an overview of studies using Benchmarks® see Leslie & Peterson, 2011) which includes 16 competency and five derailment scales. One of the competency scales measures self-awareness with four items (e.g., "Sorts out his/her strengths and weaknesses fairly accurately"). The remaining 15 competency scales contain 90 items that were used in the analyses. The derailment scales were excluded in the current study. All items had a 5-point scale ranging from 1 ("to a very little extent") to 5 ("to a very great extent").

4.2 Statistical Analyses

In order to include the complete range of leadership competencies of Benchmarks® while reducing the number of variables in our analyses, we used the 90 items of the competency scales to build three parcels of leadership competencies, containing 30 items each. Our strategy for item parceling was to derive equally balanced indicators (that is, item-to-construct balance) based on the factor loadings of the self-report (Little, Cunningham, Shahar, & Widaman, 2002). No parceling was necessary for self-awareness because this construct was measured by only four items.

We defined a series of multilevel SEM models with Mplus 8 (L. K. Muthén & Muthén, 1998-2017). The appendix includes all input files. In all models, the clustering of target managers in 117 different countries (plus one category for all managers with missing information for country) was taken into account by defining the country to be a cluster variable. Consequently, the software used robust maximum likelihood estimation (MLR) with a sandwich estimator for standard errors (Yuan & Bentler, 2000).

All models include self-ratings, boss ratings, and the individual peer and subordinate ratings. These data have a multilevel structure because peers and subordinates are nested within their respective manager. Therefore, managers are level-2 clusters (just as companies or countries in other multilevel analyses) and peers and subordinates are level-1 units in these clusters. However, unlike traditional multilevel settings, peers and subordinates define two distinct level-1 populations. They do not belong to the same group of raters but to two different ones and these two populations should be kept separate (instead of being collapsed) in a multilevel model. Mahlke et al. (2016) recently presented a multilevel model that is able to take this special structure of multisource feedback data into account. We used the technique explained in this paper in order to incorporate the two distinct groups of peers and subordinates as being nested within their respective target manager. Every target manager has one boss (instead of a group of bosses). Thus, boss ratings are not clustered within the targets and are not located on level 1 but are level-2 variables. Within this framework, we set up three different models to analyze the three research questions formulated above. All of the three models have in common:

- a multilevel structure with self-ratings and boss ratings measured on level 2 and peer ratings and subordinates ratings measured on level 1;
- a structural equation framework in which latent factors for the constructs under research are defined in order to separate true factor scores from measurement error.

4.2.1 Model 1 and model 2.

These models analyze the correlation between SOA and self-awareness (Figure 1). In model 1, the factors depicting SOA are measured by three indicators. Each of them is computed by first subtracting the others’ rating on the leadership competencies parcels from the self-rating on the leadership competencies parcels and then using the absolute value of this difference. Thus, each indicator is a manifest absolute difference variable. This is done for each others’ perspective individually and results in three factors depicting the absolute differences between self-ratings and boss ratings (level-2 factor), self-ratings and individual peer ratings (level-1 factor), and self-ratings and individual subordinate ratings (level-1 factor). Moreover, the differences between self-ratings and peer and subordinate ratings are also captured on level 2. The level-2 indicators represent the difference between self-ratings and the common part of all peers or subordinates belonging to the same target manager. Their values can be conceptualized as the difference between the self-ratings and the "average" peer or "average" subordinate ratings of the given target manager. These indicators build two level-2 factors: One factor depicts the absolute difference between the managers' self-evaluation and their "average" peer ratings on leadership competencies. The other factor depicts this difference with respect to the "average" subordinate ratings.

Model fit analyses revealed that the assumption of SOA being measured unidimensionally by the three indicators on level 2 does not hold. Therefore, we defined two additional level-2 factors that capture indicator-specific effects: One factor captures the specific effects of the second indicator and the other factor captures the specific effects of the third indicator of self-boss-agreement, of self-peer-agreement and of self-subordinate-agreement. No factor is defined for the specific effects of the first indicator of agreement because this indicator is chosen as the reference (Eid, 2000).

The second construct, self-awareness, is measured by four items and all rating perspectives are included in the model – self-ratings and others' ratings of the managers' self-awareness. This results in the following factors depicting self-awareness: On level 1, there is a factor of the individual peer ratings and a factor of the individual subordinate ratings. On level 2, there is a factor of the self-ratings, a factor of the boss ratings, and two factors for the common or "average" peer and subordinate ratings of self-awareness. In this model, we will analyze Pearson correlations between the factors of absolute SOA and self-awareness (self-rated and rated by others) in order to answer research question 1.

Model 2 is defined in the same manner with the only modification that the direction of disagreement is considered by using not the absolute value of the indicators of SOA, but the algebraic difference. We will analyze correlations between SOA and self-awareness to answer research question 2.

4.2.2 Model 3.

In model 3 (see Figure 2), we use a different operationalization of SOA. We do not collapse self- and others' ratings into a difference variable but keep all rating perspectives separate. Hence, leadership competencies are depicted as follows: On level 2, there is a factor of self-rated leadership competencies measured by the three self-report parcels. In order to depict SOA, all indicators of others' ratings (that is, the boss rating parcels, the peer rating parcels, and the subordinate rating parcels) are regressed on the factor of self-rated leadership competencies. Some part of variance in boss, peer, and subordinate ratings is shared with the manager's self-report and therefore can be explained by the self-ratings. This is the part of variance that reflects SOA. Some other part of variance in others' ratings is not shared with the manager's self-ratings. It therefore reflects the disagreement between self- and others' ratings. This variance is captured by method factors of others' ratings. These factors represent method-specific variance, that is, variance in ratings that is specific to the respective rating perspective of the boss, the peers, or the subordinates. On level 2, there are method factors of the boss and of the common or "average" peers and subordinates. The latter two refer to variance of the peer and subordinate ratings that is not shared with the manager's self-report but with the other peers or subordinates. This variance is therefore common to the peers or subordinates of a given manager but specific to that rating perspective. These factors are thus called *common method factors*. On level 1, the individual peer and individual subordinate ratings are captured by two additional method factors, the *unique method factors*. These level-1 method factors assess the variance in individual peer and subordinate ratings that is specific to the unique peer or subordinate. It is neither shared with the other peers or subordinates nor with the manager him- or herself but depicts the unique view of a peer or subordinate. Self-awareness is measured as in model 1 and model 2. In model 3, we will analyze the correlations between the level-2 method factors and the self-awareness (self-rated and rated by others). These correlations indicate whether the variance that is not shared among the manager and the others (i.e., the disagreement) is related to self-awareness.

4.3 Evaluation of Model Fit

A simulation study (Mahlke, Schultze, & Eid, in press) of the model with two distinct level-1 populations has shown that the χ^2-test of model fit is too lenient and should not be trusted for this type of model. As most other fit indices provided by Mplus are χ^2-based [Root Mean Square Error of Approximation (RMSEA), Tucker-Lewis-Index (TLI) and the Comparative Fit Index (CFI), West, Taylor, & Wu, 2012] they will also be biased. We will report the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA) but interpret them with caution. Based on the results of the simulation study (Mahlke et al., in press) we suppose that these indices have a leniency bias. Therefore, they are not appropriate to claim model fit. However, if their values indicate misfit, the postulated model should be rejected.

Mahlke et al. (in press) suggested using the Standardized Root Mean Square Residual (SRMR) in order to evaluate model fit because it is not χ^2-based. The SRMR is provided for both levels of measurement separately and values below .08 are considered to indicate model fit (Bentler, 1995). For level 1, Mahlke et al. (in press) have demonstrated that the degrees of freedom assumed by Mplus, and consequently also the SRMR_{L1}, are not correct in the specific model with two level-1 populations. The authors provide a formula to calculate an adjusted SRMR_{L1}:

$$\text{SRMR}_{\text{L1}}^{\text{adj.}} = \sqrt{\frac{p(p+1)}{p(p+1)-p_{NM_1}p_{NM_2}}\text{SRMR}_{\text{L1}}^2}$$

We will therefore analyze model fit primarily based on the adjusted $SRMR_{L1}$ and $SRMR_{L2}$.

5. Results

All three models fit the data very well (Table 1). The means and standard deviations of all indicators are printed in Table 2. In order to visualize the relationship between SOA and self-awareness in models 1 and 2 and the relationship of method factors of leadership competencies and self-awareness in model 3, scatterplots of the factor scores are shown in Figure 3. The scatterplots indicate linear relationships that we will analyze in the following section. Tables 3 and 4 display the factor correlations between SOA and self-awareness in models 1 and 2. Table 5 refers to correlations between the method factors of leadership competencies and self-awareness in model 3.

5.1 Self-Other-Agreement as Absolute Difference and Self-Awareness (Model 1)

The mean absolute deviation between self-ratings and others' ratings varied between 0.43 and 0.53 units. The lowest means (from 0.43 to 0.44) were found for SOA between the managers and their boss. The highest absolute deviation (from 0.52 to 0.53) occurred between managers and their subordinates. Thus, the managers have the highest absolute agreement with their boss and the lowest absolute agreement with their subordinates.

5.1.1 Self-awareness self-rated by the managers.

In order to interpret the following correlations note that a value of zero on absolute SOA indicates perfect agreement whereas disagreement becomes stronger the higher the absolute SOA value. The correlations between the absolute value of SOA and self-rated self-awareness in model 1 were zero to weakly negative, varying between r = -.00 and r = -.17. The strongest correlation of .-17 was found between SOA of managers and subordinates and self-awareness. The correlation between SOA of managers and their boss and self-awareness amounted to -.06. There was a zero correlation between SOA of managers and peers and self-awareness. These results indicate that there was either no association or at the maximum a slight one with higher (self-rated) self-awareness being associated with less deviations between self and others' ratings in leadership competencies.

5.1.2 Self-awareness rated by others.

Next, we analyzed the correlations between the managers' self-awareness rated by the boss, peers, or subordinates and SOA of the manager and the respective others' perspective. The correlation between SOA of managers and their boss and self-awareness rated by the boss was nearly zero (*r* = -.02). SOA of managers and their peers and self-awareness rated by the peers had a correlation of -.21. For subordinates the correlation between SOA and self-awareness was -.19. The pattern is comparable to the one found for self-rated self-awareness: Self-awareness (rated by others) was either not associated to SOA or slightly associated with higher agreement on leadership competencies for those with higher self-awareness.

5.2 Self-Other-Agreement as Algebraic Difference and Self-Awareness (Model 2)

The means in Table 2 indicate that the managers' self-ratings of leadership competencies were on average 0.01 to 0.02 units higher than their peer ratings but 0.06 to 0.07 units lower than their subordinate ratings and 0.05 to 0.07 units lower than their boss ratings. Thus, managers slightly overestimated their skills compared to their peers and slightly underestimated their skills compared to their boss and their peers. These effects are minimal though, and show that on average the managers' self-ratings and the others' ratings are highly in agreement. However, there is some variance in this agreement (from *SD* = 0.55 to *SD* = 0.67).

5.2.1 Self-awareness self-rated by the managers.

When interpreting the correlations in model 2, consider that a positive value on algebraic SOA indicates an overestimation of the manager, whereas a negative value indicates an underestimation of the manager compared to others' ratings. There were moderate to strong positive correlations between algebraic SOA and self-rated self-awareness. The strongest correlation of .63 was found between SOA of managers and their peers and self-awareness. SOA of managers and subordinates and self-awareness were also strongly correlated (*r* = .59), and SOA of managers and their boss had a correlation of .48 with self-awareness. In summary, the higher the self-rated self-awareness, the higher the overestimation of managers compared to ratings of others.

5.2.2 Self-awareness rated by others.

When analyzing self-awareness rated by others (instead of self-rated) and its relationship to SOA, the pattern was inverted as all correlations were strongly negative (instead of positive). The strongest correlation of -.67 was the one between SOA of managers and their boss and the boss rating in self-awareness. The remaining two others' perspectives had a correlation of -.55 (SOA of managers and subordinates and subordinate ratings in self-awareness) and -.50 (SOA of managers and peers and peer ratings in self-awareness). Thus, the higher the self-awareness rated by others, the higher the others' ratings of leadership competencies compared to self-ratings.

5.3 Method Factors of Leadership Competencies and Self-Awareness (Model 3)

The mean ratings of leadership competencies were high, varying between 3.89 and 4.01. As discussed in the section of model 2, the differences in mean ratings of the managers, their boss, their peers, and their subordinates were very small.

5.3.1 Self-awareness self-rated by the managers.

The correlations between method factors of the boss, of peers, and of subordinates for leadership competencies and self-rated self-awareness were all nearly zero (ranging from *r* = -.06 to *r* = -.04). Thus, when the variance in others' ratings on leadership competencies was reduced to the part that was not shared with the self-ratings, the amount of this method specific component did not covary with self-awareness. This means that the deviations of others' ratings from the manager's self-ratings in leadership competencies were not associated to the manager's self-rated self-awareness.

5.3.2 Self-awareness rated by others.

Method factors were strongly correlated to the respective others' rating on self-awareness. There was a correlation of .85 for the method factor of the boss ratings on leadership competencies and self-awareness rated by the boss. The method factor of peers and self-awareness rated by peers as well as the method factor of subordinates and self-awareness rated by subordinates had a correlation of .92 each. This finding indicates that the variance in others' ratings that cannot be explained by the self-ratings of leadership competencies is strongly associated to the others' ratings of self-awareness. The higher the method-specific ratings of a manager's competencies by others after controlling for his or her self-ratings, the higher the others' ratings on the manager's self-awareness.

6. Discussion

The relationship between SOA and self-awareness that is usually expected, that is, higher self-awareness for those managers who are in agreement with others (Church, 1997; Mersman & Donaldson, 2000; Van Velsor et al., 1993), is not found. No or only slight correlations between the absolute value of SOA and self-awareness were found. Based on our data, one cannot state that those managers with high self-awareness achieve higher SOA than managers with low self-awareness. This result is in line with Mersman and Donaldson (2000) who found no mean differences of self-monitoring between overraters, in-agreement raters, and underraters. It is however contradictory to the finding of Church (1997).

There were moderate to strong associations between SOA and self-awareness when considering the direction of disagreement by using the algebraic difference between self-ratings and others' ratings as a measure of SOA. This correlation took positive or negative values depending on whether self-awareness was self-rated or rated by others. When the self-ratings of the managers were used to assess self-awareness, we found a positive correlation to SOA. When others' ratings of the managers' self-awareness were used, the direction of the relationship to SOA was inverted. This means that the higher the manager's self-rating of self-awareness, the higher the manager's self-evaluation concerning his or her leadership skills compared to others' ratings. Contrariwise, the higher the others' ratings on the manager's self-awareness the lower the manager's self-evaluation of his or her leadership skills compared to others' ratings. We found the same pattern of correlations for all rating perspectives of others (bosses, peers, and subordinates) with only slight differences in the magnitude of the effects. While this result disagrees with the findings of Church (2000), it replicates the one of Wohlers and London (1989) concerning self-rated self-awareness and the one of Van Velsor et al. (1993) concerning self- and others' rated self-awareness.

The results of previous studies on the relationship between SOA and self-awareness are partly contradictory to each other and to the respective hypotheses. It is difficult to systematically integrate and compare existing studies for a variety of reasons: Usually only one rating perspective of others was included (e.g., peers in Mersman & Donaldson, 2000, and in Wohlers & London, 1989; subordinates in Van Velsor et al., 1993, and in Church, 1997) and some studies considered only self-ratings of self-awareness (Church, 1997; Mersman & Donaldson, 2000), while others considered also others' ratings of self-awareness

(Van Velsor et al., 1993; Wohlers & London, 1989). In many studies, categories of agreement were built instead of using a continuous indicator of agreement (e.g., Mersman & Donalson, 2000; Van Velsor et al., 1993) and to our knowledge, there is no study that considers measurement error. By including all rating perspectives, by using the algebraic as well as the absolute difference and by considering measurement error we revealed the systematic pattern of associations: The correlation between SOA and self-awareness depends on a) whether the algebraic or absolute difference is interpreted, and b) whether a self-rated or others' rated self-awareness measure is used.

Model 3, which depicts the two components of self- and others' ratings separately, helps to identify the explanation for the revealed pattern: method effects. When using only the method-specific variance in others' ratings of leadership competencies after controlling for the self-ratings there is no substantial correlation left between these method-specific components and self-rated self-awareness. Yet, these method-specific components have strong correlations to the respective others' rating of self-awareness. For example, the method-specific variance in boss ratings of leadership competencies that is not shared with the self-ratings has a high correlation with self-awareness rated by the boss. This pattern also holds for peers and subordinates. Taken together, these results indicate that all correlations found in the analyses are mainly due to shared method variance. When using the same method to assess leadership competencies and self-awareness the two constructs are highly correlated. When controlling for method effects, the shared covariance diminishes to a negligible amount.

In the traditional assessment of SOA, that is, when a (algebraic or absolute) difference score is built, these method effects arise as follows: The SOA component incorporates two methods, one being the self-rating and the other being one particular others' rating, e.g. the boss rating. When self-awareness is measured by a self-report, the correlation of self-awareness and SOA depicts not only shared variance of the two constructs, but also shared variance due to measuring the two constructs by the same method, that is, the self-report. When self-awareness is measured by the boss, the same is true but the method that leads to shared variance is the boss report. Thus, a manager who evaluates him- or herself highly positive in leadership skills has the tendency to also evaluate him- or herself highly positive in self-awareness. This explains the positive correlations in model 2. The same phenomenon appears for self-awareness rated by others. A boss, a peer, or a subordinate who evaluates a manager as highly skilled concerning his or her leadership behavior will probably also deliver a high rating of the manager’s self-awareness. These method effects explain the positive correlations of self-rated self-awareness and SOA and the negative correlations of others’ rated self-awareness and SOA in model 2. They also explain why there are only very small or even zero correlations between self-awareness and the absolute SOA in model 1. When the absolute value of SOA is used, the direction of agreement is eliminated. Consequently, the systematic tendency of each method to deliver either high or low ratings for both constructs (leadership skills and self-awareness) cannot result in a strong correlational effect. Instead, these small or none-existing correlations indicate that the supposed relationship between SOA and self-awareness is not found. Those managers who are in higher agreement with others are not more self-aware. Consequently, the degree of SOA should not be interpreted as an indicator of a manager’s self-awareness.

We therefore strongly advise against the popular routine of ascribing a lack of self-awareness to a manager whose self-ratings are not in agreement with the ratings of others in a 360-degree assessment. Instead, the differing ratings should be interpreted as resulting from different rating perspectives in the company and their additional benefit should be appreciated. If the manager’s self-awareness is of focal interest, then one should use a well validated instrument that was designed in order to assess self-awareness or a related construct (for an overview see Morin, 2011).

6.1 Limitations and Further Research

There are some limitations arising from the measures we used to assess self-awareness and SOA in leadership competencies. Self-awareness was measured by a subscale from Benchmarks®. It therefore consists of only four items and was developed to assess one subdomain of leadership behavior. The results on the relationship of self-awareness and SOA might be different if a traditional instrument of self-awareness (e.g., the Self-Consciousness-Scale [SCS]; Fenigstein et al., 1975) was used.

Additionally, it might be interesting to consider subscales of leadership behavior. In our study, we collapsed all subscales (except the one of self-awareness) and built an overall index of leadership behavior. This seemed legitimate and plausible to us, especially because previous studies have shown strong correlations between the subscales of leadership competencies (e.g., Beehr, Ivanitskaya, Hansen, Erofeev, & Gudanowski, 2001; Hoffman et al., 2010; Kets de Vries, Vrignaud, & Florent-Treacy, 2004; Mahlke et al., 2016; van der Zee, Zaal, & Piekstra, 2003), indicating that the manager him- or herself and the raters perceive the manager's behavior in a holistic fashion. However, there might be differential correlational relationships for self-awareness and leadership competencies depending on the specific leadership skill.

References

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology*, *45*, 141–164. doi:10.1111/j.1744-6570.1992.tb00848.x

Atwater, L. E., 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

Avolio, B. J., & Gardner, W. L. (2005). Authentic leadership development: Getting to the root of positive forms of leadership. *The Leadership Quarterly*, *16,* 315–338. doi:10.1016/j.leaqua.2005.03.001

Baumeister, R. F. (2005). *The cultural animal: Human nature, meaning, and social life*. New York, NY: Oxford University Press.

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

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Berson, Y., & Sosik, J. J. (2007). The relationship between self-other rating agreement and influence tactics and organizational processes. *Group and Organization Management*, *32,* 675–698. doi:10.1177/1059601106288068

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Bozeman, D. P. (1997). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior, 18,* 313–316. doi:10.1002/(SICI)1099-1379(199707)18:4<313::AID-JOB844>3.0.CO;2-O

Bratton, V. K., Dodd, N. G., & Brown, F. W. (2011). The impact of emotional intelligence on accuracy of self-awareness and leadership performance. *Leadership & Organization Development Journal, 32,* 127–149. doi:10.1108/01437731111112971

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development, 18,* 417–435. doi:10.1108/02621719910273569

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

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin, 92,* 111–135. doi:10.1037/0033-2909.92.1.111

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139174794

Church, A. H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 82,* 281–292. doi:10.1037/0021-9010.82.2.281

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74,* 68–80. doi:10.1037/h0029382

Duval, T. S., & Wicklund, R. A. (1972). *A theory of objective self-awareness*. New York, NY: Academic Press.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*, 149–167. doi:10.1016/j.leaqua.2004.09.009

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology*, *46*, 641–665. doi:10.1111/j.1744-6570.1993.tb00889.x

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, *58*, 51–100. doi:10.1006/obhd.1994.1029

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, *64*, 307–324. doi:10.1006/obhd.1995.1108

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow & N. W. Schmitt (Eds.), *Advances in measurement and data analysis* (pp. 350–400). San Francisco, CA: Jossey-Bass.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psycho-metrika*, *65*, 241–261. doi:10.1007/bf02294377

Eid, M., Geiser, C., & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science, 25*, 275–280. doi:10.1177/0963721416649624

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43,* 522–527. doi:10.1037/h0076760

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *The Leadership Quarterly, 7,* 487–506. doi:10.1016/S1048-9843(96)90003-X

Fleenor, J. W., Smither, J. W., Atwater, L. E., 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

Fletcher, C., & Bailey, C. (2003). Assessing self-awareness: Some issues and methods. *Journal of Managerial Psychology, 18,* 395–404. doi:10.1108/02683940310484008

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational and Organizational Psychology, 73,* 303–319. doi: 10.1348/096317900167047

Gentry, W. A., Ekelund, B. Z., Hannum, K. M., & de Jong, A. (2007). A study of the discrepancy between self-and observer-ratings on managerial derailment characteristics of European managers. *European Journal of Work and Organizational Psychology, 16,* 295–325. doi:10.1080/13594320701394188

Govern, J. M., & Marsch, L. A. (2001). Development and validation of the Situational Self-Awareness Scale. *Consciousness and Cognition*, *10*, 366–378. doi:10.1006/ccog.2001.0506

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. W., & Gupta, V. (Eds.) (2004). *Leadership, culture, and organizations: The GLOBE study of 62 societies.* Thousand Oaks, CA: Sage.

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

Leslie, J. B., & Peterson, M. J. (2011). *The Benchmarks sourcebook: Three decades of related research*. Greensboro, NC: Center for Creative Leadership.

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: A Multidisciplinary Journal, 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.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7,* 19–40. doi: 10.1037//1082-989X.7.1.19

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,* 91–110. doi:10.1080/10705511.2014.990153

Mahlke, J., Schultze, M., & Eid, M. (in press). Analysing multisource feedback with multilevel structural equation models – Pitfalls and recommendations from a simulation study. *British Journal of Mathematical and Statistical.* doi:10.1111/bmsp.12149

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

Mersman, J. L., & Donaldson, S. I. (2000). Factors affecting the convergence of self-peer ratings on contextual and task performance. *Human Performance*, *13*, 299–322. doi: 10.1207/s15327043hup1303_4

Morin, A. (2011). Self-awareness part 1: Definition, measures, effects, functions, and antecedents. *Social and Personality Psychology Compass*, *5*, 807-823. doi:10.1111/j.1751-9004.2011.00387.x

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,* 557–576. doi:10.1111/j.1744-6570.1998.tb00251.x

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology*, *57*, 333–375. doi:10.1111/j.1744-6570.2004.tb02494.x

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

Saal, F. E., Downey R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413–428. doi:10.1037/0033-2909.88.2.413

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

Seibert, S. E., Silver, S. R., & Randolph, W. A. (2004). Taking empowerment to the next level: A multiple-level model of empowerment, performance, and satisfaction. *Academy Of Management Journal*, *47,* 332–349. doi:10.2307/20159585

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, *30*, 526–537. doi:10.1037/h0037039

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology*, *51*, 125–139. doi: 10.1037/0022-3514.51.1.125

Taylor, S. N., Wang, M., & Zhan, Y. (2012). Going beyond self-other rating comparison to measure leader self-awareness. *Journal of Leadership Studies*, *6*(2), 6–31. doi:10.1002/jls.21235

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29. doi:10.1037/h0071663

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

Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, *32*, 249–263. doi:10.1002/hrm.3930320205

Vecchio, R. P., & Anderson, R. J. (2009). Agreement in self-other ratings of leader effectiveness: The role of demographics and personality. *International Journal of Selection and Assessment*, *17,* 165–179. doi:10.1111/j.1468-2389.2009.00460.x

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

Walumbwa, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S., & Peterson, S. J. (2008). Authentic leadership: Development and validation of a theory-based measure. *Journal of Management*, *34*, 89–126. doi:10.1177/0149206307308913

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

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

Wohlers, A. I., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42,* 235–261. doi:10.1111/j.1744-6570.1989.tb00656.x

Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as other 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

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30,* 165–200. doi:10.1111/0081-1750.00078

Appendix A

Mplus Input for Model 1 (Absolute Differences)

```
TITLE: Self-Other-Agreement on Leadership Competencies as
       Absolute Difference and Self-Awareness

DATA:

   FILE = BMK_absolute.dat;

VARIABLE:

   NAMES =   ID country
             sa1_s sa2_s sa3_s sa4_s
             sa1_b sa2_b sa3_b sa4_b
             sa1_d sa2_d sa3_d sa4_d
             sa1_p sa2_p sa3_p sa4_p
             abs1_sb abs2_sb abs3_sb
             abs1_sp abs2_sp abs3_sp
             abs1_sd abs2_sd abs3_sd;

             ! sa = self-awareness
             ! abs = absolute difference
             ! s = self; b = boss
             ! d = direct report (subordinate)
             ! p = peer
             ! sb = self-boss
             ! sp = self-peer
             ! sd = self-direct report

   USEVAR =  sa1_s sa2_s sa3_s sa4_s
             sa1_b sa2_b sa3_b sa4_b
             sa1_d sa2_d sa3_d sa4_d
             sa1_p sa2_p sa3_p sa4_p
             abs1_sb abs2_sb abs3_sb
             abs1_sp abs2_sp abs3_sp
             abs1_sd abs2_sd abs3_sd;

   BETWEEN = sa1_s sa2_s sa3_s sa4_s
             sa1_b sa2_b sa3_b sa4_b
             abs1_sb abs2_sb abs3_sb;

   MISSING = ALL(-9999);

   CLUSTER = country ID;

ANALYSIS:

   TYPE = complex twolevel;
   H1INTERATIONS = 10000;

MODEL:

   %WITHIN%
   sa_p1 by sa1_p sa2_p sa3_p sa4_p;
   sa_d1 by sa1_d sa2_d sa3_d sa4_d;

   abs_p1 by abs1_sp abs2_sp abs3_sp;
```

```
  abs_d1 by abs1_sd abs2_sd abs3_sd;

  sa_p1 with sa_d1@0 abs_d1@0;
  abs_p1 with sa_d1@0 abs_d1@0;

  %BETWEEN%
  sa_s by sa1_s sa2_s sa3_s sa4_s;
  sa_b by sa1_b sa2_b sa3_b sa4_b;
  sa_p2 by sa1_p sa2_p sa3_p sa4_p;
  sa_d2 by sa1_d sa2_d sa3_d sa4_d;

  abs_b by abs1_sb abs2_sb abs3_sb;
  abs_p2 by abs1_sp abs2_sp abs3_sp;
  abs_d2 by abs1_sd abs2_sd abs3_sd;

  is2 by abs2_sb abs2_sp abs2_sd;
  is3 by abs3_sb abs3_sp abs3_sd;

  is2 with sa_s@0 sa_b@0 sa_p2@0 sa_d2@0;
  is2 with abs_b@0 abs_p2@0 abs_d2@0;
  is3 with sa_s@0 sa_b@0 sa_p2@0 sa_d2@0;
  is3 with abs_b@0 abs_p2@0 abs_d2@0;

  sa1_p@0 sa2_p@0 sa3_p@0 sa4_p@0;
  sa1_d@0 sa2_d@0 sa3_d@0 sa4_d@0;
  abs1_sp@0 abs2_sp@0 abs3_sp@0;
  abs1_sd@0 abs2_sd@0 abs3_sd@0;

OUTPUT: STDYX CINTERVAL;

SAVEDATA:

  FILE = m1_fscores.dat;
  SAVE = fscores;
```

Appendix B

Mplus Input for Model 2 (Algebraic Differences)

```
TITLE: Self-Other-Agreement on Leadership Competencies as
       Algebraic Difference and Self-Awareness

DATA:

  FILE = BMK_algebraic.dat;

VARIABLE:

  NAMES =   ID country
            sa1_s sa2_s sa3_s sa4_s
            sa1_b sa2_b sa3_b sa4_b
            sa1_d sa2_d sa3_d sa4_d
            sa1_p sa2_p sa3_p sa4_p
            dif1_sb dif2_sb dif3_sb
            dif1_sp dif2_sp dif3_sp
            dif1_sd dif2_sd dif3_sd;

            ! sa = self-awareness
            ! dif = algebraic difference
            ! s = self; b = boss
            ! d = direct report (subordinate)
            ! p = peer
            ! sb = self-boss
            ! sp = self-peer
            ! sd = self-direct report


  USEVAR =  sa1_s sa2_s sa3_s sa4_s
            sa1_b sa2_b sa3_b sa4_b
            sa1_d sa2_d sa3_d sa4_d
            sa1_p sa2_p sa3_p sa4_p
            dif1_sb dif2_sb dif3_sb
            dif1_sp dif2_sp dif3_sp
            dif1_sd dif2_sd dif3_sd;

  BETWEEN = sa1_s sa2_s sa3_s sa4_s
            sa1_b sa2_b sa3_b sa4_b
            dif1_sb dif2_sb dif3_sb;

  MISSING = ALL(-9999);

  CLUSTER = country ID;

ANALYSIS:

  TYPE = complex twolevel;
  H1ITERATIONS = 10000;


MODEL:

  %WITHIN%
  sa_p1 by sa1_p sa2_p sa3_p sa4_p;
  sa_d1 by sa1_d sa2_d sa3_d sa4_d;
```

```
   dif_p1 by dif1_sp dif2_sp dif3_sp;
   dif_d1 by dif1_sd dif2_sd dif3_sd;

   sa_p1 with sa_d1@0 dif_d1@0;
   dif_p1 with sa_d1@0 dif_d1@0;

   %BETWEEN%
   sa_s by sa1_s sa2_s sa3_s sa4_s;
   sa_b by sa1_b sa2_b sa3_b sa4_b;
   sa_p2 by sa1_p sa2_p sa3_p sa4_p;
   sa_d2 by sa1_d sa2_d sa3_d sa4_d;

   dif_b by dif1_sb dif2_sb dif3_sb;
   dif_p2 by dif1_sp dif2_sp dif3_sp;
   dif_d2 by dif1_sd dif2_sd dif3_sd;

   is2 by dif2_sb dif2_sp dif2_sd;
   is3 by dif3_sb dif3_sp dif3_sd;

   is2 with sa_s@0 sa_b@0 sa_p2@0 sa_d2@0;
   is2 with dif_b@0 dif_p2@0 dif_d2@0;
   is3 with sa_s@0 sa_b@0 sa_p2@0 sa_d2@0;
   is3 with dif_b@0 dif_p2@0 dif_d2@0;

   sa1_p@0 sa2_p@0 sa3_p@0 sa4_p@0;
   sa1_d@0 sa2_d@0 sa3_d@0 sa4_d@0;
   dif1_sp@0 dif2_sp@0 dif3_sp@0;
   dif1_sd@0 dif2_sd@0 dif3_sd@0;

OUTPUT: STDYX CINTERVAL;

SAVEDATA:

   FILE = m2_fscores.dat;
   SAVE = fscores;
```

Appendix C

Mplus Input for Model 3 (Method Factors)

```
TITLE: Method Factors of Leadership Competencies and Self-Awareness

DATA:

  FILE = BMK_methodfactors.dat;

VARIABLE:

  NAMES =   ID country
            sa1_s sa2_s sa3_s sa4_s
            com1_s com2_s com3_s
            sa1_b sa2_b sa3_b sa4_b
            com1_b com2_b com3_b
            sa1_d sa2_d sa3_d sa4_d
            com1_d com2_d com3_d
            sa1_p sa2_p sa3_p sa4_p
            com1_p com2_p com3_p;

            ! sa = self-awareness
            ! com = leadership competencies
            ! s = self; b = boss
            ! d = direct report (subordinate)
            ! p = peer

  USEVAR =  sa1_s sa2_s sa3_s sa4_s
            sa1_b sa2_b sa3_b sa4_b
            sa1_d sa2_d sa3_d sa4_d
            sa1_p sa2_p sa3_p sa4_p
            com1_s com2_s com3_s
            com1_b com2_b com3_b
            com1_d com2_d com3_d
            com1_p com2_p com3_p;

  BETWEEN = sa1_s sa2_s sa3_s sa4_s
            sa1_b sa2_b sa3_b sa4_b
            com1_s com2_s com3_s
            com1_b com2_b com3_b;

  MISSING = ALL(-9999);

  CLUSTER = country ID;

ANALYSIS:

  TYPE = complex twolevel;
  H1ITERATIONS = 10000;

MODEL:

  %WITHIN%
  sa_p1 by sa1_p sa2_p sa3_p sa4_p;
  sa_d1 by sa1_d sa2_d sa3_d sa4_d;

  ! Define method factors of leadership competencies
  ! on level 1
```

```
   com_UMp by com1_p com2_p com3_p;
   com_UMd by com1_d com2_d com3_d;

   sa_p1 with sa_d1@0 com_UMd@0;
   com_UMp with com_UMd@0 sa_d1@0;

   %BETWEEN%
   sa_s by sa1_s sa2_s sa3_s sa4_s;
   sa_b by sa1_b sa2_b sa3_b sa4_b;
   sa_p2 by sa1_p sa2_p sa3_p sa4_p;
   sa_d2 by sa1_d sa2_d sa3_d sa4_d;

   ! Define trait factor of leadership competencies
   com_T by com1_s com2_s com3_s
            com1_b com2_b com3_b
            com1_p com2_p com3_p
            com1_d com2_d com3_d;

   ! Define method factors on level 2
   com_Mb by com1_b com2_b com3_b;
   com_CMp by com1_p com2_p com3_p;
   com_CMd by com1_d com2_d com3_d;

   com_T with com_Mb@0 com_CMp@0 com_CMd@0;

   sa1_p@0 sa2_p@0 sa3_p@0 sa4_p@0;
   sa1_d@0 sa2_d@0 sa3_d@0 sa4_d@0;
   com1_p@0 com2_p@0 com3_p@0;
   com1_d@0 com2_d@0 com3_d@0;

OUTPUT: STDYX CINTERVAL;

SAVEDATA:

   FILE = m3_fscores.dat;
   SAVE = fscores;
```

Table 1

Model Fit

Index	Model 1 (absolute differences)	Model 2 (algebraic differences)	Model 3 (method factors)
adj. SRMR_{L1}	0.015	0.011	0.013
SRMR_{L2}	0.044	0.041	0.044
CFI	1.000	1.000	1.000
RMSEA	0.000	0.000	0.000

Note. adj. SRMR_{L1}= adjusted Standardized Root Mean Square Residual for level 1, SRMR_{L2}= Standardized Root Mean Square Residual for level 2, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation.

Table 2

Indicator Means and Standard Deviations (in Parentheses)

Level 1				Level 2			
Indicator	Model 1 (absolute differences)	Model 2 (algebraic differences)	Model 3 (method factors)	Indicator	Model 1 (absolute differences)	Model 2 (algebraic differences)	Model 3 (method factors)
SOA1$_{\text{self-peer}}$	0.49 (0.39)	0.02 (0.63)	X	SOA1$_{\text{self-boss}}$	0.44 (0.35)	-0.07 (0.56)	X
SOA2$_{\text{self-peer}}$	0.49 (0.39)	0.01 (0.62)	X	SOA2$_{\text{self-boss}}$	0.44 (0.35)	-0.07 (0.56)	X
SOA3$_{\text{self-peer}}$	0.48 (0.38)	0.01 (0.61)	X	SOA3$_{\text{self-boss}}$	0.43 (0.34)	-0.05 (0.55)	X
SOA1$_{\text{self-subord}}$	0.53 (0.42)	-0.06 (0.67)	X	COM1$_{\text{self}}$	X	X	3.94 (0.38)
SOA2$_{\text{self-subord}}$	0.53 (0.41)	-0.07 (0.66)	X	COM2$_{\text{self}}$	X	X	3.90 (0.38)
SOA3$_{\text{self-subord}}$	0.52 (0.41)	-0.07 (0.66)	X	COM3$_{\text{self}}$	X	X	3.91 (0.38)
COM1$_{\text{peer}}$	X	X	3.92 (0.56)	COM1$_{\text{boss}}$	X	X	4.01 (0.48)
COM2$_{\text{peer}}$	X	X	3.89 (0.54)	COM2$_{\text{boss}}$	X	X	3.96 (0.46)
COM3$_{\text{peer}}$	X	X	3.91 (0.53)	COM3$_{\text{boss}}$	X	X	3.96 (0.47)
COM1$_{\text{subord}}$	X	X	4.00 (0.61)	SA1$_{\text{self}}$	4.11 (0.66)	4.11 (0.66)	4.11 (0.66)
COM2$_{\text{subord}}$	X	X	3.97 (0.60)	SA2$_{\text{self}}$	4.11 (0.66)	4.11 (0.66)	4.11 (0.66)
COM3$_{\text{subord}}$	X	X	3.99 (0.60)	SA3$_{\text{self}}$	3.75 (0.80)	3.75 (0.80)	3.75 (0.80)
SA1$_{\text{peer}}$	3.85 (0.85)	3.85 (0.85)	3.85 (0.85)	SA4$_{\text{self}}$	3.92 (0.65)	3.92 (0.65)	3.92 (0.65)
SA2$_{\text{peer}}$	3.91 (0.80)	3.91 (0.80)	3.91 (0.80)	SA1$_{\text{boss}}$	4.04 (0.75)	4.04 (0.75)	4.04 (0.75)
SA3$_{\text{peer}}$	3.65 (0.92)	3.65 (0.92)	3.65 (0.92)	SA2$_{\text{boss}}$	3.99 (0.75)	3.99 (0.75)	3.99 (0.75)
SA4$_{\text{peer}}$	3.83 (0.79)	3.83 (0.79)	3.83 (0.79)	SA3$_{\text{boss}}$	3.86 (0.84)	3.86 (0.84)	3.86 (0.84)
SA1$_{\text{subord}}$	3.91 (0.92)	3.91 (0.92)	3.91 (0.92)	SA4$_{\text{boss}}$	3.86 (0.77)	3.86 (0.77)	3.86 (0.77)
SA2$_{\text{subord}}$	4.00 (0.85)	4.00 (0.85)	4.00 (0.85)				
SA3$_{\text{subord}}$	3.63 (1.02)	3.63 (1.02)	3.63 (1.02)				
SA4$_{\text{subord}}$	3.93 (0.84)	3.93 (0.84)	3.93 (0.84)				

Note. SOA1 to SOA3 = indicators of self-other-agreement on leadership competencies, SA1 to SA4 = indicators of self-awareness, COM1 to COM3 = indicators of leadership competencies, subord = subordinate, X = not a factor in the model. For model 1, means and standard deviations refer to the absolute value of SOA, that is to |SOA|.

Table 3

Factor Correlations in Model 1 (Absolute Differences)

Level 1

Factor	1	2	3	4
1. $\lvert SOA_{\text{self-unique peer}}\rvert$	1			
2. $\lvert SOA_{\text{self-unique subord}}\rvert$	X	1		
3. $SA_{\text{unique peer}}$	-.16 [-.20, -.12]	X	1	
4. $SA_{\text{unique subord}}$	X	-.12 [-.15, -.09]	X	1

Level 2

Factor	1	2	3	4	5	6	7
1. $\lvert SOA_{\text{self-boss}}\rvert$	1						
2. $\lvert SOA_{\text{self-common peer}}\rvert$	.48 [.47, .50]	1					
3. $\lvert SOA_{\text{self-common subord}}\rvert$	.37 [.35, .39]	.73 [.72, .75]	1				
4. SA_{self}	-.06 [-.11, -.00]	-.00 [-.10, .10]	-.17 [-.24, -.10]	1			
5. SA_{boss}	-.02 [-.06, .02]	-.13 [-.18, -.09]	-.06 [-.09, -.03]	.10 [.08, .12]	1		
6. $SA_{\text{common peer}}$	.02 [-.00, .04]	-.21 [-.25, -.17]	-.09 [-.12, -.06]	.19 [.16, .22]	.48 [.45, .51]	1	
7. $SA_{\text{common subord}}$	.01 [-.00, .03]	-.11 [-.14, -.08]	-.19 [-.23, -.15]	.17 [.16, .19]	.32 [.29, .36]	.67 [.65, .70]	1

Note. All values are Pearson correlations *r* with 95% CIs in brackets. |SOA| = self-other-agreement on leadership competencies as absolute difference, SA = self-awareness, subord = subordinate, X = nonadmissible correlations. Grey cells contain those correlations that are interpreted in order to answer the research questions. For simplicity reasons, we do not depict the correlation (*r* = .48) of the two indicator-specific factors of |SOA|.

Table 4

Factor Correlations in Model 2 (Algebraic Differences)

Level 1 Factor	1	2	3	4
1. SOA$_{\text{self-unique peer}}$	1			
2. SOA$_{\text{self-unique subord}}$	X	1		
3. SA$_{\text{unique peer}}$	-.90 [-.91, -.90]	X	1	
4. SA$_{\text{unique subord}}$	X	-.93 [-.93, -.92]	X	1

Level 2 Factor	1	2	3	4	5	6	7
1. SOA$_{\text{self-boss}}$	1						
2. SOA$_{\text{self-common peer}}$	.70 [.68, .71]	1					
3. SOA$_{\text{self-common subord}}$	.61 [.60, .62]	.85 [.84, .86]	1				
4. SA$_{\text{self}}$	.48 [.45, .51]	.63 [.61, .66]	.59 [.58, .61]	1			
5. SA$_{\text{boss}}$	-.67 [-.69, -.66]	-.25 [-.27, -.24]	-.17 [-.19, -.16]	.11 [.08, .13]	1		
6. SA$_{\text{common peer}}$	-.28 [-.32, -.24]	-.50 [-.53, -.47]	-.33 [-.36, -.29]	.19 [.16, .22]	.48 [.46, .51]	1	
7. SA$_{\text{common subord}}$	-.17 [-.22, -.13]	-.29 [-.33, -.25]	-.55 [-.57, -.52]	.18 [.16, .20]	.33 [.30, .36]	.66 [.64, .69]	1

Note. All values are Pearson correlations *r* with 95% CIs in brackets. SOA = self-other-agreement in leadership competencies as algebraic difference, SA = self-awareness, subord = subordinate, X = nonadmissible correlations. Grey cells contain those correlations that are interpreted in order to answer the research questions. For simplicity reasons, we do not depict the correlation (*r* = .48) of the two indicator-specific factors of SOA.

Table 5

Factor Correlations in Model 3 (Method Factors)

Level 1

Factor	1	2	3	4
1. UM$_{\text{unique peer}}$	1			
2. UM$_{\text{unique subord}}$	X	1		
3. SA$_{\text{unique peer}}$	.90 [.89, .91]	X	1	
4. SA$_{\text{unique subord}}$	X	.93 [.92, .93]	X	1

Level 2

Factor	1	2	3	4	5	6	7	8
1. T$_{\text{self}}$	1							
2. M$_{\text{boss}}$	X	1						
3. CM$_{\text{common peer}}$	X	.49 [.45, .52]	1					
4. CM$_{\text{common subord}}$	X	.38 [.35, .41]	.73 [.71, .76]	1				
5. SA$_{\text{self}}$	.84 [.84, .85]	-.04 [-.06, -.03]	-.04 [-.05, -.03]	-.06 [-.08, -.05]	1			
6. SA$_{\text{boss}}$	.07 [.05, .09]	.85 [.84, .86]	.44 [.42, .47]	.31 [.28, .33]	.11 [.08, .13]	1		
7. SA$_{\text{common peer}}$	.16 [.13, .19]	.38 [.33, .42]	.92 [.91, .93]	.61 [.57, .65]	.20 [.17, .22]	.45 [.42, .48]	1	
8. SA$_{\text{common subord}}$	.20 [.18, .22]	.31 [.27, .34]	.66 [.64, .68]	.92 [.91, .93]	.19 [.17, .21]	.32 [.30, .35]	.67 [.65, .70]	1

Note. All values are Pearson correlations *r* with 95% CIs in brackets. T = trait factor of self-reported leadership competencies, UM / CM / M = unique method factor / common method factor / method factor of leadership competencies, SA = self-awareness, subord = subordinates, X = nonadmissible correlations. Grey cells contain those correlations that are interpreted in order to answer the research questions.

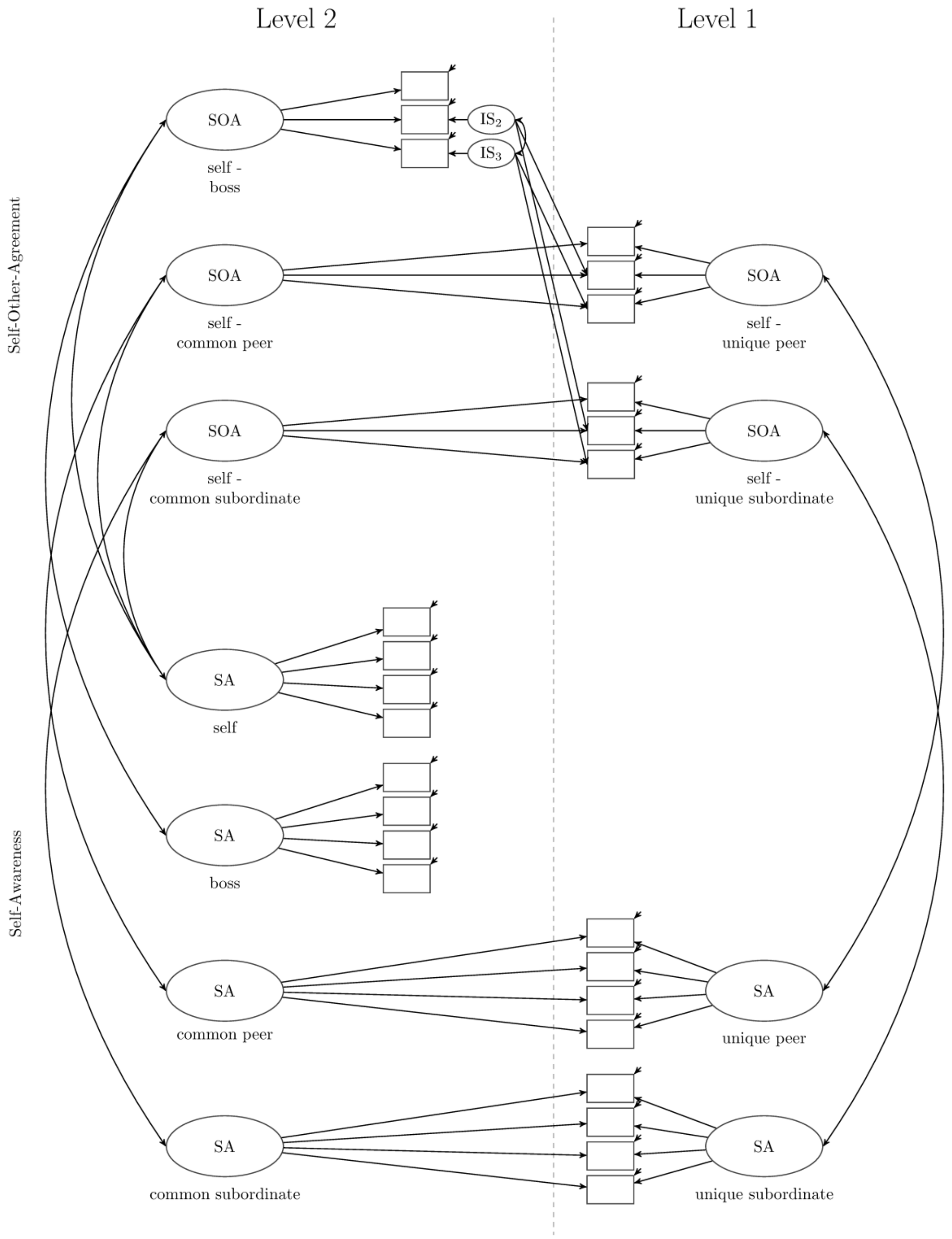

Figure 1. Illustration of model 1 and model 2. SOA = self-other-agreement on leadership competencies, SA = self-awareness, IS$_2$/IS$_3$ = indicator-specific factors. In model 1, SOA is the algebraic difference between self- and others' ratings. In model 2, SOA is the absolute difference between self- and others' ratings. Small arrows to the indicators indicate error variance. For simplicity reasons, we depict only those correlations that will be interpreted in the analysis. However, all factors per level are allowed to correlate.

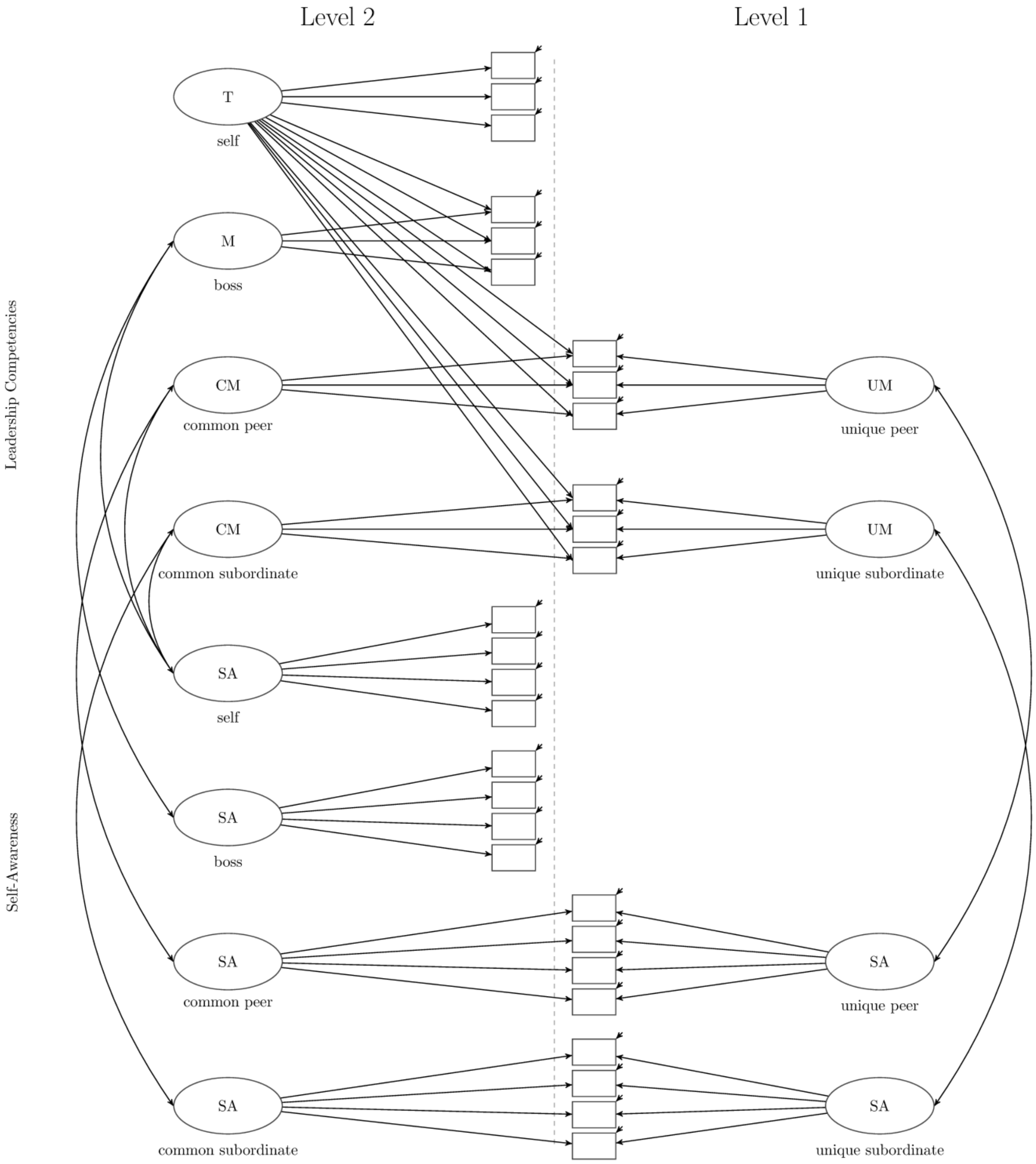

Figure 2. Illustration of model 3. T = trait factor of leadership competencies measured by the self-ratings, M = method factor of leadership competencies, CM = common method factor of leadership competencies, UM = unique method factor of leadership competencies, SA = self-awareness. Small arrows to the indicators indicate error variance. For simplicity reasons, we depict only those correlations that will be interpreted in the analysis. However, all factors per level are allowed to correlate except for the trait factor that does not correlate with the method factor and the common method factors.

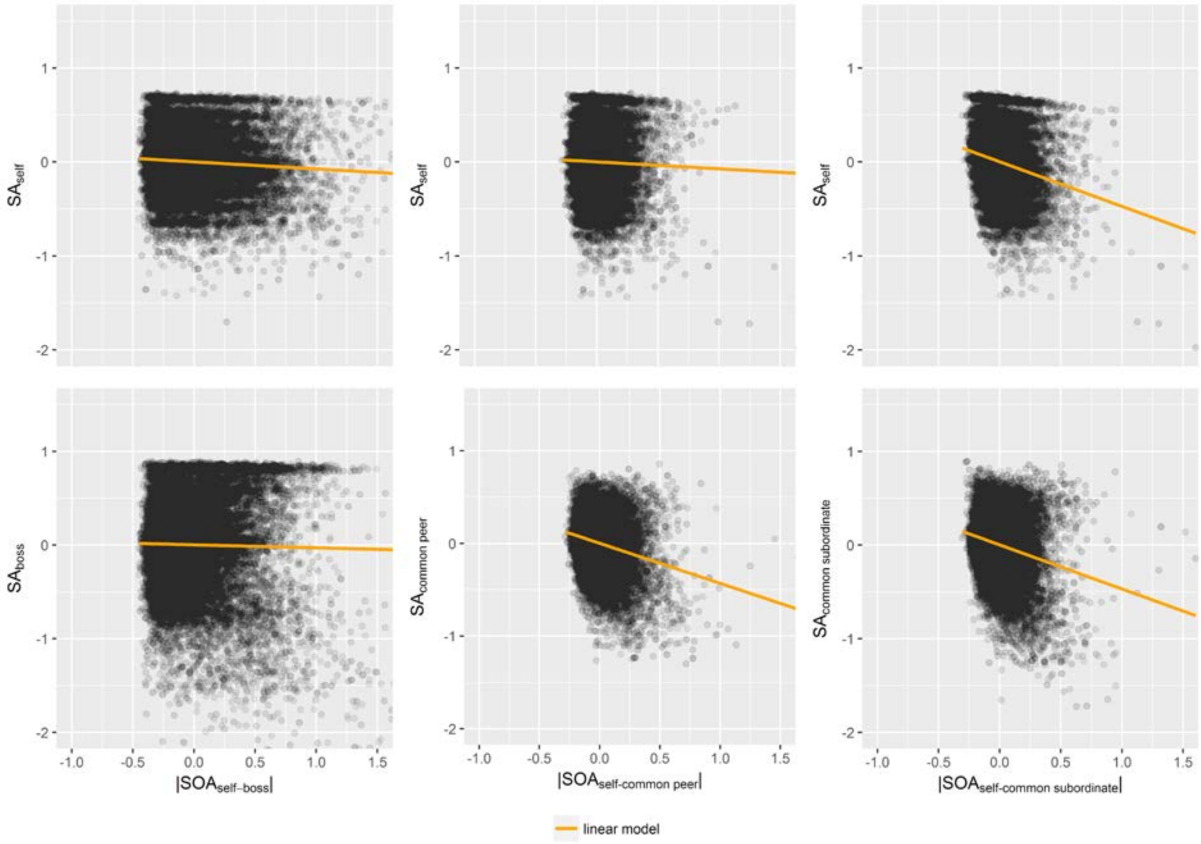

Figure 3a

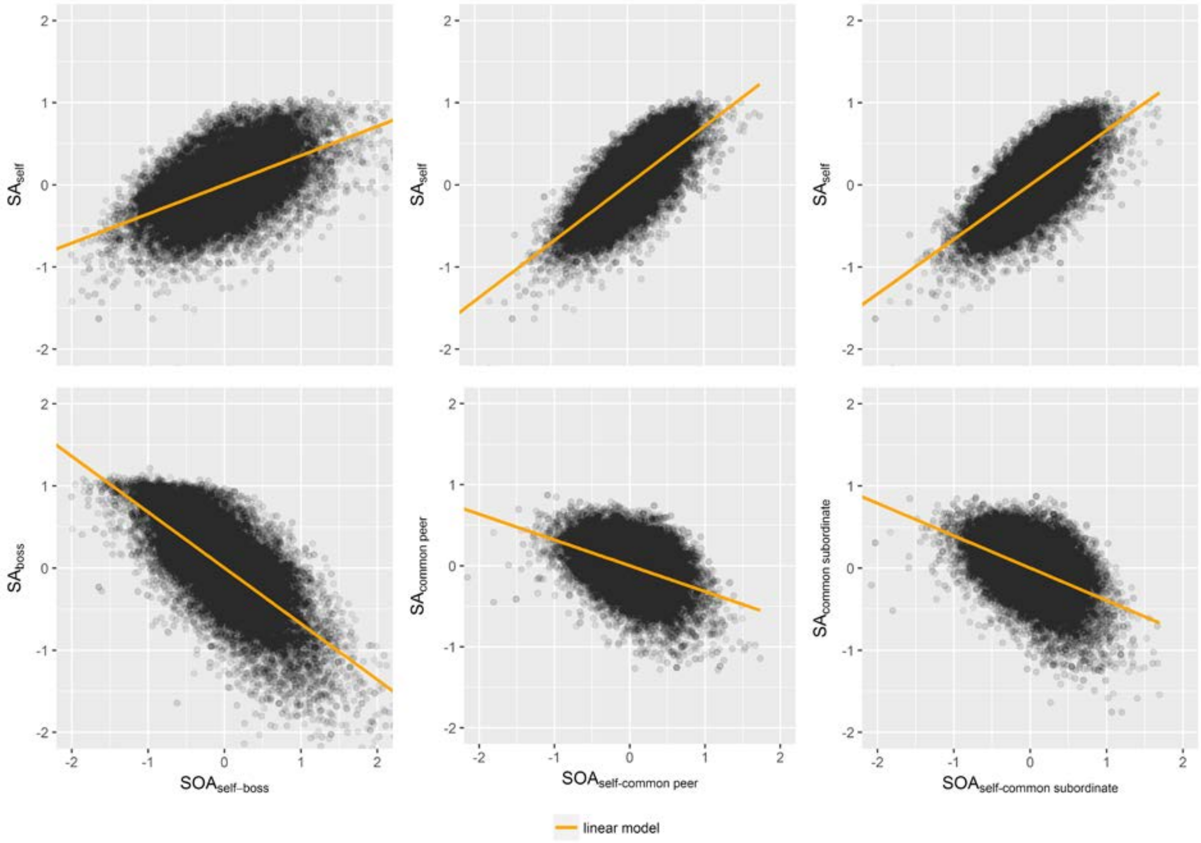

Figure 3b

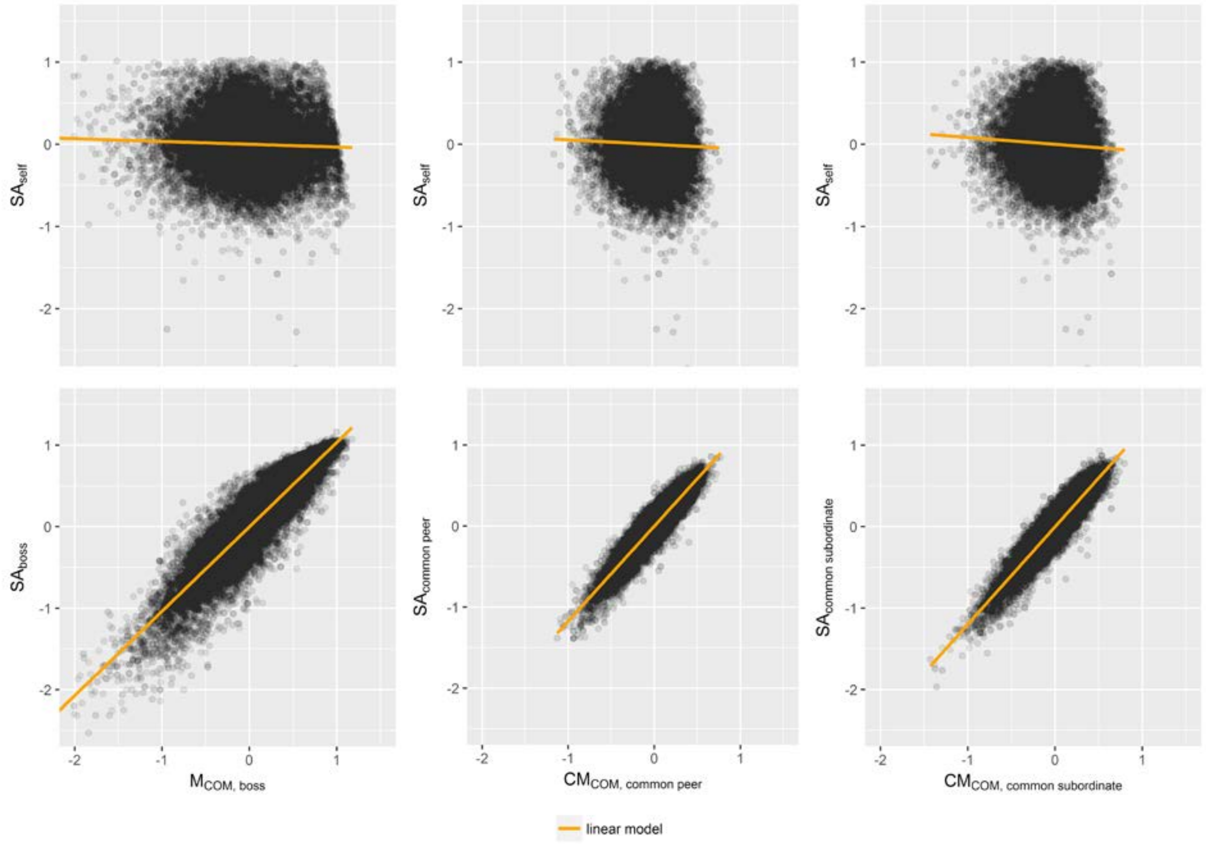

Figure 3c

Figure 3a-c. Scatterplots for a) model 1, b) model 2, and c) model 3. |SOA| = absolute self-other-agreement, SOA = algebraic self-other-agreement, SA = self-awareness, M = method factor and CM = common method factor of leadership competencies (COM) after controlling for self-ratings. Note that all plotted variables have a mean of zero because they are latent variables in Mplus.

Chapter 5

General Discussion

1. Conclusions From the Previous Chapters

The purpose of this thesis was to develop a flexible statistical model that allows integrating all rating perspectives of 360-degree feedback assessment. Such a model did – to my knowledge – not yet exist but is urgently needed when a 360-degree feedback instrument is to be validated or when other research questions in this context, e.g. on SOA in leadership ratings, are investigated. The development of this model was successfully accomplished by using the multilevel CFA-MTMM model for the combination of structurally different and interchangeable methods (Eid et al., 2008) as a starting point and extending it to multiple sets of interchangeable methods on level 1.

Chapter 2 presented the definition of the multilevel CFA-MTMM model for two sets of interchangeable methods and one structurally different method, its implementation in the software Mplus (L. K. Muthén & Muthén, 1998–2017), and an empirical application. The key advance of this model is that it incorporates two different level-1 populations that are both nested within the same clusters. The main challenge in the model development process was to make the software treat the data so that it corresponds to the model definition. This was achieved by using a planned missing data format for level 1 and then defining a traditional multilevel structural equation model with a full information maximum likelihood (FIML) estimator.

In Chapter 3, a Monte Carlo simulation study demonstrated that the model converges in almost all cases and that parameters and standard errors were estimated with high validity. These results strongly supported the newly developed model and its implementation in the software. The distribution of χ^2-values, however, was biased and shifted leftward. This would lead to an overestimation of the goodness of fit and the acceptance of too many falsely specified models. The χ^2-test therefore should not be trusted for this specific type of model.

Chapter 4 is an empirical application of the model that analyzed the relationship between SOA in leadership ratings and self-awareness. It revealed that the correlations between the two constructs were almost completely attributable to method effects. If this shared method variance between the two constructs was explicitly captured, there was no substantial relationship remaining between SOA and self-awareness. This result questions the routine of blaming a manager for lacking self-awareness when his or her self-evaluations are in low agreement with others' ratings.

In the following, I will discuss the results of the previous chapters from a broader perspective. This will include highlighting the scientific contribution of this thesis and pointing out possible applications and extensions as well as limitations of the model.

2. Scientific Contribution

When I started working on this thesis, there have already been many methods applied in 360-degree feedback research. The magnitude of different models can partly be explained by the wide scope of research questions that are aimed to answer. However, two more reasons for this inconsistency are possible. On the one hand, it can indicate that researchers struggle to identify the method that is best suited to model the complex data structure at hand. Eid et al. (2008) claimed that "most applied researchers are still uncertain as to which model should be chosen in a specific application" (p. 231). According to the authors, the decision for a specific model is often driven by model fit after having defined all available MTMM models instead of being based on reflections on the specific measurement design. On the other hand, it can indicate that there simply does not exist an adequate model so that each researcher tries to find the most suitable model out of a range of possible ones, each one having its own shortcomings.

The thesis at hand aims to address both of these issues. First, the structure of 360-degree feedback data has been thoroughly identified following the concept of structurally different vs. interchangeable methods (Eid et al., 2008) and the allocation of ratings to two levels of measurement in a multilevel framework. The need for the development of a new model that is able to consider peers and subordinates as two sets of interchangeable methods was stressed. Second, this new model was presented, simulated, and applied. The model offers a flexible framework for the analysis of MTMM data and overcomes many of the issues of previous models discussed in the introduction.

2.1 Correction for Unreliability of Ratings

As the model is built in the structural equation modeling framework, it explicitly models measurement error so that all variance components can be reported error-free and are estimated with higher validity as it is possible with manifest measures. In the analysis in Chapter 2, between nine and 37% of variance in ratings was due to measurement error. In a model using manifest measures, this error variance would remain unmodeled and be a potentially biasing component of parameter estimates. Contrariwise, the variance components (consistency, unique and common method specificity, discriminant validity) reported in Chapter 2 are free of measurement error.

2.2 Well-Defined Components of Trait, Method, and Error Variance

Previous research in the field predominantly used the CT-CM model (e.g., Hoffman et al., 2010; Hoffman & Woehr, 2009; Mount et al., 1998). As discussed in the introduction, this model has some drawbacks. For example, Scullen, Mount, and Goff (2000) obtained improper solutions when trying to fit a CT-CM model to 360-degree feedback data and therefore switched to a different model. Instead of relying on a CT-CM structure, the model presented in this thesis uses a CT-C(*M*-1) structure on level 2. Accordingly, trait, method, and error components of ratings cannot only be separated but are, in contrast to the CT-CM model, well-defined. This enables a clear interpretation of the model parameters. Additionally, the rate of improper solutions in the simulation study (Chapter 3) was very low. Only 2% of all models were affected and most of them did not show improper solutions that are typically of relevance to the user. The simulation study in Chapter 3 has also shown that model convergence, another issue that is often encountered with the CT-CM model (Castro-Schilo et al., 2013), was very satisfying (99.98%).

2.3 Multilevel Data Structure

The model considers the multilevel structure of 360-degree feedback ratings. It locates self-ratings and boss ratings on level 2 and peer and subordinate ratings on level 1. The clustering of peers and subordinates in target managers enables the researcher to include a varying number of peers and a varying number of subordinates. It is therefore no longer necessary to aggregate the peer and subordinate ratings (Atwater et al., 2009; Church, 1997; Flechter & Baldry, 2000) or to randomly choose a fixed number of peers and subordinates as it was often done (Diefendorff, Silverman, & Greguras, 2005; Hoffman et al., 2010; Scullen et al., 2000; Woehr et al., 2005). As the data used in this field of research typically stem from real applications of 360-degree feedback programs (instead of being collected explicitly for research purposes), the number of peers and subordinates often varies between target managers because simply all available and willing raters are included. For example, in the data set used for the study presented in Chapter 2, the number of peers varied between zero and 17 and the number of subordinates varied between zero and 38. All of these ratings could be included in the analyses so that no available information was disregarded. Instead, the variance of peers and the variance of subordinates given a target manager was explicitly modeled and further analyzed on level 1.

Recent research (Hoffmann et al., 2010; Mount et al., 1998; Scullen et al., 2000, Woehr et al., 2005) has tried to disentangle variance in others' ratings into a part that is attributable to the specific rating source and a part that is attributable to the individual rater. Thus, the variance that was elsewhere only generally attributed to the rating method has been split into the variance due to the unique method, that is, the unique peer or subordinate (*idiosyncratic rater effects*), and the common method, that is, the group of peers or the group of subordinates (*rating source effects*). While some authors found that method effects in 360-degree feedback ratings were nearly completely due to idiosyncratic rater effects (Mount et al., 1998; Scullen et al., 2000), others confirmed the existence of rating source effects (Hoffmann et al., 2010; Woehr et al., 2005). The average proportion of variance in ratings due to idiosyncratic rater effects varied between 55% (Hoffmann et al., 2010) and 58% (Scullen et al., 2000). In the study of Woehr et al. (2005), idiosyncratic rater effect and measurement error effects were confounded and accountable for a combined variance proportion of 45% on average. The average variance proportion due to rating source effects was considerably smaller (8% in Scullen et al., 2000; 21% in Woehr et al., 2005; 22% in Hoffmann et al., 2010).

This ongoing debate about the (non-)existence of rating source effects has the potential to question the meaningfulness of 360-degree feedback assessments. Considering all rating perspectives would be neither necessary nor reasonable from an economical point of view, if these different perspectives did not make a unique contribution. In light of this momentous debate, it is deplorable that available peer and subordinate ratings were disregarded because they could not be included in models that ignore the multilevel data structure. The variance of peer ratings and of subordinate ratings is needed to reliably distinguish idiosyncratic (or unique method) effects and rating source (or common method) effects.

In the study presented in Chapter 2, the error-free variance due to idiosyncratic rater effects (i.e., the unique method specificity) was between 64% and 73%. The remaining variance in peer and subordinate ratings, that is, 27% to 36%, was shared among the peers or among the subordinates and therefore indicates the presence of rating source effects. This latter variance proportion is higher than it was found before for two reasons. First, in contrast to other studies, the variance reported is free of measurement error. It would also be possible to define the proportion of this variance in relation to the total variance, i.e., including the error variance. However, when one is interested in defining the true amount of rating source effects, it is more reasonable to estimate this component free of measurement error. Second, the variance due to rating source effects reported above includes both – variance shared between the group of peers or subordinates and the self-report, i.e., the consistency, and variance shared among the group of peers or subordinates but not with the self-report, i.e., the common method specificity. Both of these components reflect variance that is shared among the group of peers or among the group of subordinates and therefore they both indicate rating source effects. Previous studies were based on modeling approaches that did not consider these two components that were summed up here.

Another advantage of modeling the multilevel structure is that the convergent validity of self-ratings and peer or subordinate ratings cannot only be defined as a proportion of the variance of the single peer or subordinate ratings (level-1 variance) but also as a proportion of the variance of the “average” peer or subordinate ratings (level-2 variance). While self-ratings predict only five to nine percent of the reliable variance of the single peer or subordinate ratings in Chapter 2, they predict 18% to 26% of the variance of the “average” peer or subordinate ratings. This indicates low convergent validity on the basis of the single peer or subordinate ratings but much higher convergence between a manager’s self-ratings and the “average” peer or subordinates of the given manager. This result therefore supports the idea of gathering information from multiple peers and multiple subordinates.

2.4 Peers and Subordinates as two Sets of Interchangeable Methods

Given that peer *and* subordinate ratings are nested within the target manager, the focal challenge of this thesis was how to include both of these ratings on level 1 while taking into account that they stem from two populations. This was achieved by organizing the data in a special missing by design structure and then defining a standard multilevel structural equation model. Within this approach it is ensured that there are no level-1 covariances between peer and subordinate ratings so that these ratings are treated as stemming from two distinct populations. All model parameters are estimated for both populations separately. On level 1, the unique view of peers and the unique view of subordinates is captured by two unique method factors that are not allowed to correlate because each level-1 rating either stems from a peer or from a subordinate. Therefore, there is no shared variance on level 1. On level 2, however, the correlation of the common method factors of peers and the common method factor of subordinates is admissible. It shows the degree to which common method effects generalize across peers and subordinates.

In the application in Chapter 2, there was a strong tendency that managers who were overestimated by peers were also overestimated by subordinates (between r = .70 and r = .73) and vice versa. The same correlational strength (r = .73) was found in Chapter 4, revealing a high convergent validity of the two different rater groups. These correlations should be considered in the discussion on the existence of rating source effects. While the decomposition of variance components above has revealed that a relevant amount of variance was shared among the group of peers or among the group of subordinates, the correlations between the method factors indicate that this variance is partly also shared between the two groups of peers and subordinates. This part of shared level-2 variance is therefore not unique to the rating source of peers or to the rating source of subordinates but common to some global “others’ perspective”.

3. Scope of Applications in 360-Degree Feedback Research

The scope of research questions in the context of 360-degree feedback ratings is wide and different statistical approaches have been typically chosen in the different research areas. One stream of research focusses on classical validation questions, i.e. on the different sources of variance that affect ratings in 360-degree feedback assessments in an MTMM framework. These studies quantify the reliability as well as the effects of rating method and of rating dimension (e.g., Hoffman et al., 2010; Hoffman & Woehr, 2009; Mount et al., 1998; Scullen et al., 2000; Woehr et al., 2005).

Other studies concentrate on identifying correlates of ratings in 360-degree feedback assessments or of SOA in these ratings. Among these correlates have been job performance or effectiveness (e.g., Atkins & Wood, 2002; Atwater, Waldman, Ostroff, Robie, & Johnson, 2005; Atwater et al., 2009; Fleenor et al., 1996; Ostroff et al., 2004), cultural characteristics (e.g., Atwater et al., 2005; Eckert et al., 2010; Gentry et al., 2007; Gentry, Yip, & Hannum, 2010), or personality (e.g., Bergner, Davda, Culpin, & Rybnicek, 2016).

The statistical approaches used in the field can be broadly classified into manifest vs. latent models. Table 1 shows the most prominent techniques in the different research areas. All these techniques have their unique advantages and many of the studies applying them made significant contributions to the field. However, all manifest approaches suffer from not taking measurement error into account. Therefore, in the last decades latent counterparts to all manifest approaches in Table 1 have been provided. The last column in Table 1 refers to the multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods. It highlights that the new model offers a comprehensive framework to address many of the research questions in this field.

Chapter 2 presented an application of the model in the context of a typical validation analysis. Two subscales of Benchmarks® (Lombardo, McCauley, McDonald-Mann, & Leslie, 1999), Leadings Employees and Participative Management, were analyzed. The item parcels had reliabilities between .63 and .91. The remaining – that is, reliable – variance of the peer and subordinate ratings was decomposed in the following components: Five to nine percent of reliable variance was shared with the target managers' self-ratings. This indicates low convergent validity between self-ratings and others' ratings. There were no obvious differences in convergent validity between peers and subordinates or between the two subscales. Twenty-two to twenty-six percent of variance was shared among the peers or among the subordinates but not with the target manager. Adding these two sources of variance leads to 27% to 36% of rating source effects, that is, variance shared among the peers or shared among the subordinates. Sixty-four to seventy-three percent of variance was due to idiosyncratic rater effects. Factor correlations revealed low discriminant validity between the two traits (r = .84 to r = .95). This result is in line with previous research that found high intercorrelations of subscales in leadership assessments and a strong general factor (e.g., Beehr, Ivanitskaya, Hansen, Erofeev, & Gudanowski, 2001; Hoffmann et al., 2010; for a meta-analysis see Visweswaran, Schmidt, & Ones, 2005).

Besides the traditional validation analysis, it is also possible to include further variables and to correlate them to the scales measured by the 360-degree feedback instrument. Chapter 4 demonstrates this approach and contributes to the debate on the meaning of SOA in leadership ratings. It defines the correlation between self-awareness, i.e., an external variable, and the trait and common method factors of leadership competency ratings. Applying the multilevel CFA-MTMM model helps to identify that the correlational pattern found between SOA and self-awareness is almost completely due to shared method variance.

There are many variables besides self-awareness that are discussed to be related to SOA in leadership competencies. Whenever it is reasonable to assume that shared method variance has an effect on the correlation between two constructs, the approach presented in Chapter 4 can be used to analyze the association between SOA in leadership competencies and a further variable after controlling for method effects. For example, there is a large amount of studies on the relationship between SOA on leadership competencies and leader effectiveness (e.g., Atwater et al., 2005; Atwater & Yammarino, 1992; Fleenor, McCauley, & Brutus, 1996; Ostroff et al., 2004). In these studies, leader effectiveness often is assessed by boss reports that are also used to build an index of SOA. Therefore, the covariance of the two constructs for which the relationship is analyzed is not only due to the two related constructs but also due to shared method variance. Using the approach presented in Chapter 4 allows disentangling these two sources that cause covariation.

The analysis in Chapter 4 could also be extended to a regression model, so that the latent trait or method factors serve as independent or dependent variables. For example, self-awareness could be used to predict the latent method factors of leadership competencies. If one wants to predict the common method factors by an independent variable that correlates also with the latent trait variable, this leads to a suppression structure and potentially to serious methodological problems (Koch, Kelava, & Eid, 2018). In this case, the explanatory variables should be residualized as described in Koch et al. (2018), so that they do no longer correlate with the latent trait variable.

Table 1

Statistical Approaches to Typical 360-Degree Feedback Research Questions

Manifest approaches	**Latent approaches**	**Multilevel CFA-MTMM model**
Amount of the different variance components in ratings		
Correlation analyses to quantify the consistency between different rating methods (for a meta-analysis see Conway & Huffcut, 1997), or Within and between analysis (WABA; Yammarino & Markham, 1992) that uses ANOVA and ANCOVA procedures to define (co-)variation of variables within and between groups	CFA models to quantify the proportion of variance attributable to rating method, to rating dimension, and to measurement error (e.g., Hoffman et al., 2010; Scullen et al., 2000)	Estimation of the proportion of variance attributable to consistency, to unique method specificity, to common method specificity, and to measurement error (Chapter 2)
Correlates of ratings or of SOA in ratings: Correlation analyses		
Continuous difference scores or categories of agreement to measure SOA and correlate it to external variables (e.g., Atwater & Yammarino, 1992; Fletcher & Baldry, 2000)	CFA models in which the different variance components of ratings are correlated to external variables (Hoffman & Woehr, 2009)	Possible inclusion of further variables that are correlated to the model's trait or method factors, e.g., self-awareness (Chapter 4)
Correlates of ratings or of SOA in ratings: Regression analyses		
Multivariate regression where self-ratings and others' ratings are jointly analyzed as outcome variables to predict SOA by an external variable (Gentry et al., 2007), or	Latent regression techniques that are direct adoptions of the manifest regression models into the structural equation modeling framework (Edwards, 2009)	Possible extension of the model to regression analyses with the model's trait or method factors as outcomes or predictors. If common method factors are predicted, the

Table 1 (continued)

Manifest approaches	**Latent approaches**	**Multilevel CFA-MTMM model**
Polynomial regression where an outcome variable is predicted by self-ratings and others' ratings including quadratic effects to determine the shape of the underlying relationship (e.g., Atwater et al., 1998)		approach presented by Koch, Kelava & Eid (2018) should be considered (for a discussion see this chapter)

Note. The last column refers to the multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods.

CFA = confirmatory factor analysis; MTMM = multitrait-multimethod; SOA = self-other-agreement.

3.1 Dealing With Variations of the Data Structure

In the analysis in Chapter 4, the direct supervisor's ratings were included in the analysis and there typically is exactly one such supervisor for every target manager. These ratings were therefore considered as a further structurally different method and modeled on level 2. However, there are situations where more than one boss delivers ratings for a given target manager (e.g., two bosses in Hoffman et al., 2010, and in Mount et al., 1998). In this case, it has to be carefully considered whether the multiple bosses are interchangeable or structurally different methods. If all bosses rating one target manager have the same rating perspective, e.g., because they are both located on the same hierarchical level in the organization and have the same access to the manager's behavior, they stem from one population. Their ratings are then interchangeable and should be located as a third set of level-1 methods in addition to peer and subordinate ratings. If the bosses have different rating perspectives, e.g., because there is one direct supervisor and one higher-ranking supervisor, they are not interchangeable but structurally different. They should be included as two additional level-2 methods.

A full 360-degree feedback assessment typically also considers ratings of clients, customers, or other external parties. If they need to be included in the analysis, the same questions as outlined above for boss ratings have to be answered to decide whether multiple clients of one target manager are interchangeable or structurally different. In most cases, the client ratings of one manager will be interchangeable and accordingly are an additional set of level-1 methods.

One data situation often encountered when using data from organizations that conducted a 360-degree feedback assessment is that some raters will deliver ratings for more than one person. There might for example be employees that deliver ratings for several of their peers, who are the target managers in the assessment. This is not consistent with the definition of the model as it violates the assumption of independent raters across targets (Meiser & Steinwascher, 2014; Schultze, Koch, & Eid, 2015). As a result the parameter estimates and standard errors are potentially biased if this nonindependence is not modeled (Schultze et al., 2015). Therefore, Koch et al. (2016) recently developed a cross-classified multilevel CFA-MTMM for structurally different and nonindependent interchangeable methods. It incorporates a latent interaction factor to take into account that the rater sets of different targets are (partly) overlapping. The concept of this approach can be adapted to the model presented in this thesis if there are peers or subordinates that rate more than one target manager. However, this extension leads to a very complex model and Koch et al. (2016) recommend using a Bayesian estimator which has not yet been used for the model presented in this thesis. It therefore should carefully be ascertained in further research whether this model extension is realizable.

There exist more possible scenarios of dependencies in the data set that violate the model assumptions. For example, an employee can be a target manager A and at the same time be a peer of target manager B, a subordinate of target manager C, and the boss of target manager D. Accordingly, this employee would have four roles and could deliver ratings from each of these roles in the assessment: Target manager A could provide self-ratings, ratings as a peer, ratings as a subordinate, and ratings as a boss. Unfortunately, it is not possible to identify if such a scenario exists in the data that was used for the analyses in Chapters 2 and 4, because there is no variable in the data set to identify the single raters. In order to identify and quantify this potential nonindependence of raters, it would be necessary and extremely helpful if this (anonymized) information was included in further assessments. However, there is, to my knowledge, no available strategy how to deal with such dependencies in a comparable CFA-MTMM framework besides the one proposed by Koch et al. (2016) for the special case of nonindependent interchangeable methods, that is, peers or subordinates who rate multiple target managers. Impulses for such a model could be derived from the statistical models used for social relations analyses of round-robin designs where each member of a group interacts and rates all other group members (Kenny, 1994; Olsen & Kenny, 2006; Schönbrodt, Back, & Schmukle, 2012; Snijders & Kenny, 1999).

The model presented can also be used in cross-cultural research on leadership. In this framework, the most typical research question is the analysis of SOA differences between countries (Atwater et al., 2005; Atwater et al., 2009; Eckert et al., 2010; Gentry et al., 2007; Gentry et al., 2010), but one could also test the measurement equivalence of a 360-degree feedback instrument across different countries (or cultures). There are two approaches to analyze cross-cultural research questions (Eid & Lischetzke, 2013). If there are few countries that are explicitly chosen to be compared, a multigroup model should be applied. This approach can directly be realized for the model presented in this thesis by including the country as a grouping variable in the software. It is then possible to constrain model parameters to be equal across countries and to test whether model fit suffers significantly.

The second approach applies if there are many countries included in the analysis and if these countries can be considered as a random sample of possible countries. In this case, the countries constitute another level of measurement and should be considered as a cluster (instead of a grouping) variable. Again, this strategy can directly be applied to the model presented in this thesis by defining a three-level model in the software. Peers and subordinates (level 1) are clustered in target managers (level-2 cluster variable). These target managers are clustered in countries (level-3 cluster variable). However, such a model is very complex and therefore probably requires large samples. Eid and Lischetzke (2013) stressed that there are no general rules for sample sizes applicable because the minimal number of countries depends on the complexity of the structural equation model. While the sample size requirements for level 1 and level 2 have been analyzed in Chapter 3, a further Monte Carlo simulation study is recommendable in order to define the number of countries that is needed before such a three-level model is applied.

Most companies that conduct a 360-degree feedback assessments aim to foster personnel development of their managers. Benchmarks®, the instrument that was used in Chapter 2 and 4, was explicitly designed to diagnose developmental needs (McCauley, Lombardo, & Usher, 1989). In order to investigate whether this change or development really occurs, the assessment often is repeated after some time has passed or is even conducted on a regular, e.g., yearly, basis. The analysis of change and stability in such repeated measurements requires a longitudinal modeling approach. Koch et al. (2014) recently proposed a longitudinal CFA-MTMM model for structurally different and interchangeable methods. This approach can be adapted to the model with one additional set of interchangeable methods presented in this thesis. However, as already pointed out for the extension to three-levels of measurement, the target model will be very complex and a simulation study should analyze the requirements on sample size.

4. Adapting the Model to Other Contexts

As pointed out in the introduction, there are many fields in which multisource feedback is assessed and can be considered as MTMM data. The new multilevel CFA-MTMM model can be applied whenever one wants to model data that includes one (or more) structurally different method(s) and two sets of interchangeable methods. Figure 1 shows some hypothetical applications and reveals that the model is not restricted to the analysis of 360-degree feedback data.

A concrete example for a possible adoption of the model to personality research is the analysis of Wessels, Zimmermann, Biesanz, and Leising (in press). The authors assessed self-

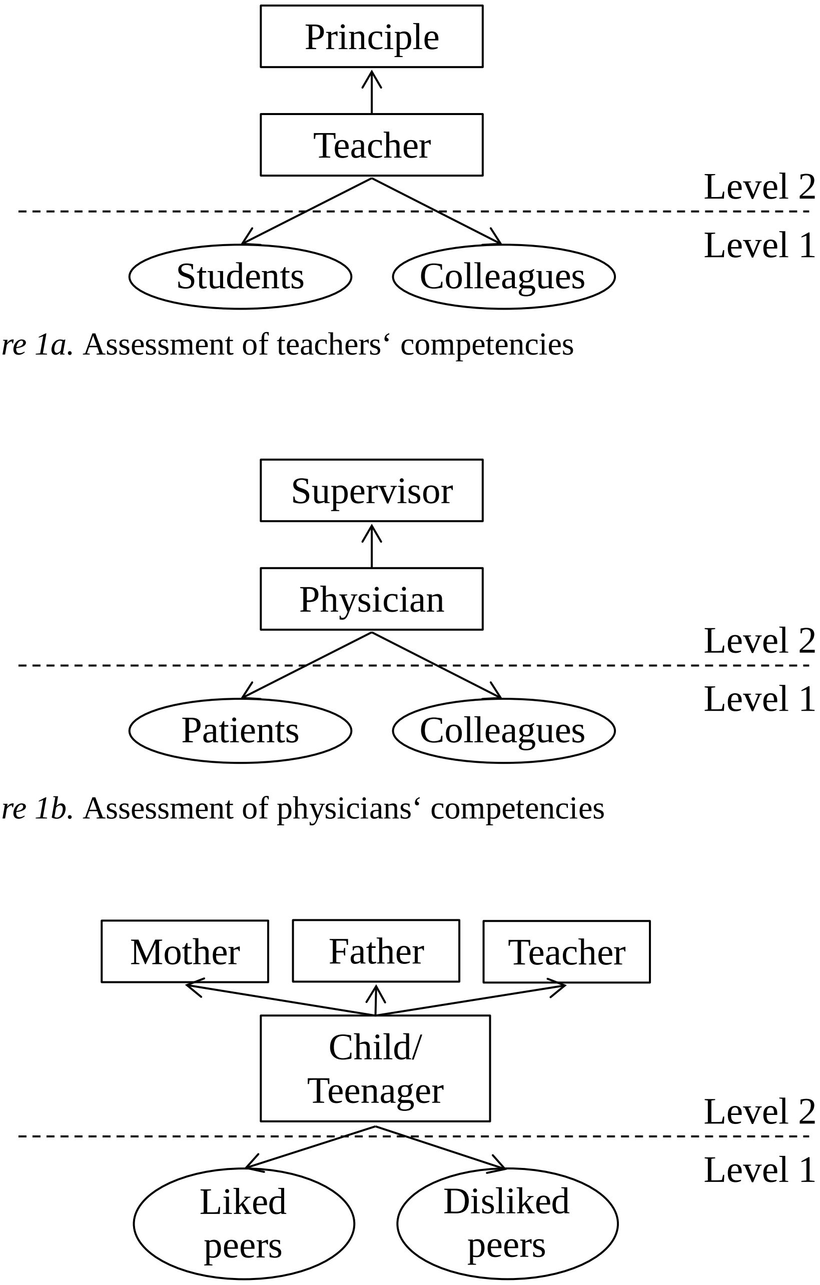

Figure 1a. Assessment of teachers‘ competencies

Figure 1b. Assessment of physicians‘ competencies

Figure 1c. Assessment of children‘s or teenagers‘ psychological attributes

Figure 1. Examples of possible multisource feedback applications with structurally different methods (boxes) and two sets of interchangeable methods (ellipses).

rated personality and others’ ratings of personality, stemming from two different groups of others. The first group is a group of fellow student informants and the second group is a group of target-nominated informants. These two different sets of informants’ ratings are predicted by the self-ratings in a multilevel profile analysis. Additionally, all informants rate the degree to which they know the target (knowing) and the degree to which they like the target (liking). These ratings are included as predictors as well as a dummy-coded variable to test for differential effects of the two groups of informants. The authors analyze the effects of these predictors on rating accuracy. The research questions addressed in this study could also be analyzed by using the multilevel CFA-MTMM model for one structurally different method and two sets of interchangeable methods.

In this modeling framework, the self-rated personality would be chosen as the reference method and would be located on level 2. The two sets of informants would both be modeled on level 1 as two distinct populations nested within targets. The method factors would reflect the degree of rating accuracy, i.e., agreement between self-ratings and others' ratings, on both levels. Latent regressions of these method factors on predictors such as knowing and liking can be included.

When taking a broader perspective of the modeling approach presented in this thesis, it is important to highlight that the proposed planned missing data format is applicable whenever there are two sets of level-1 populations – also in cases where no CT-C(M-1) structure is used. For example, the three models defined in Chapter 4 measure self-awareness without separating trait from method components, so that there is one level-2 factor for each method (self-rated self-awareness, boss ratings, common peer ratings, and common subordinate ratings of the managers' self-awareness) as well as a level-1 factor of unique peer ratings and a level-1 factor of unique subordinate ratings. This makes it possible to include peers and subordinates as two different level-1 populations by using the planned missing data format while no CT-C(M-1) model is defined on level 2. This perspective on the planned missing data modeling approach extends the scope of possible applications considerably as it addresses researchers applying various structural equation models. However, it should be kept in mind that the results of the Monte Carlo simulation study in Chapter 3 apply only to the model that was specified in the simulation. Using a different structural equation model will, for example, probably yield different requirements on sample size.

5. Limitations

The main limitation of the proposed model concerns the test of model fit. The Monte Carlo simulation study (Chapter 3) has revealed that the χ^2-test of model fit is too lenient and should not be trusted. As the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis-Index (TLI) and the Comparative Fit Index (CFI) are based on χ^2-values (West, Taylor, & Wu, 2012), these indices should also not be considered. It was therefore recommended in Chapter 3 to use the Standardized Root Mean Square Residual (SRMR). This fit index quantifies the difference between the observed and the model-implied covariances and is not affected by χ^2-bias. It is separately reported for level 1 and for level 2. However, as thoroughly described in Chapter 3, the planned missing data format makes the software use a wrong number of degrees of freedom on level 1. Therefore, a corrected SRMR for level 1 ($\mathrm{SRMR}_{\mathrm{L1}}^{\mathrm{adj.}}$) has been proposed in Chapter 3. The recommendation to use the $\mathrm{SRMR}_{\mathrm{L1}}^{\mathrm{adj.}}$ and the $\mathrm{SRMR}_{\mathrm{L2}}$ is not based on simulation results, though. It is not possible to draw conclusions on the appropriateness of these fit indices because their sampling distribution is (in contrast to the one of the χ^2-index) unknown. In order to analyze the appropriateness of the $\mathrm{SRMR}_{\mathrm{L1}}^{\mathrm{adj.}}$ and the $\mathrm{SRMR}_{\mathrm{L2}}$ and to quantify potential bias, it would be necessary to fit misspecified models to the simulated data.

The reason for the χ^2-bias, yet, is only partly known. Correcting the degrees of freedom reduces the bias between the expected and the observed χ^2-distribution. However, the bias remains substantial. Model decisions based on the χ^2-test statistic will result in too many falsely accepted models and reduced power. This is challenging for everyone who wants to apply the model because it remains a focal objective in the application of structural equation models to show that the selected model is able to reproduce the observed data (West, Taylor, & Wu, 2012). An explanation for and a strategy how to avoid the bias are therefore urgently needed. The question arises whether this bias is specific to the model proposed in this thesis and due to the planned missing data format or if it is also present in other structural equation models. Searching the literature reveals that other authors encountered the same bias (e.g., Koch et al., 2014; Ulitzsch, Holtmann, Schultze, & Eid, 2017).

The most promising idea is that it is associated to the unidimensionality of the common method factors. In the simulated model, the target-specific level-2 variables are unidimensionally captured by a common method factor. At the same time, their residual variances are fixed to zero, implying perfectly correlated variables. The model that was simulated by Koch et al. (2014) also used perfectly correlated latent variables and concluded that the χ^2-test was too liberal. Ulitzsch et al. (2017) found χ^2-bias in a model with unidimensional trait factors but not in a model with indicator-specific trait factors. An explanation for these findings can be derived from Stoel, Garre, Dolan, and Van Den Wittenboer (2006). The authors demonstrated that the χ^2-distribution is shifted leftward when parameters are constrained to a boundary value, e.g. a correlation of -1 or +1. It is argued that in structural equation models with boundary values the underlying distribution of the likelihood ratio test does not follow a central χ^2-distribution with the traditionally assumed degrees of freedom. A stepwise procedure to obtain the correct distribution by conducting a Monte Carlo simulation is proposed.

This explanation and the comparable results of other authors (Koch et al., 2014; Ulitzsch et al., 2017) suggest that the χ^2-bias is not associated to the special planned missing data format but is present in all models that explicitly or implicitly impose boundary parameters such as correlations of one, e.g., by defining unidimensional trait factors. Further research is urgently needed here. Especially, the strategy proposed by Stoel et al. (2006) should be adapted to the multilevel CFA-MTMM model for structurally different methods and two sets of interchangeable methods. On the one hand, this is necessary to check whether the approach is successfully applicable for this model. On the other hand, researchers who want to apply the model and need to define its fit would benefit from an exemplarily application.

Another issue concerns the model estimation process with missing values. The missing by design format that was imposed to take account of the special data structure is a missing completely at random (MCAR) condition (Little & Rubin, 2002). The probability that an observation is missing on a peer or subordinate variable *X* depends neither on the values of the variable *X* itself nor on the values of other variables in the data set (Rubin, 1976). A FIML estimator was used because it has been shown to yield unbiased estimates under MCAR conditions (e.g., Enders, 2001) and the results of the simulation study support this approach (Chapter 3). However, throughout the analyses in this thesis, there were no other missing values than the ones that were imposed by the special missing data structure of this model. All raters that were included in the analyses had available values on all parcels or items. This strategy was chosen in order to not confound two sources of missingness. Therefore, it is not yet known if there is a biasing effect of additional (nonplanned or "true") missingness in the model. As Enders (2001) concluded that the FIML method yields unbiased estimates under MCAR or missing at random (MAR) conditions (i.e., missingness of variable *X* depends on other variables in the data set, but not on the values of the variable *X* itself), no bias is expected when the additional missingness is under MCAR or MAR assumptions. However, the effect of combining two sources of missingness should be further analyzed, e.g., by means of a Monte Carlo simulation.

Last, it is important to me to point out a potential antagonism that I encountered throughout the work on this thesis. On the one hand, the proposed model is complex. Using it or even decoding its result in detail requires deepened statistical knowledge. On the other hand, the model was developed as a tool for those who are dedicated to the applied research in the field of leadership and 360-degree feedback assessment. This target audience necessarily has its expertise in other domains than the one of statistical modeling. When the manuscripts in Chapters 2 and 3 were submitted to journals with an applied (instead of statistical) focus, anonymous reviewers constantly had difficulties with understanding the modeling approach. Taking the reviewers' perspective, these difficulties are comprehensible and required repeated rephrasing of the manuscripts. Nonetheless, the two chapters that have been published (Chapter 2) or accepted (Chapter 3) so far are placed in statistical journals (*Journal of Structural Equation Modeling: A Multidisciplinary Journal* and *British Journal of Mathematical and Statistical Psychology*), so that the reach among the target audience is probably not as broad as it would have been in an applied journal.

However, the development of new statistical methods is never an end in itself. Methods are developed to be applied. I therefore claim that both sides should take the challenge: Those who develop a new approach should undertake best efforts to present it in a way that the target audience can understand and adopt it. Comprehensible model presentation, illustrative applications, and provision of software code are needed. Additionally, guidelines when to choose which model are helpful for those who are overwhelmed by the magnitude of available models. Those who aim to answer research questions in their applied field, on the other hand, should be ambitioned to keep up to date with methodological and statistical developments. They should make an effort to adopt new approaches – even if they are complex – so that their research benefits from statistical progress. I hope that my thesis meets these requirements and will help others to drive research and scientific findings forward.

References

Antonioni, D. (1996). Designing an effective 360-degree appraisal feedback process. *Organizational Dynamics, 25,* 24–38. doi:10.1016/S0090-2616(96)90023-6

Atkins, P. W., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55,* 871–904. doi: 10.1111/j.1744-6570.2002.tb00133.x

Atwater, L. E., 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. E., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings of leadership. *Personnel Psychology, 48,* 35–59. doi:10.1111/j.1744-6570.1995.tb01745.x

Atwater, L. E., Waldman, D., Ostroff, C., Robie, C., & Johnson, K. M. (2005). Self-other agreement: Comparing its relationship with performance in the U.S. and Europe. *International Journal of Selection and Assessment, 13,* 25–40. doi:10.1037/a0014561

Atwater, L. E., 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,* 876–886. doi:10.1037/a0014561

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45,* 141–164. doi:10.1111/j.1744-6570.1992.tb00848.x

Atwater, L. E., & Yammarino, F. J. (1997). Self-other rating agreement. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 15, pp. 121−174). Greenwich, CT: JAI Press.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6,* 15–43. doi:10.1177/1094428102239424

Baumeister, R. F. (2005). *The cultural animal: Human nature, meaning, and social life.* New York, NY: Oxford University Press.

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

Bergner, S., Davda, A., Culpin, V., & Rybnicek, R. (2016). Who overrates, who underrates? Personality and its link to self-other agreement of leadership effectiveness. *Journal of Leadership & Organizational Studies, 23,* 335–354. doi:10.1177/1548051815621256

Berk, R. A. (2009). Using the 360 multisource feedback model to evaluate teaching and professionalism. *Medical Teacher, 31,* 1073–1080. doi:10.1080/01421590802572775

Berson, Y., & Sosik, J. J. (2007). The relationship between self-other rating agreement and influence tactics and organizational processes. *Group and Organization Management, 32,* 675–698. doi:10.1177/1059601106288068

Besco, R. O., & Lawshe, C. H. (1959). Foreman leadership as perceived by superiors and subordinates. *Personnel Psychology, 12,* 573–582. doi:12. 10.1111/j.1744-6570.1959.tb01344.x

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

Bratton, V. K., Dodd, N. G., & Brown, F. W. (2011). The impact of emotional intelligence on accuracy of self-awareness and leadership performance. *Leadership & Organization Development Journal, 32*, 127–149. doi:10.1108/01437731111112971

Brutus, S., Fleenor, J. W., McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development, 18,* 417–435. doi:10.1108/02621719910273569

Campbell, J. J., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). Managerial behavior, performance, and effectiveness. New York, NY: McGraw-Hill.

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-multimethod 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,* 181–207. doi:10.1080/10705511.2013.769385

Church, A. H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 82,* 281–292. doi:10.1037/0021-9010.82.2.281

Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360-degree feedback: Guest editors' comments on the research and practice of multirater assessment methods. *Group & Organization Management, 22,* 149–161. doi:10.1177/1059601197222002

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22,* 139–162. doi:10.1177/014920639602200106

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

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74,* 68–80. doi:10.1037/h0029382

Diefendorff, J. M., Silverman, S. B., & Greguras, G. J. (2005). Measurement equivalence and multisource ratings for non-managerial positions: Recommendations for research and practice. *Journal of Business and Psychology, 19,* 399–425. doi:10.1007/s10869-004-2235-x

Donnon, T., Al Ansari, A., Al Alawi, S., & Violato, C. (2014). The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. *Academic Medicine, 89,* 511–516. doi:10.1097/ACM.0000000000000147

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

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology*, *46,* 641–665. doi:10.1111/j.1744-6570.1993.tb00889.x

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, *58*, 51–100. doi:10.1006/obhd.1994.1029

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes, 64*, 307–324. doi:10.1006/obhd.1995.1108

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow & N. W. Schmitt (Eds.), *Advances in measurement and data analysis* (pp. 350–400). San Francisco, CA: Jossey-Bass.

Edwards, J. R. (2009). Latent variable modeling in congruence research: Current problems and future directions. *Organizational Research Methods, 12,* 34–62. doi:10.1177/1094428107308920

Edwards, M. R., & Ewen, A. J. (1996). How to manage performance and pay with 360-degree feedback. *Compensation and Benefits Review, 28*(3), 41–46. doi:10.1177/088636879602800308

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

Eid, M., Geiser, C., & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science, 25,* 275–280. doi:10.1177/0963721416649624

Eid, M., & Lischetzke, T. (2013). Statistische Methoden der Auswertung kulturvergleichender Studien. In P. Genkova, T. Ringeisen & F.T.L. Leong (Eds.), *Handbuch Stress und Kultur: interkulturelle und kulturvergleichende Perspektiven* (pp. 189–206). Wiesbaden, Germany: VS.

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 CTC(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

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics, 3,* 1–21. doi:10.2307/3001534

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 128–141. doi:10.1207/S15328007SEM0801_7

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, *86,* 215–227. doi:10.1037/0021-9010.86.2.215

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *The Leadership Quarterly, 7,* 487–506. doi: 10.1016/S1048-9843(96)90003-X

Fleenor, J. W., Smither, J. W., Atwater, L. E., 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

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational and Organizational Psychology, 73,* 303–319. doi:10.1348/096317900167047

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

Gentry, W. A., Hannum, K. M., Ekelund, B. Z., & de Jong, A. (2007). A study of the discrepancy between self- and observer-ratings on managerial derailment characteristics of European managers. *European Journal of Work and Organizational Psychology, 16,* 295–325. doi:10.1080/13594320701394188

Gentry, W. A., Yip, J., & Hannum, K. M. (2010). Self-observer rating discrepancies of managers in Asia: A study of derailment characteristics and behaviors in Southern and Confucian Asia. *International Journal of Selection and Assessment, 18,* 237–250. doi:10.1111/j.1468-2389.2010.00507.x

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

Hensel, R., Meijers, F., van der Leeden, R., & Kessels, J. (2010). 360 degree feedback: How many raters are needed for reliable ratings on the capacity to develop competences, with personal qualities as developmental goals? *The International Journal of Human Resource Management, 21,* 2813–2830. doi:10.1080/09585192.2010.528664

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

Hoffman, B. J., & Woehr, D. J. (2009). Disentangling the meaning of multisource performance rating source and dimension factors. *Personnel Psychology, 62*, 735–765. doi:10.1111/j.1744-6570.2009.01156.x

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133. doi:10.1007/bf02291393

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

Kenny, D. A. (1994). *Interpersonal perceptions: A social relations analysis.* New York, NY: Guilford.

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

Koch, T., Kelava, A., & Eid, M. (2018). Analyzing different types of moderated method effects in confirmatory factor models for structurally different methods. *Structural Equation Modeling: A Multidisciplinary Journal, 25,* 179–200. doi:10.1080/10705511.2017.1373595

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

Koch, T., Schultze, M., Jeon, M., Nussbeck, F. W., Praetorius, A.-K. & Eid, M. (2016). A cross-classified CFA-MTMM model for structurally different and nonindependent interchangeable methods. *Multivariate Behavioral Research, 51*, 67–85. doi:10.1080/00273171.2015.1101367

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology, 51*(5p1), 369–381. doi:10.1037/h0025095

Lee, G. G. (2015). Caution required: Multirater feedback in the army. *Military Review*, *95*(4), 58–67.

Lepsinger, R., & Lucia, A. D. (2009). *The art and science of 360 degree feedback* (2nd ed.). San Francisco, CA: Wiley.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Lombardo, M. M., McCauley, C. D., McDonald-Mann, D., & Leslie, J. B. (1999). *Benchmarks® developmental reference points.* Greensboro, NC: Center for Creative Leadership.

London, M., & Smither, J. W. (1995). Can multisource feedback change self-awareness and behavior? Theoretical applications and directions for research. *Personnel Psychology*, *48*, 803–840. doi:10.1111/j.1744-6570.1995.tb01782.x

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7*, 19–40. doi: 10.1037//1082-989X.7.1.19

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

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* (pp. 177–198). Thousands Oaks, CA: Sage.

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

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,* 1–3. doi:10.3389/fpsyg.2014.00615

Mersman, J. L., & Donaldson, S. I. (2000). Factors affecting the convergence of self-peer ratings on contextual and task performance. *Human Performance, 13,* 299–322. doi:10.1207/s15327043hup1303_4

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. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37*, 687-702. doi: 10.1111/j.1744-6570.1984.tb00533.x

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,* 557–576. doi:10.1111/j.1744-6570.1998.tb00251.x

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Olsen, J. A., & Kenny, D. A. (2006). Structural equation modeling with interchangeable dyads. *Psychological Methods, 11,* 127–141. doi:10.1037/1082-989X.11.2.127

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology, 57,* 333–375. doi:10.1111/j.1744-6570.2004.tb02494.x

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

Rigdon, E. E. (1994). Demonstrating the effects of unmodeled random measurement error. *Structural Equation Modeling: A Multidisciplinary Journal, 1,* 375–380. doi:10.1080/10705519409539986

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592. doi:10.1093/biomet/63.3.581

Ruch, W., Köhler, G., & van Thriel, C. (1996). Assessing the temperamental basis of the sense of humor: Construction of the facet and standard trait forms of the State-Trait Cheerfulness Inventory-STCI. *International Journal of Humor Research, 9,* 303–339. doi:10.1515/humr.1996.9.3-4.303

Schmitt, N., & Stults, D.M. (1986). Methodological review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10,* 1–22. doi:10.1177/014662168601000101

Schönbrodt, F.D., Back, M.D. & Schmukle, S.C. (2012). TripleR: An R package for social relations analyses based on round-robin designs. *Behavior Research Methods, 44,* 455–470. doi:10.3758/s13428-011-0150-4

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

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

Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships, 6,* 471–486. doi: 10.1111/j.1475-6811.1999.tb00204.x

Stoel, R. D., Garre, F. G., Dolan, C., & Van Den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11,* 439–455. doi: 10.1037/1082-989x.11.4.439

Stone, L. L., Otten R., Engels, R. C. M. E., & Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review, 13,* 254–274. doi:10.1007/s10567-010-0071-2

Ulitzsch, E., Holtmann, J., Schultze, M., & Eid, M. (2017). Comparing multilevel and classical confirmatory factor analysis parameterizations of multirater data: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 24,* 80-103. doi:10.1080/10705511.2016.1251846

Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management, 32,* 249–263. doi:10.1002/hrm.3930320205

Vecchio, R. P., & Anderson, R. J. (2009). Agreement in self-other ratings of leader effectiveness: The role of demographics and personality. *International Journal of Selection and Assessment, 17,* 165–179. doi: 10.1111/j.1468-2389.2009.00460.x

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. (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

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (in press). *Differential associations of knowing and liking with accuracy and positivity bias in person perception.* doi:10.31234/osf.io/5dg94

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). New York, NY: Guilford.

Wherry, R. J. & Fryer, D. H. (1949). Buddy ratings: Popularity contest or leadership criteria? *Personnel Psychology, 2,* 147–159. doi:10.1111/j.1744-6570.1949.tb01395.x

Williams, S. B., & Leavitt, H. J. (1947). Group opinion as a predictor of military leadership. *Journal of Consulting Psychology, 11,* 283–291. doi:10.1037/h0056512

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

Wohlers, A. I., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42,* 235–261. doi:10.1111/j.1744-6570.1989.tb00656.x

Yammarino, F. J. (1998). Multivariate aspects of the varient/WABA approach: A discussion and leadership illustration. *The Leadership Quarterly, 9,* 203–227. doi:10.1016/S1048-9843(98)90005-4

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. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management, 32,* 231–247. doi:10.1002/hrm.3930320204

Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as other 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

Yammarino, F. J., & Markham, S. E. (1992). On the application of within and between analysis: Are absence and affect really group-based phenomena? *Journal of Applied Psychology, 77,* 168–176. doi:10.1037/h0090363

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics: Handbook of Statistics* (Vol. 26, pp. 45–79). Amsterdam, The Netherlands: Elsevier.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, 01.12.2018 Jana Mahlke