Kent Academic Repository

Full text document (pdf)

Citation for published version

Brown, Anna and Inceoglu, Ilke and Lin, Yin (2016) Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. Organizational Research Methods, 20 (1). pp. 121-148. ISSN 1094-4281.

DOI

https://doi.org/10.1177/1094428116668036

Link to record in KAR

http://kar.kent.ac.uk/56849/

Document Version

Author's Accepted Manuscript

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version.
Users are advised to check **http://kar.kent.ac.uk** for the status of the paper. **Users should always cite the published version of record.**

Enquiries
For any further enquiries regarding the licence status of this document, please contact:
researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at **http://kar.kent.ac.uk/contact.html**

KAR Kent Academic Repository

University of Kent

Preventing Rater Biases in 360-Degree Feedback by Forcing Choice

Anna Brown

University of Kent

Ilke Inceoglu

University of Surrey

Yin Lin

University of Kent, CEB SHL Talent Measurement Solutions

Author Note

Anna Brown is Senior Lecturer in Psychological Methods and Statistics at School of Psychology, University of Kent, Canterbury, Kent, CT2 7NP, United Kingdom. Tel: +44 (0)1227 823097. E-mail: a.a.brown@kent.ac.uk.

Ilke Inceoglu is Senior Lecturer in Organizational Behavior and HRM and Associate Dean (Research) at Surrey Business School, University of Surrey, Guildford, GU2 7XH, Surrey, United Kingdom. Tel: +44 (0)1483 682018. Email: i.inceoglu@surrey.ac.uk.

Yin Lin is PhD Researcher at School of Psychology, University of Kent, and Senior Research Scientist at CEB SHL Talent Measurement Solutions, 1 Atwell Place, Thames Ditton, KT7 0NE, Surrey, United Kingdom. Tel: +44 (0)0208 3358000. Email: yin.lin@shl.com.

Correspondence concerning this article should be addressed to Anna Brown, School of Psychology, University of Kent, Canterbury, Kent, CT2 7NP, United Kingdom.
E-mail: a.a.brown@kent.ac.uk

Abstract

We examined the effects of response biases on 360-degree feedback using a large sample (N=4,675) of organizational appraisal data. Sixteen competencies were assessed by peers, bosses and subordinates of 922 managers, as well as self-assessed, using the Inventory of Management Competencies (IMC) administered in two formats – Likert scale and multidimensional forced choice. Likert ratings were subject to strong response biases, making even theoretically unrelated competencies correlate highly. Modeling a latent common method factor, which represented non-uniform distortions similar to those of "ideal-employee" factor in both self- and other assessments, improved validity of competency scores as evidenced by meaningful second-order factor structures, better inter-rater agreement, and better convergent correlations with an external personality measure. Forced-choice rankings modelled with Thurstonian IRT yielded as good construct and convergent validities as the bias-controlled Likert ratings, and slightly better rater agreement. We suggest that the mechanism for these enhancements is finer differentiation between behaviors in comparative judgements, and advocate the operational use of the multidimensional forced-choice response format as an effective bias prevention method.

Keywords: multisource feedback, halo effect, rater biases, forced choice, Thurstonian IRT model, ideal-employee factor

Preventing Rater Biases in 360-Degree Feedback by Forcing Choice

Three-hundred-and-sixty degree appraisals are widely used in organizations, and the basic idea is to capture distinct perspectives on a set of employee behaviors thought to be important in their job roles. For feedback emerging from such process to be useful, it must converge across raters of the same target (although some differences are expected and even welcome), and it must be differentiated by behavior. Unfortunately, the opposite is often the case – assessments of conceptually distinct behaviors by the same rater are too similar (Murphy, Jako & Anhalt, 1993), while assessments of the same behavior by different raters are too diverse (Adler et al., in press; Borman, 1997; Conway & Huffcutt, 1997). Furthermore, rater perspective (e.g. subordinates versus peers of target X) explains little of the typically observed overall variability in 360-degree ratings (e.g. Scullen, Mount & Goff, 2000). Thus, there seems to be little value in the usual practice of separating information from specific rater perspectives (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). Although some plausible explanations of the low rater agreement have been suggested over time, such as different observability of behaviors (Warr & Bourne, 2000), rater personality (Randall & Sharples, 2012), differences in organizational level (Bozeman, 1997) and cultural values (Eckert, Ekelund, Gentry, & Dawson, 2010), many researchers have serious doubts about the validity of 360-degree ratings. For example, summarizing articles of an ORM special issue dedicated to modern data analytic techniques of 360-degree feedback data, Yammarino (2003, p. 9) concluded that “construct validity of multisource ratings and feedback is faulty or at least highly suspect”.

In the present article, we argue that the major issue often overlooked in studies using multi-source assessments is the problem presented by response biases. Many studies apply measurement assuming that only substantive constructs and random error sources influence

responses, and ignore systematic sources of error (biases). Studies that do consider biases tend to either examine effects of one specific type of bias (e.g. rater leniency; Barr and Raju, 2003) however small it may be, or, conversely, to assess the overall "rater idiosyncratic" variance overlooking the nature and type of biases contributing to it (e.g. Scullen, Mount, & Goff, 2000). Both directions of research are valuable; however, it is important to know what types of bias are likely to have the greatest impact on ratings in a specific context. To address this question, we use a large dataset of operational 360-degree appraisals to search for most potent biasing effects in multiple-rater assessments empirically, and assess two ways of overcoming them – by modeling biases after they occurred, and by preventing them from occurring in the first place.

The paper is organized as follows. First, we define the type of evaluations that can be reasonably expected in 360-degree appraisals. Second, we discuss potential threats to this objective, namely response biases, and outline two conceptually distinct ways of overcoming biases – modeling them statistically after the event, and preventing them with forced-choice response formats. We then identify the nature of biasing effects found in our empirical data, and evaluate the two alternative approaches to bias control by comparing the scores obtained with these methods in terms of their construct validity. We conclude with recommendations for research and practice.

Constructs Measured in 360-Degree Appraisals

Three-hundred-and-sixty degree assessments typically aim to measure competencies, commonly understood as "sets of behaviors that are instrumental in the delivery of desired results or outcomes" (Bartram, Robertson, & Callinan, 2002; p. 7). The important question is then whether and to what extent self- and other ratings measure behaviors of targets. We concur with Van Der Heijden and Nijhof's (2004) conclusion that raters certainly do not contribute

objective information on a target's behaviors. Instead, **subjective** evaluations of behaviors are being captured, which can be reasonably considered the intended constructs in 360-degree feedback as they have a value in themselves. Thus, the validity of the method can be defined as the extent to which 360 ratings reflect actual perceptual judgements of the rater (or recall of actual behaviors; Keller Hansborough, Lord, & Schyns, 2015), as opposed to nuisance factors, which have the potential to contribute to both random and systematic errors in ratings.

Response Biases in 360-Degree Appraisals and Ways to Counteract Them

The literature on 360-degree assessments has predominantly focused on Likert-type question formats (also referred to as single-stimulus formats). A multitude of biases elicited by the use of single-stimulus questions is well known and documented. Both self-reported and other-reported questionnaires can be subject to response styles. Acquiescence, often dubbed "yeah-saying", is a response style primarily caused by inattention or lack of motivation (Meade & Craig, 2012); therefore, it is unlikely to be a major threat in organizational appraisals, which typically constitute medium-stakes assessments. Extreme /central tendency responding is the individual tendency to use primarily the extreme / middle response options, which is thought to be related to personality and culture (Van Herk, Poortinga, & Verhallen, 2004). Specific to ratings by external observers is the leniency / severity bias, where some raters are lenient and some are severe in their ratings of all targets (e.g. Murphy & Balzer, 1989; Murphy & Cleveland, 1995). Barr and Raju (2003) investigated the leniency bias in 360-degree feedback, and found that despite statistical significance, its effect size on observed ratings was small.

Judging by extant research, much more potent in multiple-rater assessments is the tendency to maintain cognitive consistency in ratings of all behaviors guided by affect felt toward the target – the so-called 'halo' effect (Thorndike, 1920; Kahneman, 2011). This

unmotivated but powerful cognitive bias results in high dependencies between all assessed qualities – even conceptually unrelated ones. Moreover, external raters may be influenced by different goals and political pressures (Murphy, Cleveland, Skattebo & Kinney, 2004), while the assessment targets may be keen to present a picture of an "ideal employee" (Schmit & Ryan, 1993) – all examples of motivated processes. Unmotivated or motivated, response distortions in organizational appraisals may render ratings invalid – because we cannot assume that perceptual judgments of behavior are measured.

Statistical correction of response biases. One way to overcome response biases has been the application of statistical correction after distortions have taken place. In the past, a commonly used approach was to quantify the biasing effect by calculating a simple index (for example, the number of times a person used the extreme response options), which was then used to partial out the effect from the observed scores (Webster, 1958). Advances in item response modeling have allowed controlling for some types of response distortions by incorporating them in the response model (e.g. Böckenholt, 2012, 2014; Bolt & Newton, 2011; Bolt, Lu, & Kim, 2014; Maydeu-Olivares & Coffman, 2006). For example, if a bias can be conceptualized as a random additive effect f, the psychological values (or utilities) that a respondent expresses for items A and B, t'_A and t'_B, are a combination of "true" utilities t_A and t_B that the respondent feels for these items, and the added effect:

$$t'_A = t_A + \lambda_A f, \qquad t'_B = t_B + \lambda_B f \ . \tag{1}$$

Such processes can be modeled in a confirmatory factor analysis (CFA) framework, whereby a latent variable f influences all item responses, over and above any common factors that are assumed to underlie the true item utilities. The extent to which item A is sensitive to the biasing effect f is described by the respective factor loading, λ_A. Uniform response biases assume

factor loadings equal for all items; non-uniform biases assume varying loadings. For instance, acquiescence and rater leniency are uniform additive effects; they can be modelled by adding a common random intercept with equal loadings for all items (e.g. Barr & Raju, 2003; Maydeu-Olivares & Coffman, 2006). The "ideal-employee" factor as conceptualized by Klehe at al. (2012), on the other hand, is a non-uniform additive effect whereby more desirable indicators show higher loadings.

Benefits of explicitly modeling biasing effects are twofold. First, the bias is conceptualized and measured as a construct, and can then be explored in relation to other psychological variables. For example, Klehe et al. (2012) controlled for score inflation often observed in high-stakes personality assessments by modeling an "ideal-employee" factor, and found that the construct's relationship with job performance was fully explained by the applicant's ability to identify performance criteria. This example also illustrates the second benefit of modelling bias: "purification" of the measured constructs, which now can be analyzed without the distortions clouding our understanding of their validity.

Prevention of response biases. Another way of overcoming biases has been employing research designs that would prevent them from occurring in the first place. One popular design is collecting data using forced-choice formats, whereby a number of items are presented together in blocks, and respondents are asked to rank the items within each block. The format makes it impossible to endorse all items, thereby counteracting acquiescence and improving differentiation (thus directly combating halo effects). Moreover, since direct item comparison requires no rating scale, extreme / central tendency response styles cannot occur.

It has been shown that the forced-choice format eliminates all effects acting uniformly across items under comparison (Cheung & Chan, 2002). The mechanism for this elimination is

very simple. According to Thurstone's (1927) law of comparative judgment, choice between items A and B is determined by the difference of utilities that a respondent feels for A and for B, $t_A - t_B$. If this difference is positive, $t_A - t_B > 0$, then A is preferred to B, and if it is negative, $t_A - t_B < 0$, then B is preferred to A. Because the outcome only depends on the sign of the utility difference, it is invariant to any transformation of the utilities as long as their difference remains of the same sign. For example, if item utilities are biased so that the expressed utilities t' are linear combinations of the true utilities t with fixed coefficients c and d,

$$t'_A = ct_A + d, \qquad t'_B = ct_B + d, \tag{2}$$

then the difference of the "biased" utilities has the same sign as the difference of the true utilities when $c > 0$,

$$t'_A - t'_B = \left(ct_A + d\right) - \left(ct_B + d\right) = c\left(t_A - t_B\right). \tag{3}$$

It can be seen that any additive and multiplicative effects (terms d and c in Equation (2) respectively) are eliminated by the forced-choice format. Importantly, it is not necessary that the effects are uniform across **all** items in a questionnaire. It is sufficient that the coefficients c and d are constant within each block, but they can vary across blocks. This feature has been used by researchers as the basis for creating forced-choice designs robust to motivated response distortions such as impression management or faking (e.g. Stark, Chernyshenko & Drasgow, 2011). Indeed, if faking can be conceptualized as an additive effect f as in Equation (1), then careful matching of items within blocks on the extent they are susceptible to faking (i.e. on their factor loadings λ) should remove or greatly reduce the effect,

$$t'_A - t'_B = \left(t_A + \lambda_A f\right) - \left(t_B + \lambda_B f\right) = t_A - t_B + \left(\lambda_A - \lambda_B\right) f\ . \tag{4}$$

On the contrary, combining items with very different susceptibility to faking within one block, for example a positive indicator and a negative indicator of two desirable traits would predictably result in a failure to control faking (e.g., Heggestad, Morrison, Reeve & McCloy, 2006).

Previous research with forced-choice instruments in performance measurement (Bartram, 2007) demonstrated better discrimination between behaviors and improved predictor-criterion relationships. However, older studies employing the forced-choice format used the classical scoring method, yielding ipsative (relative to self, inter-personally incomparable) data, with its many spurious effects (Brown & Maydeu-Olivares, 2013) and hence the true effectiveness of the forced-choice method in 360-degree feedback is unknown.

Separating Substantive and Biasing Effects

In their critical analyses of the literature on halo bias, Murphy, Jako and Anhalt (1993) called for a moratorium on the use of "halo indices" derived from observed scores. One example of such index is the overall score on all measured dimensions, which Landy, Vance, Barnes-Farrell and Steele (1980) suggested to partial out of the dimension-specific scores to overcome halo. The problem with this approach, as Murphy and colleagues (1993) rightly identified, was that such indices cannot separate the cognitive bias of inflated coherence in judgements due to affect felt for a target (aka halo; or "illusory" halo as often specified in the literature) from the conceptual overlap between assessed traits due to competence in a wider domain or overall job competence (specified as "true" halo). While the former might be an interesting psychological phenomenon to study, it is nuisance to organizational appraisals of competencies. The latter, on the other hand, is probably one of the most important variables in organizational psychology, and removing it from data would amount to throwing the baby out with the bathwater.

Fortunately, huge advances that latent variable modeling has made in the last 20 years allowed researchers to move from indices based on observed scores to modelling biases as latent variables, for example assuming a model such as described by Equation (1). The question of such models' ability to separate the substantive and biasing effects is then one of **model identification**. Are the substantive factors underlying the true utilities t and the biasing factor f separately identified? Generally, one can control for "any systematic variance among the items independent of the covariance due to the constructs of interest" (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, page 894). This assumption of independence of the biasing factor from the substantive factors is quite reasonable in many contexts. However, the constructs of interest must be allowed to correlate with each other, which is necessary to capture the substantive overlap between them – for example, due to overall job competence[i]. A suitable model is presented in Figure 2 (panel 2), with a latent common method factor independent from multiple correlated constructs of interest. This model is generally identified, although empirical under-identification may occur – for instance, when the number of measured traits is large but the number of items measuring them is small (Podsakoff et al., 2003). The model allows capturing common method effects without specifying and measuring them; however, to identify several biasing effects that act simultaneously, further constraints and special research designs are required.

The use of forced choice has a good potential to reduce the inflated coherence ("illusory" halo) by directly forcing raters to differentiate. But can we ensure that at the same time, the "true" halo be retained? Or similarly, while making it impossible to endorse all desirable behaviors in self-assessment (and thus reducing the "ideal-employee" factor), can we ensure that any overlap due to genuine clustering of certain behaviors in the same people be retained? The traditional approach to scoring forced-choice data resulting in ipsative data made this objective

impossible. Ipsative scoring removed the person overall baseline (all dimension scores add to a constant for everyone) and consequently the common basis for scores overlap. One immediate consequence is that the average correlation between ipsative scores is always negative (Brown & Maydeu-Olivares, 2013). The same spurious effect is observed when subtracting the person total score from their dimension scores as suggested by Landy et al. (1980), which effectively ipsatizes the scores. Understandably controversial in the past, the forced-choice formats are rapidly gaining in popularity since the development of item response theory (IRT) models that infer proper measurement from them. It has been recently shown that, with the use of Thurstonian IRT models (Brown & Maydeu-Olivares, 2011) such as one illustrated in Figure 2 (panel 3), the still-widespread concern among researchers about removal of any common source of item variance by the forced-choice format can be finally put to rest. Brown (2016) demonstrated that under general conditions, the covariance structure underlying the measured traits is preserved in forced-choice data. Thus, forced choice can be used to remove or greatly reduce any common method effects according to Equation (4), while preserving factorial structures of substantive traits that are the focus of measurement.

Objectives and Hypotheses

The aim of the present study is to address some methodological limitations of previous research, and examine the effects of bias control and prevention on validity of 360-degree organizational appraisals. The specific objectives are to evaluate the forced-choice method, and to explicitly compare this bias prevention method to either “doing nothing” (i.e. scoring single-stimulus responses without any bias control), or statistically controlling for biases after they have occurred. We apply model-based measurement to enable proper scaling of competencies based on forced-choice (FC) assessments and thereby avoid ipsative data (Brown & Maydeu-Olivares,

2013), and to separate the method-related (biasing) factor from the substantive factors (i.e. factors we are interested in measuring) in single-stimulus (SS) assessments.

The default assumption in traditional scoring protocols of appraisal tools is that any similarities between observed competency ratings are due only to the substantive overlap between competencies (overall competence, or "true halo"). Based on previous research (e.g. Bartram, 2007; Landy & Farr, 1980; Scullen, Mount, & Goff, 2000), we contest this assumption and hypothesize that rater idiosyncratic biases will also play a significant role. More formally:

Hypothesis 1. When the SS format is used, behavior ratings will indicate not only their designated competency domains, but also a common method factor, which will explain a substantial proportion of the variance.

Based on consistent reports of the "ideal-employee" factor emerging in high and medium-stakes self-assessments (e.g. Klehe et al., 2012; Schmit & Ryan, 1993), and widely reported effects of exaggerated coherence ("illusory halo") in assessments by external observers (e.g. Kahneman, 2011; Landy et al., 1980; Murphy et al., 1993), we further hypothesize that the common method factor influencing SS ratings will be mostly due to these influences.

Hypothesis 2a. In SS self-assessments, the common method factor will reflect the extent of selective overreporting on behaviors associated with those of "ideal employee";

Hypothesis 2b. In SS assessments by others, the common method factor will reflect the extent of indiscriminate overreporting on all behaviors, due to cognitive bias of exaggerated coherence.

When biasing effects are present, either statistical control or prevention is necessary to reduce the irrelevant variance in measured competency constructs. Based on the previous reports of model-based bias control in the SS format (e.g. Böckenholt, 2014; Bolt, Lu, & Kim, 2014) and

bias prevention using the FC format (e.g. Bartram, 2007; Salgado & Táuriz, 2014), we predict that both methods will be effective in reducing the irrelevant variance and therefore increasing construct validity of measured competencies.

Hypothesis 3: Construct validities of competency scores based on either bias-controlled SS ratings or FC rankings will be better than those of straight SS ratings. Specifically,

H3a: Internal (factorial) structure of competency assessments will show better differentiation of behavioral domains;

H3b: Agreement between raters will increase, indicating improved convergent validity of rater assessments;

H3c: Convergent correlations of competency scores with similar external constructs will increase, while correlations with dissimilar constructs will stay low.

If the statistical control and prevention methods are shown to be effective, it would be of major importance to compare their effectiveness. To our knowledge, this has not been done before – instead, separate research studies compared either method to the straight SS ratings. Therefore, we have no specific hypotheses with regard to relative effectiveness of the two methods, and approach this question in exploratory fashion.

Method

Participants

Participants in this study were from 21 organizations located in the UK. This was a convenience sample, comprising archival appraisals data collected between 2004 and 2011. Of the assessed N = 922 managers, 65% were male, 92% identified as White. All working ages were represented, with the largest age groups being 30-34 years old (18.8%), 35-39 years old (18.1%) and 40-44 years old (17.2%). The best represented were not-for-profit organizations including

education, government and healthcare (67.5% of all participants), and private companies in finance and insurance (17% of all participants).

The managers were assessed on key competencies by 795 bosses, 1,149 peers and 1,857 subordinates, as well as 874 managers providing self-assessments (total N = 4,675 assessments). Not every target was assessed from all rater perspectives, including some absent self-ratings. The numbers of raters per target were variable, ranging from 0 to 3 for bosses, from 0 to 10 for peers, and from 0 to 12 for subordinates. Whenever a particular rater category was present for a target, the average number of boss raters was 1.08 per target, the average number of peers was 2.44, and the average number of subordinates was 2.59.

Measures

Inventory of Management Competencies (IMC). The Inventory of Management Competencies (IMC; SHL, 1997) was administered to all raters to seek assessments on 16 competency domains listed in Appendix. The IMC consists of 160 behavioral statements (for example, "identifies opportunities to reduce costs"), with 10 statements measuring each competency. The statements are presented in 40 blocks of four, with statements within one block indicating different competencies. Responses are collected using a unique response format, comprising single-stimulus (SS) ratings and forced-choice (FC) rankings. The SS format requires respondents to rate every statement in the block using a 5-point frequency rating scale ("hardly ever" – "seldom" – "sometimes" – "often" – "nearly always"). The FC format requires respondents to perform partial ranking of four statements in the block, selecting one statement as "most" representative of the target's behavior, and one statement as "least" representative of his/her behavior. Thus, the SS and FC formats are used in the IMC questionnaire side by side,

with every four items being rated and ranked in one presentation. Here is a sample block with example responses:

Mr John Smith …	Rating (SS)	Ranking (FC)
…is entrepreneurial	hardly ever	least
…draws accurate conclusions	often	most
…leads the team	often	
…produces imaginative solutions	seldom	

The IMC has been shown to be a robust instrument for measuring managerial competencies in multiple studies (e.g. Bartram, 2007; Warr & Bourne, 1999). The IMC scales yield internally reliable scores; reported alphas for the SS format in the manual (SHL, 1997) range between .83 and .91 (median .87). In the present study, alphas for the SS format ranged from .84 to .92 for self-assessments, from .89 to .94 for bosses, from .87 to .93 for peers, and from .86 to .93 for subordinates.

Occupational Personality Questionnaire (OPQ32). To examine the convergent and discriminant validity evidence for the IMC competency scores, we used self-reported personality assessments with the Occupational Personality Questionnaire available for N = 213 targets in the present study. The OPQ32 is a well-established measure of 32 work-relevant personal styles, with a wealth of materials on its reliability and validity available (Bartram, Brown, Fleck, Inceoglu, & Ward; 2006). The forced-choice OPQ32i version was used here, which consists of 104 blocks of four, with statements within one block indicating different personality traits. Responses are collected using partial ranking, selecting one statement in each block as “most” and one statement as “least” descriptive of self. Normally this version is associated with ipsative scores, for this study however raw responses were used to estimate Thurstonian IRT scores on the 32 traits for each target – the methodology yielding scores free of problems of ipsative data.

To this end, we used an approach that was later adopted for the successor of the OPQ32i, the OPQ32r and is published elsewhere (Brown & Bartram, 2009-2011).

Analyses

All analyses in this paper, unless stated otherwise, were performed in Mplus 7.2 (Muthén & Muthén, 1998-2015).

INSERT FIGURE 1 ABOUT HERE

Data considerations. In 360-feedback data, independent sampling required by many statistical tests cannot be assumed. Instead, targets are the primary sampling units, rater perspectives are the secondary sampling units, and the raters within the perspectives are the third-level sampling units. Figure 1 illustrates this nested structure. Here, individual assessors (level 1) are nested within rater perspectives (level 2), which are in turn nested within targets (level 3).

With nested data, two types of effects might be of interest. First, the researcher may be interested in pooled effects resulting from variation at all levels – variation due to idiosyncratic rater differences (level 1), due to differences between rater perspectives (level 2), and due to individual differences between targets (level 3). In the present study, the pooled effects were considered when deriving scores on the 360-degree appraisal tool (see the section "Fitting measurement models"). To account for non-independence of observations due to cluster sampling when computing standard errors and a chi-square test of model fit, the Mplus feature TYPE=COMPLEX was used (Asparouhov, 2006). Second, the researcher may be interested in separating the effects due to specific nesting levels. In the present study, the extent to which the different sampling levels (for example, being any rater of target X versus being a peer rater of

target X) influenced similarity of ratings was of interest. We used multilevel (three-level) modeling to this end (see the section "Evaluating rater agreement").

INSERT FIGURE 2 ABOUT HERE

Fitting measurement models to each rater perspective. To infer measurement from the observed responses of selves, bosses, peers and subordinates to SS and FC questions, we fitted appropriate CFA models with categorical variables (aka IRT models), using the Unweighted Least Squares estimator with robust standard errors (denoted ULSMV in Mplus). Since the IMC was designed to measure 16 related competencies, all measurement models comprised 16 correlated latent traits. In the SS format, the competencies were indicated by their respective item responses. In the FC format, the competencies were indicated by the respective latent utilities of items, which in turn were indicated by dummy coded rankings of items within blocks (see the subsections "SS response formats" and "FC response format" for detail).

To test the hypotheses, two alternative models were fitted to the SS responses, and one model was fitted to the FC responses. The models are illustrated schematically in Figure 2. The **SS model** comprising 16 correlated latent traits reflected the view whereby only substantive competency perceptions cause variability in SS ratings. The **SS-Method model,** comprising a common "Method" factor in addition to the 16 correlated latent traits, reflected the alternative view whereby both competency perceptions and common-to-all-items biases cause variability in SS ratings. The Method factor was assumed uncorrelated with the 16 competency factors; however, the competency factors were freely correlated to allow the substantive overlap between them (for instance, due to overall job competence). Finally, the **FC model,** comprising the 16

correlated latent traits, reflected the view that only competency perceptions cause variability in FC rankings. Specific details of the SS and FC measurement models are described below.

Model fit (here, how well the model reproduces the observed tetrachoric / polychoric correlations) was assessed by the chi-square test (χ^2), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). For RMSEA, the cut-off .05 has been suggested for close fit (MacCallum, Browne, & Sugawara, 1996). For SRMR, a value of .08 or less is considered indicative of an acceptable fit (Hu & Bentler, 1999).

SS response format. Since the single-stimulus response options comprised five ordered categories, Samejima's (1969) graded response model was used to link the item responses to their respective latent traits. Thus, for each item, four thresholds were estimated, and one factor loading on the respective competency factor; plus an additional factor loading on the Method factor in the SS-Method model.

FC response format. To fit Thurstonian IRT models to forced-choice IMC responses, we followed the standard procedure recommended for partial ranking data, and used a macro published by Brown and Maydeu-Olivares (2012) to create an Mplus syntax[ii]. First, rankings of four items (A, B, C, D) in each forced-choice block were coded as six binary dummy variables, representing all pairwise comparisons between items {A,B}, {A,C}, {A,D}, {B,C}, {B,D}, {C,D}. Each dummy variable was coded "1" if the first item in the pair was preferred to the second item and "0" otherwise. Since the outcome of comparison between two items not selected as "most" or "least" was not observed, it was coded missing. Here is an example partial ranking and corresponding dummy codes:

Partial ranking				Dummy variables					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
least	most			0	0	0	1	1	.

Overall, 240 binary dummy variables were created (6 per each of the 40 blocks). Each dummy variable {A, B} was modelled as categorization of the difference between two latent variables – the utility for item A, t_A, and the utility for item B, t_B. The item utilities, in turn, were modelled as the indicators of their respective competency factors. These features of the FC measurement model are illustrated in Figure 2.

When complete rankings are collected, fitting the FC model is almost as straightforward as fitting the SS model. However, because the IMC collects partial rankings, missing responses must be imputed to avoid bias in parameter estimation (for full detail of the missingness mechanism in partial rankings – MAR but not MCAR, and of the problem with estimating MAR missing data with limited information estimators such as Unweighted Least Squares used in this study, see Brown & Maydeu-Olivares, 2012). The multiple imputation method implemented in Mplus is based on a Bayesian estimation of an unrestricted model (i.e. model in which the observed dummy-coded binary variables are correlated freely), which is then used to impute the missing values (Asparouhov & Muthén, 2010). Ten imputed data sets (independent draws from the missing data posterior) were created, and the Thurstonian model was then fitted to each set. The estimated parameters in the 10 sets were averaged, yielding the final item parameters (factor loadings, thresholds, and error variances) and the latent trait parameters (i.e. means, variances, and covariances).

Exploring the latent trait covariance structures. To establish the structural validity of competency assessments in each measurement model, the correlations between the 16 latent competency traits were factor analyzed. Looking at the latent correlations provides the benefit of estimating the relationships between error-free theoretical constructs, rather than the attenuated

correlations between estimated scores. To establish the number of dimensions, optimal implementation of Parallel Analysis (Timmerman & Lorenzo-Seva, 2011) was used, which is based on comparison of the observed correlations to the 95th percentile of the randomly generated data with the same type of correlations (here, Pearson's), and the same type of underlying dimensions (here, factors). Then, Unweighted Least Squares factor analysis with oblique Oblimin rotation was carried out to find a simple structure with correlated factors. The SS, SS-Method, and FC latent competency correlations were examined in this fashion, using the software FACTOR 9.2 (Lorenzo-Seva & Ferrando, 2013).

Estimating competency scale scores. To enable analyses of rater agreement and convergent and discriminant validities (see further sections), it was necessary to work with estimated rather than latent competency scores[iii]. To place the scores from all the rater perspectives on the same scale, which is essential in computation of intraclass correlations as well as averaging scores from all the external observers (Woehr, Sheehan, & Bennett Jr., 2005), we fitted the SS, SS-Method and FC models described above using multiple-group CFA. Single-group models used so far are not suitable for these types of analyses since they reset the origin and unit of measurement to each of the rater perspectives. The technical detail of multiple-group CFA, sample Mplus syntax for conducting such analyses and scoring responses, and the results of multiple-group analyses of the present study are described in the downloadable Supplement to this article.

Model-based competency factor scores were estimated using the Bayes estimation maximizing the mode of the posterior distribution[iv] (Maximum a Posteriori, or MAP). Multivariate normal distributions of the 16 traits with covariances equal to the estimated values in the respective model were used as the prior. After the scoring procedures were applied to each

of the three measurement models (SS, SS-Method and FC), each target had three sets of competency scores for self-assessment (if completed); three sets of scores for each (if any) participating boss; three sets of scores for each (if any) participating peer; and three sets of scores for each (if any) participating subordinate.

Evaluating rater agreement. To quantify rater agreement on the traits measured by each of the alternative models, Intraclass Correlation Coefficients (ICC) of the estimated scale scores were computed at two hierarchical levels of nesting, as illustrated in Figure 1. The ICC is a measure of homogeneity among scores within units of analysis (Hox, 2010) – therefore it reflects the extent of absolute agreement between raters. In three-level models, the intraclass correlation is calculated based on the intercept-only model of assessment scores, which partition the total variance into three components – between targets (σ^2_{target}) at level 3, between perspectives on the same target ($\sigma^2_{\text{perspective}}$) at level 2, and between individual raters from the same perspective (σ^2_{rater}) at level 1. With this, the ICC at the highest nesting level 3 (assessment targets),

$$\rho_{\text{target}} = \frac{\sigma^2_{\text{target}}}{\sigma^2_{\text{target}} + \sigma^2_{\text{perspective}} + \sigma^2_{\text{rater}}}, \quad (5)$$

is an estimate of the population correlation between **two randomly chosen raters of the same target**. Because raters from the same perspective of the target are necessarily nested within the target, the ICC at level 2 (rater perspective),

$$\rho_{\text{perspective}} = \frac{\sigma^2_{\text{target}} + \sigma^2_{\text{perspective}}}{\sigma^2_{\text{target}} + \sigma^2_{\text{perspective}} + \sigma^2_{\text{rater}}}, \quad (6)$$

is an estimate of the population correlation between **two randomly chosen raters from the same perspective on the target** (Hox, 2010). We note that $\rho_{\text{perspective}}$ cannot be lower than ρ_{target}. The

larger the difference between the ICCs at two hierarchical levels, the more variation in ratings can be attributed to the rater perspective.

Computing convergent and discriminant validities of competency scores. To provide evidence for convergent and discriminant validity of the traits measured by the alternative IMC models, their correlations with personality traits as measured by the OPQ32 were computed. To match the IMC competencies with the conceptually concordant OPQ32 personality traits, each of the authors undertook independent reviews of the construct descriptions, followed by the panel discussion. Due to a substantial parallelism of both assessment tools, IMC and OPQ32, the matching procedure was straightforward. For each IMC competency, one or two most concordant personality construct measured by the OPQ were identified (for example, IMC Persuasiveness and OPQ Persuasive). All but one IMC competency (Written Communication) yielded at least one matching personality construct. For Resilience, the OPQ trait Tough Minded was hypothesized concordant with the self-assessments of Resilience, while OPQ trait Emotionally Controlled was hypothesized concordant with the others assessments (i.e. behavioral expressions of Resilience are likely to look as Emotional Control to others). For the matched Competency-Personality Trait pairs, the correlation coefficients were computed between the OPQ32 score and the IMC self-assessment for N=202 targets, and between the OPQ32 score and the average IMC assessment by external observers of N=208 targets.

Results

Hypothesis 1: Emergence of a Common Method Factor in SS Ratings

H1 proposed that in the SS response format, behavior ratings would indicate not only their designated competency domains but also a method factor common to all items. To test this hypothesis, we compared the parameters and goodness of fit of the two alternative measurement

models for rating scale data – the SS model, and the SS-Method model. If the SS-Method model, which assumes a common-to-all-items latent factor in addition to 16 correlated traits (competencies), fits the observations better than the SS model and the Method factor explains significant proportions of variance in responses, we have evidence for H1.

The SS and SS-Method models converged for all rater perspectives. Goodness of fit indices for the respective models[v] are given in Table 1. Despite the significant χ^2 statistics (not surprising considering the very large sample size in this study), both SS and SS-Method models had a good exact fit according to the SRMR, which was comfortably below.08 for all rater perspectives. The RMSEA values below.05 also suggested a good approximate fit for both models. Because the SS model is nested in the SS-Method model, we performed the difference of χ^2 tests. As can be seen from the $\Delta\chi^2$ results presented in Table 1, controlling for the Method factor improved the model fit. The improvement was not only highly statistically significant (which is easily achievable with large samples), but also practically important judging by the non-trivial improvements in the χ^2/df ratios, the RMSEA and particularly the SRMR, which is an absolute measure of correspondence between the observed and predicted polychoric correlations.

INSERT TABLE 1 ABOUT HERE

Appraising the SS-Method model parameters, all the item loadings on their respective competency factors were positive and significant. Importantly, all the Method factor loadings were also positive and significant, and of approximately the same magnitude on average as the substantive factor loadings (the factor loadings will be considered in more detail in the results related to Hypothesis 2). The competency factors explained slightly **more** variance than the Method factor in the average item for self-ratings (30% versus 25%) and boss ratings (35%

versus 29%), and slightly **less** variance than the Method factor in the average item for peer ratings (29% versus 32%) and subordinate ratings (24% versus 36%).

Taken together, the better goodness of fit of the SS-Method model and the substantial amount of variance explained by the Method factor comparable to the variance explained by competency factors, support Hypothesis 1. Controlling for the Method factor is indeed important in modeling SS data.

Hypothesis 2: Nature of the Common Method Factor

H2 proposed that the common Method factor influencing SS ratings would be mostly due to "ideal-employee" type distortion for selves, and due to exaggerated coherence ("illusory halo") for others. To see what construct the Method factor represented for each rater perspective, we explored the factor loadings of its indicators. The Method factor loadings were all positive, significant, but varied in magnitude greatly (the minimum unstandardized loadings were 0.3–0.4 depending on rater perspective, and the maximum were 1.4–1.7). The Method factor had non-uniform effect on item responses, thus unlikely representing rating style biases.

H2a: Self-ratings

For the self-perspective, items with the largest loadings on the Method factor were "Produces imaginative solutions" and "Generates imaginative alternatives" measuring Creativity & Innovation (unstandardized loadings 1.41 and 1.31 respectively). These were followed by "Is committed to achieving high standards" (Quality Orientation), "Thinks in strategic terms" (Strategic Orientation) and "Motivates others to reach team goals" (Leadership). The lowest loading items included "Uses correct spelling and grammar" (unstandardized loading 0.38) measuring Written Communication, "Works long hours" measuring Personal Motivation, and "Fits in with the team" measuring Interpersonal Sensitivity. Based on these findings, the Method

factor in self-assessments emphasized the most desirable managerial behaviors across different measured competencies, and deemphasized unimportant or less desirable behaviors (although some of these behaviors may be important for employees at non-managerial levels). The common method effect had all the features of the "ideal-employee" factor described by Schmit and Ryan (1993), and in the present study of managers it could be labelled the "ideal-manager" factor. Thus, hypothesis H2a was supported.

H2b: Ratings by external observers

For bosses, items with the largest loadings on the Method factor were "Is effective in leading others" and "Builds effective teams" measuring Leadership (unstandardized loadings 1.74 and 1.63 respectively), followed by several items measuring Quality Orientation (e.g. "Sets high standards"). Exactly the same items had the largest loadings for subordinates (unstandardized loading for "Is effective in leading others" was 1.77). For peers, items measuring Quality Orientation (e.g. "Produces high quality results", 1.64) had the largest loadings on the Method factor, followed by some items measuring Leadership, Action Orientation, and Creativity and Innovation. The items with lowest loadings were similar across the external rater perspectives. Item "Pays attention to the political process" had the lowest loading for bosses and peers (0.36 and 0.39 respectively), and item "Takes a radical approach" had the lowest loading for subordinates (0.45). For all external perspectives, items "Works long hours", "Is profit conscious" and "Identifies opportunities to reduce costs", and "Uses correct spelling and grammar in writing" were among the least salient.

These findings suggest that the Method factor captured a similar construct for the ratings by external observers and selves, with most desirable managerial behaviors affected most and to the same extent across the rater perspectives. Table 2 provides the correlations between the

Method factor loadings for each pair of rater perspectives. It can be seen that the loadings were indeed very similar, with closest correspondence between bosses and peers (r = .90), and least correspondence between selves and subordinates (r =.74).

INSERT TABLE 2 ABOUT HERE

These results suggest that the Method factor had a very similar meaning for self-ratings and other ratings. For the external rater perspectives, the effect does not appear fully consistent with the definition of "illusory halo" as the cognitive bias of exaggerated coherence, because the latter should lead to uniform overreporting or underreporting of all characteristics depending on whether the target is appraised positively or negatively in general (Murphy et al., 1993). Instead, we observed greater distortion of characteristics that the assessors know to be more important for the rated manager's appraisal outcome (we will hypothesize possible mechanisms for this in the Discussion). This process may have been supplemented by the bias of exaggerated coherence ("illusory halo"), which we hypothesized. However, it is impossible to separate the two processes with the present design, and due to the overriding non-uniform effect, we have to reject H2b.

Hypothesis 3: Improved Construct Validity with the Bias Control and Prevention Methods

H3a. Factorial structure of competency domains

H3a proposed that both bias-controlled SS ratings and FC rankings would yield more meaningful factorial structure (i.e. differentiated behavioral domains in line with theoretical expectations) than straight SS ratings. To test this hypothesis, we compared the covariance structures of the 16 latent competency traits emerging from each measurement model.

The **SS model** was characterized by a strong positive manifold of all competency correlations for every rater perspective. Table 3 reports the average off-diagonal latent (error free) correlations, which ranged from r = .51 for self-ratings to an astonishing r = .65 for subordinates. Not surprisingly, the latent traits were highly suitable for factor analysis – the Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy were "good" or "very good". Parallel Analysis revealed that just one factor was sufficient to describe the variability in competency constructs, explaining over 50% of variance for all rater perspectives (see Table 3).

In the ***SS-Method model***, the average off-diagonal latent correlations were positive but much lower than in the SS model (see Table 3). While correlations between conceptually concordant competencies remained as strong as in the SS model (for example, Commercial Awareness and Strategic Orientation still correlated around .60–.70 for all rater perspectives), correlations between conceptually unrelated competencies disappeared (for example, Specialist Knowledge and Resilience correlated around .40–.50 in the SS model, but became zero in the SS-Method model). The data were no longer well suited for factor analysis – the KMO measures were barely over .5 and classified as "bad" (see Table 3). Despite this result, we explored the factorial structure underlying the competency constructs. Parallel Analysis suggested four factors for self-assessments, and three factors for bosses, peers and subordinates. The three-factor solutions for external ratings yielded very strong conceptual similarities. The three factors could be labelled: 1) "*executing*", indicated by Problem Solving and Analysis, Specialist Knowledge, Planning and Organizing, Written Communication and Quality Orientation; 2) "*getting ahead*", indicated by Commercial Awareness, Creativity and Innovation, Strategic Orientation, Action Orientation, Oral Communication, Persuasiveness and Personal Motivation; and 3) "*getting along*", indicated by Interpersonal Sensitivity, Leadership, Flexibility, and Resilience. Self-

ratings yielded one further factor – "*communicating*", separating Oral Communication and Persuasiveness into a distinct domain.

INSERT TABLE 3 ABOUT HERE

The **FC model** converged for the self- and peer perspectives; however, some additional constraints on factor loadings were necessary to avoid Heywood cases (i.e. negative variance estimates) in models for bosses and subordinates. Table 1 reports goodness of fit indices for the FC models. It can be seen that despite the significant χ^2, the χ^2/df ratios were better than for the SS and SS-method models, and approximate fit according to the RMSEA was good. The SRMR also indicated acceptable fit for all rater perspectives except bosses, for whom this index was only slightly above the cut-off.

The FC models yielded patterns of the latent competency correlations similar to the SS-Method models. Table 3 shows that the average off-diagonal correlations were positive but low[vi]; however, strong positive relationships between conceptually related competencies were preserved. The FC competencies were slightly more suitable for factor analysis than the SS-Method competencies – the KMO was "bad" for selves, but "mediocre" for all external rater perspectives. Parallel Analysis suggested four factors for self-assessments and three factors for external perspectives. The three factors for the external perspectives were "getting ahead", "executing" and "getting along", with the same main competencies indicating them as in the SS-Method model, but some differences in salience of various competencies.

To summarize the factor analysis results, the SS-Method and FC models yielded well-differentiated behavioral domains in line with theoretical expectations. Specifically, conceptually related competencies correlated strongly while conceptually unrelated competencies did not

correlate. This is in contrast to straight SS ratings, which yielded competency scores correlating strongly regardless of conceptual similarity, with just one common factor underlying them. The theoretically justified patterns of competency correlations, as well as the meaningful factorial structures emerging from the method-controlled SS ratings and the FC rankings confirm the hypothesis.

H3b. Rater agreement

H3b proposed that both bias-controlled SS ratings and FC rankings would yield better rater agreement compared to straight SS ratings. Intraclass correlations of estimated competency scores in three-level models are reported in Table 4. We remind the reader that only external rater perspectives were included in these analyses (self-assessments were not included). In the **SS model**, agreement between two random raters of the same target was moderate (average ρ_{target} = .25). The common perspective improved the within-target agreement a little, yielding the average $\rho_{perspective}$ = .32. However, only for Strategic Orientation and Action Orientation the increment $\rho_{perspective}-\rho_{target}$ was substantial (.12 and .13 respectively).

In the **SS-Method model**, agreement between two random raters of the same target was better than in the SS model (competency average was ρ_{target} = .31). The common perspective again improved the within-target agreement, yielding the average $\rho_{perspective}$ = .39. Four competencies achieved increments over .1; these were Strategic Orientation ($\rho_{perspective}-\rho_{target}$ = .23), Planning and Organizing (.13), Action Orientation (.15) and Persuasiveness (.13).

INSERT TABLE 4 ABOUT HERE

For the **FC model**, agreement due to common target was substantially better than for the SS model, and slightly better than for the SS-Method model (average ρ_{target} = .33). The common

perspective improved the within-target agreement, reaching the average $\rho_{perspective}$ = .41. Five competencies showed increment over .1; these were Strategic Orientation ($\rho_{perspective}-\rho_{target}$ = .22), Action Orientation (.19), Planning and Organizing (.16), Persuasiveness (.11) and Specialist Knowledge (.10).

To summarize, the SS model yielded the lowest agreement values; the SS-Method model yielded substantially better agreement values, and the FC model performed best. For 12 out of 16 competencies, the FC model yielded small increments in agreement over the SS-Method model (most of which were statistically significant given the very large sample size for these analyses). To avoid multiple significance testing but provide a test of overall model performance, we computed inter-rater agreement based on the sum of the 16 competency scores, representing the overall appraisal score for each alternative model. While the differences between the SS and SS-Method models were small ($\rho_{perspective}$ = .29 and .31 respectively), the FC model yielded a substantially better agreement ($\rho_{perspective}$ = .40) over targets' overall performance appraisal. This confirms the hypothesis, and also provides evidence for the relative performance of the bias control (SS-Method model) and prevention (FC model) methods.

H3c. Convergent and discriminant validity

H3c proposed that both bias-controlled SS ratings and FC rankings would yield higher correlations with similar external constructs while not increasing correlations with dissimilar constructs. Table 5 presents the correlations between the IMC competencies estimated using the alternative measurement models, and conceptually concordant personality traits measured by the self-reported OPQ32. As expected, self-assessments of competencies yielded substantial correlations with self-reported personality. The **SS model** fared worst (average convergent correlation r = .35), while the **SS-Method model** and **FC model** did better (average r = .42 and r

= .40 respectively), although most differences between correlations were not statistically significant with the current sample size (N = 202).

Assessments by external observers correlated weaker with self-reported personality; nevertheless, most hypothesized correlations were statistically significant and positive. Here, the advantages of either modeling biases or preventing them were even more obvious than in the self-assessments, improving the average validity of the **SS model** (average r = .14) more substantially (and yielding statistically significant improvements for several competencies), reaching the average validity r = .25 for the **SS-Method model**, and r = .26 for the **FC model**. However, just like for the self-assessments, differences between the convergent correlations in the SS-Method and FC models were inconsistent across competencies and insignificant.

While improving the convergent correlations, the bias control and prevention methods did not inflate the correlations between conceptually unrelated constructs of IMC and OPQ – the heterotrait-heteromethod correlations for self-assessed and other-assessed competencies were low, and much lower than the convergent correlations (see Table 5). The convergent and discriminant validity evidence of the improvements achieved by the use of bias control and prevention methods confirms the hypothesis. We also have evidence for the approximately equal performance of both methods in relation to convergent and discriminant validities.

INSERT TABLE 5 ABOUT HERE

Discussion

The objective of this study was to examine the extent to which 360-degree appraisals of competencies are subject to biases, examine the nature of these biases, and test whether validity gains can be achieved by either statistically controlling for biases or preventing them with

forced-choice formats. We compared operational appraisals data collected using Likert scales (SS format) with multidimensional forced-choice (FC format) data, by applying model-based measurement. We systematically compared three methods of inferring measurement: 1) assuming that only substantive constructs (competencies) are captured by the SS format; 2) modeling a method factor to control for biasing effects in the SS format; and 3) preventing the occurrence of biases by employing the FC format. To our knowledge, the present study is the first to apply Thurstonian IRT modeling (Brown & Maydeu-Olivares, 2011) to multisource feedback collected using multidimensional forced choice, to overcome the problems of ipsative data and ensure proper scaling of competencies.

The results suggested that SS responses were subject to strong common method biases at the item level, making behavior ratings across all competencies highly similar. When ignoring these effects in scoring, one factor was sufficient to explain the variability in theoretically distinct and diverse 16 competencies. Clearly, administering a long instrument (the IMC comprises 160 items) to measure just one construct defeats the purpose of a differentiated 360-degree assessment; it is a waste of time for everyone involved in the process.

The important question is what caused this similarity of assessments on all competencies – substantive overlap (i.e. real clustering of competencies in the same individuals or “true halo”), rater biases (including seeming clustering of competencies in the same individuals or “illusory halo”, and desired clustering of competencies or “ideal-employee” factor), or both substantive overlap and rater biases? From the analyses in which statistical modeling of a common method factor was attempted, we found that both causes were at play, with rater biases having a greater influence. Indeed, while the competency factors and the Method factor explained approximately equal amounts of variance in the average item (around 30% each), only about half of the

competency-related variance could be attributed to broader domains of competence – the substantive overlap (see the results of factor analysis in Table 3, column "SS-Method"). By explicitly modeling rater biases, competency perceptions and errors of measurement in SS ratings, this study contributes to the debate of differentiating between illusory and true halo (Murphy et al., 1993).

The next important question is the nature of the detected rater biases. The evidence suggests that the Method factor represented a meaningful construct, which was remarkably similar across the rater perspectives. We found that the distortion was non-uniform, with the most salient indicators of the Method factor representing the most desirable leadership behaviors spanning many competency constructs, including transformational and transactional leadership qualities (Judge & Piccolo, 2004). The salient indicators incorporated positively charged words such as "reach team goals", "high standards", "motivate others", "committed", "effective", etc. The weakest indicators of the Method factor represented behaviors not typically associated with effective leadership, such as "writes in a fluent manner" or "works long hours". For self-assessments, this was not surprising since the "ideal-employee" factor, previously found and replicated in the literature (Klehe et al., 2012; Schmit & Ryan, 1993) has these features of emphasizing the important job characteristics. Interestingly, however, external appraisal ratings had the same features. Instead of, or in addition to expected cognitive bias of exaggerated coherence (which presumably influences all behaviors uniformly), we observed particular overreporting of behaviors that made the target look like a more effective manager ("ideal-manager" factor). Interestingly, the extent of overreporting was similar across raters of the same target – this is evident from a non-ignorable nesting effects of the estimated Method factor scores (ρ_{target} = .24 and $\rho_{\text{perspective}}$ = .32). It appears that the external observers overreported behaviors of

particular managers, and that they were selective about which behaviors to overreport. Two mechanisms might have been at play here. First, motivated distortions similar to impression management but on behalf of another person might have taken place. This is in line with earlier work by Murphy and Cleveland (1995) and also Murphy et al. (2004), who showed that raters do manipulate appraisal ratings in pursuit of their own goals, for example overreport others' behaviors when seeking greater harmony with colleagues (also Randall & Sharples, 2012), or distort selected behaviors of executives pursuing own political goals. Second[vii], distortion may have resulted from raters applying their own implicit theories of what makes an effective leader (e.g., Eden & Leviatan, 1975; Rush, Thomas & Lord, 1977); with the behaviors most central to perceived leadership effectiveness affected most. Given that observing and evaluating leadership behavior is a complex cognitive process and followers are unlikely to observe and accurately recall all behaviors, it has been argued that the reliance on implicit leadership theories reduces the amount of information processing involved (e.g. Rush et al., 1977). We believe that the Method factor identified in the present study contributes to future research on non-uniform manipulation of behavioral ratings by external assessors, be they driven by rater goals or implicit leadership theories or both. We cannot support or disprove either explanation based on the data we have; experimental manipulations or external covariates controlling for rater goals or the level of familiarity with the target of assessment could delineate the possible causes. Future research should investigate the specific motivations underlying such distortions, particularly with respect to different rater perspectives.

When biasing effects were explicitly modelled, the "purified" competency constructs captured intended behaviors more closely. The strong positive correlations between conceptually concordant competencies remained intact, while correlations between unrelated competencies

disappeared, supporting the expected relatedness between some behaviors but also distinctiveness of domains underlying managerial performance. As a result, meaningful second-order competency domains emerged. Three easy-to-interpret dimensions were identified in the present study, which we labelled "executing", "getting ahead" and "getting along", after the distinct vectors described by Hogan and Shelton (1998). The bias-controlled competency scores also yielded substantially better interrater agreement and convergent validities than the straight SS scores. To summarize, the researcher can make a better use of the collected SS ratings by explicitly modeling the common method factor causing all items to overlap.

Like the statistical bias control method, the prevention method using the FC format was effective in improving validities of the competency constructs. In terms of structural validity of FC rankings, the competency correlations were in line with theoretical expectations, and the second-order competency domains were meaningful. This is not surprising since sufficient evidence already exists for good structural validity of properly scaled FC rankings (e.g. Brown & Maydeu-Olivares, 2013; Brown & Maydeu-Olivares, 2011). Furthermore, the similarity of the SS-Method and FC factor structures supports the validity of the SS-Method model, and reinforces its effectiveness in separating the substantive and the biasing effects. In terms of rater agreement, the FC rankings yielded substantial improvements compared to straight SS ratings, and small improvements compared to bias-controlled SS ratings. When the choice between behaviors was forced, impressive levels of agreement were achieved for some competencies (e.g. $\rho_{\text{perspective}} = .58$ for Commercial Awareness, and $\rho_{\text{perspective}} = .54$ for Strategic Orientation). Most importantly, further gains were made in perspective-related agreement, where previous research struggled to find any non-ignorable effects (LeBreton et al., 2003; Yammarino, 2003). Five competencies showed perspective-related increment in agreement exceeding .1, and two of them

reached the increment around .2. Such substantial differences support the practice of separating raters by perspective, while negligible differences provide evidence against this practice. Importantly, any discussions about the behaviors for which raters have similar perceptions, or for which rater perspective matters, should be based on scores that are as free from rating distortions as possible.

In terms of relationships with external personality measures, the FC rankings performed slightly better on average than the straight SS ratings and on par with the bias-controlled SS ratings. These findings corroborate earlier meta-analytic results (Salgado & Táuriz, 2014) on similar or slightly higher convergent correlations of forced-choice questionnaires compared to single-stimulus questionnaires. The unique contribution of the present study is the demonstration of convergent and discriminat validities of properly scaled (i.e. not ipsative) forced-choice rankings in the condition of high biases. To our knowledge, this is the first study to show that reduction of biases with the forced-choice format can actually translate into better correlations with external measures in the context of high response distortions. Previous research in a low-bias context found the FC format performing slightly worse than the SS format (Brown & Maydeu-Olivares, 2013).

Overall, the FC format demonstrated substantial gains over the straight SS ratings in all aspects of construct validity, and further small gains over the bias-controlled SS ratings with respect to inter-rater agreement. This is a significant achievement considering that the rater agreement analysis was based on estimated scores, and FC scores are typically less reliable that SS scores derived from the same items (Brown & Maydeu-Olivares, 2013). Furthermore, the forced-choice format is not immune to non-uniform distortions within the same block. What might be the mechanism for the small incremental gains of FC rankings over bias-controlled SS

ratings that we observed? Kahneman (2011) argued that explicit comparisons (here, comparisons between behaviors from different behavioral domains) engage the cognitive "System 2", which comprises slow and rational thinking processes. This contrasts the fast and intuitive "System 1", which is engaged by default, resulting in many cognitive biases. Kahneman attributed finer discrimination achieved by using comparative judgments in his research to engagement of System 2 instead of System 1. The present research may have evidence for such a differentiating and contrasting process having taken place. Despite the strong similarities between the SS-Method and FC factor structures underlying the competency constructs, the FC model comprised a larger number of small negative cross loadings. For example, boss appraisal of a target's Interpersonal Sensitivity that primarily loaded on "getting along" in the SS format, kept its central role to that factor in the FC format but also acquired a weak negative loading on "getting ahead". Thus, ranking interpersonal sensitivity above other competencies indicated not only the boss's high appraisal of the target's ability to "get along", but also slightly lesser appraisal of the target ability to "get ahead", thus demonstrating how explicit contrasts with other behaviors may have enhanced cognitions.

Conclusions and Recommendations

Today the evidence is overwhelming that response processes involved in 360-degree appraisals are complex, and are affected by response distortions, likely both unmotivated and motivated. From the analyses of a large sample of responses to a comprehensive 360-degree assessment tool in this study, we obtained the evidence that these distortions are so strong as to substantially deteriorate the score validity. We therefore argue that whenever ratings are collected using Likert scales, model-based control of biases at the item level is necessary. In the present study, we modeled the common Method factor, retaining substantive overlaps between

measured traits (i.e. “true halo”) while removing biases affecting all items (for example, due to motivated distortions aimed to present a picture of an “ideal manager” by self or others, unmotivated cognitive bias of exaggerated coherence or implicit leadership theories held by external observers). Following guidance on model assessment by Williams, Edwards and Vandenberg (2003), we obtained the evidence that the Method factor 1) accounted for a very sizeable proportion of variance in responses; 2) yielded significant improvements in terms of model fit; 3) yielded significant improvements in terms of convergent and discriminant validity.

Another and possibly even better alternative is bias prevention by collecting forced-choice rankings of targets’ behaviors. The impressive validity gains obtained by using rankings when compared to straight ratings, and small further gains in inter-rater agreement when compared to ratings statistically controlled for the Method factor, suggest that forcing differentiation between facets of assessment is a viable and effective method, which is well placed to capture the essence of 360-degree feedback – perceptual judgements of competencies. In practice, the FC method can be implemented operationally so that the model parameters are established once (based on the multiple-group CFA), and then applied automatically to all new assessments yielding estimated scores on traits of interest. This is in contrast to impracticalities of modelling the Method factor in SS ratings, which cannot be done once and for all, since different assessment contexts are likely to change the model parameters severely. To make the best use of the forced-choice method, careful matching on desirability of behaviors within each block is strongly recommended to minimize non-uniform response distortions; other guidelines for creating good forced-choice designs must also be followed (e.g. Maydeu-Olivares & Brown, 2010; Brown & Maydeu-Olivares, 2011; Brown, 2016). Importantly, a suitable IRT-based model must be used to score the forced-choice responses to avoid ipsative data, such as Thurstonian

IRT models for dominance items (Brown & Maydeu-Olivares, 2011), or Multi-Unidimensional Pairwise Preference models for ideal-point items (Stark, Chernyshenko, & Drasgow, 2005) . Practical implications are far-reaching as we can make better use of information collected in 360-degree assessments but also improve performance appraisals and leadership assessment in organizations more widely, which suffer from the same problems (Adler et al., in press; Landy & Farr, 1980).

Endnotes

[i] This is in contrast to orthogonal bifactor models, which assume one general and several specific factors mutually uncorrelated with each other. The general factor in such models is the only source of shared variance between items measuring different traits, and the specific factors capture the residual variance within specific domains not explained by the general factor. This approach would be suitable for modeling "true halo" – the substantive common cause of the overlap between competencies. In fact, any data where a common second-order factor explains the covariation of measured traits can be presented in the bifactor form (Rindskopf & Rose; 1988). Model presented in Figure 2 (panel 2) is different in that it allows the competency traits to correlate; therefore, the general "method" factor (e.g. "illusory halo") is not the only source of shared variance between the items measuring different traits; it is in addition to any covariance structure underlying the competencies (e.g. "true halo").

[ii] Sample Mplus syntax for analysis of partial rankings in forced-choice blocks of size 4 is available for download from the online version of Brown and Maydeu-Olivares (2012); the macro writing Mplus syntax for analysis of forced-choice questionnaires of different configurations is available for download from http://annabrown.name.

[iii] Multilevel modelling of the latent traits is possible; however, it is not computationally feasible in this study due to the very large number of items and traits in the measurement part of the model.

[iv] To estimate the forced-choice scores, the average model parameters across 10 imputations were applied to the original dataset with one of every six responses missing by design.

[v] Unfortunately, current computing capabilities are prohibitive of obtaining goodness of fit for models with a very large number of categorical outcomes, such as the models in our study. To overcome this problem and obtain a reasonable indication of how the alternative models fared in comparison, we obtained fit indices for models including only the first half of observed responses (80 items rather than 160 items). Conveniently, the IMC questionnaire employs a balanced design, with the first half containing 20 blocks of 4 items, and exactly 5 items measuring each competency. This allowed testing the measurement models with the full hypothesized latent structure, but the reduced number of observed indicators.

[vi] This is in contrast to ipsative scores, which always yield negative average off-diagonal correlation. For 16 scales, the off-diagonal correlations would necessarily average at –.07 (Brown & Maydeu-Olivares, 2013).

[vii] We would like to thank one of the anonymous reviewers for pointing out the possible link with implicit leadership theories.

References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. Industrial and Organizational Psychology, 9(2), 219-252. doi: 10.1017/iop.2015.106

Asparouhov, T. (2006). General multi-level modeling with sampling weights. Communications in Statistics: Theory and Methods, 35, 439-460. doi: 10.1080/03610920500476598

Asparouhov, T. and Muthén, B. (2010). Multiple Imputation with Mplus (Version 2). Retrieved from http://www.statmodel.com

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. Organizational Research Methods, 6(1), 15-43. doi: 10.1177/1094428102239424

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. International Journal of Selection and Assessment, 15 (3), 263-272. doi: 10.1111/j.1468-2389.2007.00386.x

Bartram D., Brown A., Fleck S., Inceoglu I., & Ward K. (2006). OPQ32 Technical Manual. Surrey, UK. SHL Group.

Bartram, D., Robertson, I., & Callinan, M. (2002). A framework for examining organizational effectiveness. In Robertson, I. T., Callinan, M., & Bartram, D. (Eds.), Organizational effectiveness: The role of psychology (pp. 227-255). John Wiley & Sons.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. Psychological Methods, 17(4), 665-678. doi: 10.1037/a0028111

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. Psychometrika, 79(3), 515-537. doi: 10.1007/s11336-013-9390-9

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. Psychological Methods, 19(4), 528-541. doi: 10.1037/met0000016

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. Educational and Psychological Measurement, 71(5), 814-833. doi: 10.1177/0013164410388411

Borman, W. C. (1997). 360 degree ratings: An analysis of assumptions and a research agenda for evaluating their validity. Human Resource Management Review, 7, 299–316. doi: 10.1016/S1053-4822(97)90010-3

Bozeman, D. P. (1997). Interrater agreement in multi-source performance appraisal: A commentary. Journal of Organizational Behavior, 18, 313–316. Stable URL: http://www.jstor.org/stable/3100178

Brown, A. (2016). Item Response Models for Forced-Choice Questionnaires: A Common Framework. Psychometrika, *81(1),* 135–160. doi: 10.1007/s11336-014-9434-9

Brown, A. & Bartram, D. (2009-2011). OPQ32r Technical Manual. Surrey, UK. SHL Group.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement, 71, 460–502. doi: 10.1177/0013164410375112

Brown, A. & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. Behavior Research Methods, 44, 1135–1147. doi:10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods, 18, 36-52. doi: 10.1037/a0030641

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. Structural Equation Modeling, 9, 55-77. doi: 10.1207/S15328007SEM0901_4

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(3), 259-278. doi: 10.1080/13594320802678414

Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. Journal of Applied Psychology, 60(6), 736. doi: 10.1037/0021-9010.60.6.736

Hansbrough, T.K., Lord, R. & Schyns, B. (2015). Reconsidering the Accuracy of Follower Leadership Ratings. The Leadership Quarterly, 26(2), 220-237. doi: 10.1016/j.leaqua.2014.11.006

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. Journal of Applied Psychology, 91, 9-24. doi: 10.1037/0021-9010.91.1.9

Hogan, R., & Shelton, D. (1998). A Socioanalytic Perspective on Job Performance. Human Performance, 11(2), 129–144. doi:10.1207/s15327043hup1102&3_2

Hox, J. J. (2010). Multilevel analysis: Techniques and applications (Second Edition). Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55. doi: 10.1080/10705519909540118

Judge, T. A. & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. Journal of Applied Psychology, 89, 755-768. doi: 10.1037/0021-9010.89.5.755

Kahneman, D. (2011). Thinking, fast and slow. London, UK: Allen Lane.

Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. Human Performance, 25(4), 273-302. doi: 10.1080/08959285.2012.703733

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107. doi: 10.1037/0033-2909.87.1.72

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology, 65,* 501-506. doi: 10.1037/0021-9010.65.5.501

LeBreton, J. M., Burgess, J. R., 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(1), 80-128. doi: 10.1177/1094428102239427

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. Applied Psychological Measurement, 37(6), 497-498. doi: 10.1177/0146621613487794

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. Psychological Methods, 11(4), 344-362. doi: 10.1037/1082-989x.11.4.344

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1, 130-149. doi: 10.1037/1082-989X.1.2.130

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological methods, 17(3), 437. doi: 10.1037/a0028085

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74, 619 – 624. doi: 10.1037/0021-9010.74.4.619

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

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. Journal of Applied Psychology, 89(1), 158-164. doi: 10.1037/0021-9010.89.1.158

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. Journal of Applied Psychology, 78, 218-225. doi: 10.1037/0021-9010.78.2.218

Muthén, L. K., & Muthén, B. O. (1998 – 2015). *Mplus user's guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879. doi: 10.1037/0021-9010.88.5.879

Randall, R., & Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. Journal of Occupational and Organizational Psychology, 85(1), 42-59. doi: 10.1348/2044-8325.002002

Rindskopf, D. & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. Multivariate Behavioral Research, 23, 51–67. doi: 10.1207/s15327906mbr2301_3

Rush, M. C., Thomas, J. C., & Lord, R. G. (1977). Implicit leadership theory: A potential threat to the internal validity of leader behavior questionnaires. Organizational Behavior and Human Performance, 20(1), 93-110. doi: 10.1016/0030-5073(77)90046-0

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. European Journal of Work and Organizational Psychology, 23(1), 3-30. doi: 10.1080/1359432X.2012.716198

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. Journal of Applied Psychology, 78(6), 966. doi: 10.1037/0021-9010.78.6.966

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. Journal of Applied Psychology, 85(6), 956. doi: 10.1037/0021-9010.85.6.956

SHL. (1997). Inventory of Management *Competencies: Manual and User's Guide.* Surrey, UK. Saville & Holdsworth ltd.

Stark, S., Chernyshenko, O., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. Applied Psychological Measurement, 29, 184–203. doi: 10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2011). Constructing fake-resistant personality tests using item response theory. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), New perspectives on faking in personality assessment (pp. 214-239). London: Oxford University Press.

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

Thurstone, L.L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286. doi: 10.1037/h0070288

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. Psychological Methods, 16, 209-220. doi: 10.1037/a0023353

Van der Heijden, B., & Nijhof, A. (2004). The value of subjectivity: Problems and prospects for 360-degree appraisal systems. The International Journal of Human Resource Management, 15(3), 493-511. doi: 10.1080/0958519042000181223

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. Journal of Cross-Cultural Psychology, 35(3), 346-360. doi: 10.1177/0022022104264126

Warr, P., & Bourne, A. (1999). Factors influencing two types of congruence in multirater judgments. Human Performance, 12(3-4), 183-210. doi: 10.1080/08959289909539869

Warr, P., & Bourne, A. (2000). Associations between rating content and self-other agreement in multi-source feedback. European Journal of Work and Organizational Psychology, 9(3), 321-334. doi: 10.1080/135943200417948

Webster, H. (1958). Correcting personality scales for response sets or suppression effects. Psychological Bulletin, 55(1), 62-64. doi: 10.1037/h0048031

Williams, L. J., Edwards, J. R., & Vandenberg, R. J. (2003). Recent advances in causal modeling methods for organizational and management research. Journal of Management, 29(6), 903-936. doi: 10.1016/S0149-2063_03_00084-9

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(3), 592. doi: 10.1037/0021-9010.90.3.592

Yammarino, F.J. (2003). Modern Data Analytic Techniques for Multisource Feedback. Organizational Research Methods, 6(1), 6-14. doi: 10.1177/1094428102239423

Appendix. IMC Competencies

1. **Commercial Awareness**- Identifying opportunities for new business and for cost savings. Taking account of revenue and cash flow. Showing awareness of competitor activity.
2. **Specialist Knowledge**- Demonstrating specialist knowledge in job. Keeping up to date with advances in own area of expertise. Quickly assimilating new technical information.
3. **Problem Solving & Analysis**- Making rational judgments. Drawing appropriate conclusions from information provided. Integrating data from different sources. Effective problem solving.
4. **Creativity & Innovation**- Generating creative ideas. Coming up with imaginative solutions and fresh insights to work-related issues.
5. **Strategic Orientation**- Understanding organizational strategy and corporate aims. Relating own work or that of teams to long-term organizational goals.
6. **Planning & Organizing**- Effective planning and organizing. Setting up and monitoring timescales and plans, allocating realistic time scales for activities, keeping track of activities.
7. **Action Orientation**- Making decisions without delay. Making decisions under pressure. Taking initiative to act.
8. **Oral Communication**- Speaking clearly and confidently. Presenting in a compelling manner to groups. Expressing ideas clearly. Articulating key points of an argument concisely. Responding to feedback from an audience.
9. **Written Communication**- Writing clearly and succinctly. Using correct spelling and grammar. Producing written communication that is easy to follow.
10. **Interpersonal Sensitivity**- Supporting others in their work. Interacting with others in a sensitive way. Listening and showing concern for others.

11. **Persuasiveness**- Persuading others to own viewpoint. Convincing with counter-arguments. Lobbying effectively. Negotiating well. Changing opinions of others.
12. **Leadership**- Coordinating group activities. Building effective teams. Motivating and empowering individuals or teams to reach organizational goals. Identifying development opportunities for staff.
13. **Quality Orientation**- Setting high standards. Paying attention to quality issues. Producing high quality results. Encouraging a sense of high standards in others.
14. **Flexibility**- Adapting own behavior to new circumstances. Reacting positively to change. Supporting change initiatives.
15. **Resilience**- Staying calm under pressure. Keeping control in stressful situations. Working effectively under pressure.
16. **Personal Motivation**- Showing drive and determination. Taking on new work. Seeking career progression. Seeking responsibility. Working long hours.

Table 1

Goodness of fit for alternative measurement models in perspective-specific analyses

Model	SS	SS-Method		FC**
Observed variables	80 (5 categories)	80 (5 categories)		120 (2 categories)
degrees of freedom	2,960	2,880	Δdf [80]	6,800***
χ^2			$\Delta\chi^2$	
Self	6,627	5,675	[787]*	9,914
Boss	7,497	6,311	[1,160]*	10,051
Peers	7,615	6,278	[1,225]*	9,694
Subordinates	11,184	8,889	[2,047]*	11,463
RMSEA (90% CI)				
Self	.038 (.036–.039)	.033 (.032–.035)		.023 (.022–.024)
Boss	.044 (.043–.045)	.039 (.037–.040)		.025 (.024–.025)
Peers	.037 (.036–.038)	.032 (.031–.033)		.019 (.018–.020)
Subordinates	.039 (.038–.039)	.034 (.033–.034)		.019 (.018–.020)
SRMR				
Self	.061	.052		.080
Boss	.064	.052		.086
Peers	.059	.047		.077
Subordinates	.052	.043		.064

Note. The models were limited to the first half of the questionnaire to enable calculation of χ^2 and RMSEA (see Endnote v); SRMR were calculated on the full questionnaire. 90% CI = 90 percent confidence interval. * $\Delta\chi^2$ is not equal to the difference of the χ^2 because the difference of χ^2 are not distributed as chi-square when estimators with robust errors are used (such as the ULSMV used in this study), and adjustments need to be made (the Mplus function DIFFTEST accomplishes that). ** Values for the first imputed dataset are reported. *** The degrees of freedom in the FC models were adjusted for redundancies among the thresholds and tetrachoric correlations estimated from the binary outcome variables (Maydeu-Olivares, 1999). There are 4 redundancies per block of 4 items, thus the degrees of freedom printed by Mplus were reduced by 4*20=80.

Table 2

Correlations between the Method factor loadings in perspective-specific analyses

	Self	Boss	Peers
Boss	.78		
Peers	.84	.90	
Subordinates	.74	.83	.83

Note. Correlations are computed across k = 160 item factor loadings.

Table 3

Exploratory factor analyses of the 16 latent competency constructs by measurement model in perspective-specific analyses

Model	SS	SS-Method	FC
Average off-diagonal correlation			
Self	.51	.07	.04
Boss	.54	.12	.06
Peers	.57	.07	.05
Subordinates	.65	.06	.04
Kaiser-Meyer-Olkin (KMO) test			
Self	.88	.53	.57
Boss	.87	.58	.60
Peers	.89	.54	.63
Subordinates	.91	.57	.60
Number of factors (% of variance explained)			
Self	1 (59%)	4 (50%)	4 (58%)
Boss	1 (58%)	3 (50%)	3 (52%)
Peers	1 (64%)	3 (48%)	3 (54%)
Subordinates	1 (70%)	3 (47%)	3 (49%)

Table 4

Agreement among external observers (boss, peers and subordinates) within corresponding sampling units, by multiple-group measurement model

Measurement model	SS		SS-Method		FC	
Sampling unit	ρ_{target}	$\rho_{perspective}$	ρ_{target}	$\rho_{perspective}$	ρ_{target}	$\rho_{perspective}$
Commercial Awareness	.29	.38	.45	.51	.51	.58
Specialist Knowledge	.32	.35	.36	.43	.36	.46
Problem Solving and Analysis	.22	.27	.33	.36	.37	.42
Creativity and Innovation	.20	.28	.29	.33	.30	.37
Strategic Orientation	.19	.31	.26	.49	.32	.54
Planning and Organizing	.28	.36	.31	.44	.31	.47
Action Orientation	.22	.35	.29	.43	.27	.47
Oral Communication	.24	.31	.28	.33	.26	.31
Written Communication	.25	.30	.30	.34	.30	.33
Interpersonal Sensitivity	.30	.34	.42	.45	.40	.45
Persuasiveness	.20	.29	.23	.36	.22	.33
Leadership	.30	.37	.31	.37	.34	.41
Quality Orientation	.29	.33	.28	.33	.32	.37
Flexibility	.19	.26	.30	.36	.35	.40
Resilience	.20	.26	.22	.26	.26	.32
Personal Motivation	.30	.36	.40	.44	.36	.38
Competency average	**.25**	**.32**	**.31**	**.39**	**.33**	**.41**
Overall appraisal score	.21	.29	.22	.31	.31	.40

Note. Overall performance is computed as the sum of the 16 competency scores. Intraclass correlations ρ_{target} and $\rho_{perspective}$ are calculated using Equations (5) and (6). Increments $\rho_{perspective} - \rho_{target} > .1$ are underlined.

Table 5

Correlations between OPQ32 traits and IMC competencies scored by multiple-group measurement models

		Self (N=202)			Others, mean (N=208)		
IMC	OPQ32	SS	SS-Method	FC	SS	SS-Method	FC
Commercial Awareness	Competitive	.27**	.30**	.31**	.19**	.23**	.19**
Specialist Knowledge	Data Rational	.21**	.35**	.31**	-.02	.14*	.19**
Problem Solving and Analysis	Evaluative	.37**	.50**	.49**	.13	.25**	.28**
	Data Rational	.21**	.42**	.43**	-.03	.22**	.31**
Creativity and Innovation	Innovative	.67**	.69**	.66**	.22**	.43**	.36**
	not Conventional	.38**	.37**	.35**	.25**	.32**	.19**
Strategic Orientation	Forward Thinking	.35**	.40**	.34**	-.01	.13	.20**
Planning and Organizing	Conscientious	.39**	.40**	.35**	.22**	.32**	.32**
	Detail Conscious	.34**	.49**	.49**	.20**	.45**	.49**
Action Orientation	Decisive	.39**	.39**	.48**	.28**	.29**	.35**
Oral Communication	Socially Confident	.42**	.46**	.38**	.07	.15*	.12
Interpersonal Sensitivity	Caring	.31**	.44**	.45**	.12	.24**	.24**
Persuasiveness	Persuasive	.42**	.46**	.34**	.11	.25**	.17*
Leadership	Controlling	.48**	.32**	.37**	.29**	.12	.24**
Quality Orientation	Detail Conscious	.34**	.46**	.38**	.23**	.39**	.41**
Flexibility	Adaptable	.04	.18*	.20**	.15*	.11	.08
Resilience	Tough Minded	.29**	.34**	.31**			
	Emotionally Controlled				.13	.10	.14*
Personal Motivation	Achieving	.44**	.57**	.54**	.23**	.41**	.41**
Average convergent validity		**.35**	**.42**	**.40**	**.14**	**.25**	**.26**
Average discriminant validity	raw values	.03	.01	-.01	.01	.00	-.01
	absolute values	.12	.12	.12	.07	.09	.09

Note. * Correlation is significant at 0.05 level (2-tailed); ** Correlation is significant at 0.01 level (2-tailed).

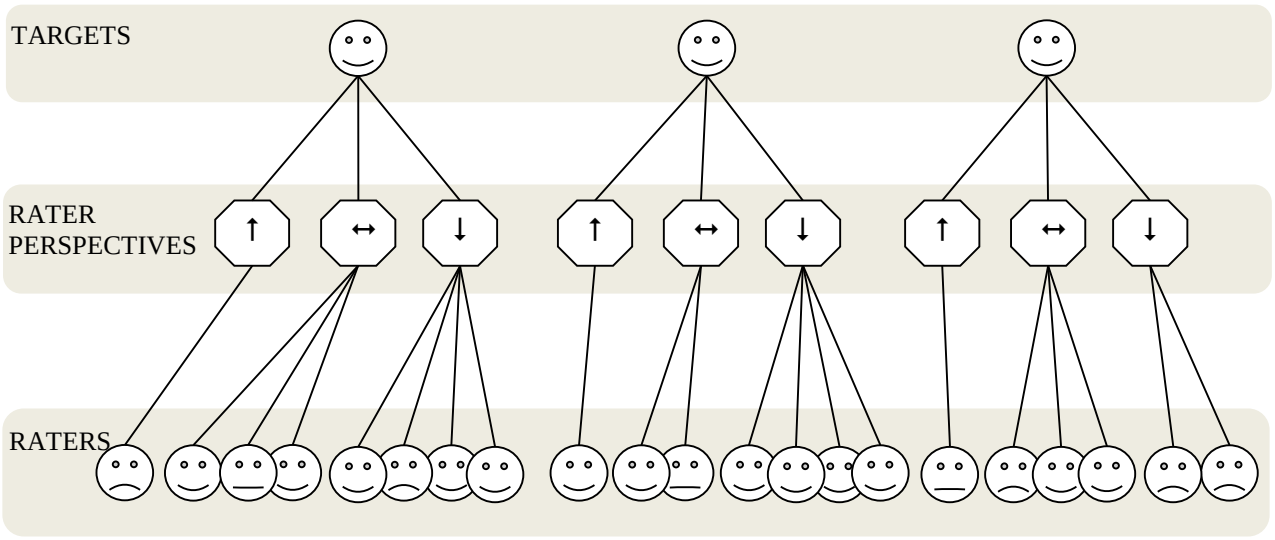
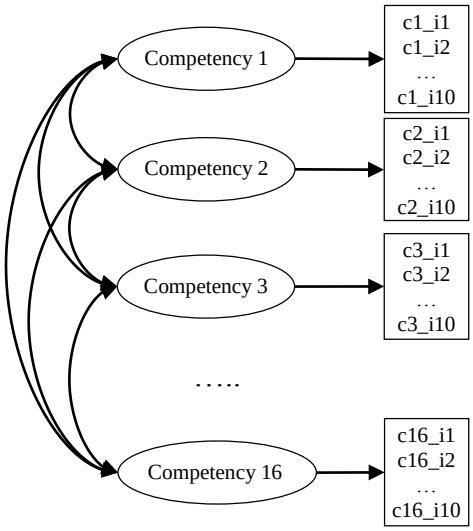
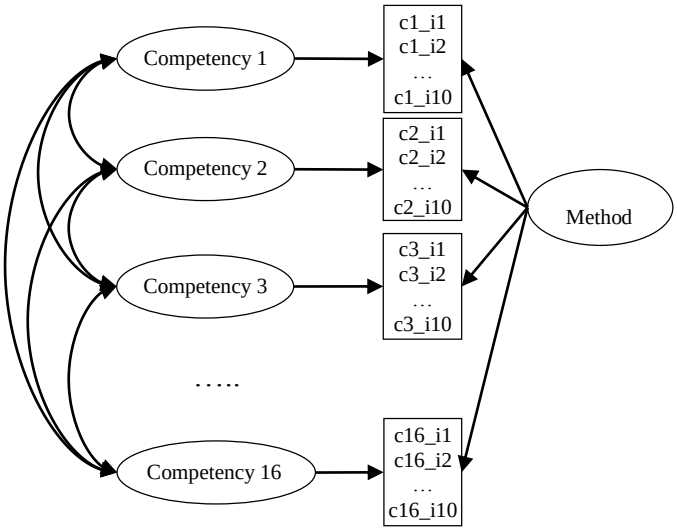


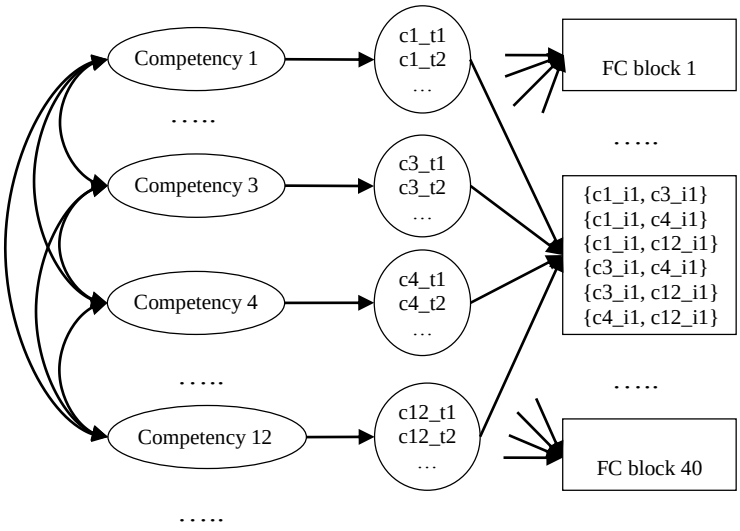
Figure 1. Hierarchical nesting of external raters. Raters are nested within perspectives (boss =↑, peers = ↔ and subordinates = ↓), and perspectives are nested within targets of assessment.



1. SS (Single-Stimulus)



2. SS-Method (Single-Stimulus with Method factor)



3. FC (Forced-Choice)

Figure 2. **A**lternative measurement models for IMC item responses