Agreeing to Disagree?
Explaining Self–Other Disagreement on Leadership Behaviour

Dominik Vogel

Alexander Kroll

Agreeing to Disagree?
Explaining Self–Other Disagreement on Leadership Behaviour

Leadership research tends to treat differences among ratings of the same leaders as measurement error. Our study makes such varying perceptions of leadership behaviour its main phenomenon of investigation. We conceptualize divergent leadership ratings based on the difference between managers' self-ratings and team members' assessments of leadership behaviour. Using data from three German public organizations on 51 teams and 190 leader-follower dyads, we find that divergent leadership ratings are a function of managers' motivation, their use of managerial reflection routines, and team members' personality. The findings point to the importance of using multisource feedback and developing managers' self- and other-awareness.

Introduction

Imagine a public manager who considers him or herself to be a great leader, but their team members do not agree. False self-perceptions of leadership can be just as detrimental to organizations as poor leadersh1ip behaviour itself. Inaccurate self-ratings are not only associated with lower performance, but may also have a negative effect on subordinates' job satisfaction and organizational commitment (Fleenor et al. 2010, 1019). Divergent leadership ratings among managers and employees are an indicator of lacking self-awareness, which is seen as an important antecedent of leadership effectiveness (Atkins and Wood 2002; Atwater and Yammarino 1992). Or, as Fletcher (1997, 186) states: if self-perceptions differ from those of others, "[…] then it is difficult to see how one can manage work relationships successfully, how one can contribute well as a team member, and how one can adapt one's behavior to the circumstances and individuals."

If managers are bad leaders, they receive training, mentoring, or are assigned less supervisory responsibilities. If managers are bad leaders, but they think they do a great job and hence overestimate their leadership skills, however, organizations may be even worse off. In such cases in which managers lack self-awareness, leadership problems may go undetected or behavioural change may require coercion as managers do not see the need to alter their behaviour (Atwater et al. 2005). Team members will become frustrated due to inadequate or a lack of leadership, and they may leave the team or agency, or simply respond by showing low levels of motivation, engagement, or performance (Bass and Yammarino 1991; Lee and Carpenter 2018).

Practitioners of human resource management and leadership development are aware of the self–other disagreement problem (Gentry, Cullen, and Altman 2016; Sala 2003). More than twenty-five years ago, employing the Johari Window for mapping what is known or unknown to a person's self and to others was a de rigueur aspect in leadership training. Since then, training and development specialists have promoted the use of multisource feedback to offer managers a more realistic view of their leadership behaviour (Kutcher, Donovan, and Lorenzet 2010). So-called 180 and 360 (degree) reviews have become popular components of public agencies' leadership development programs (Van Wart 2003, 220; U.S. Office of Personnel Management 2017; Hunt and Ivergard 2007). The former type of reviews refers to assessments that include feedback from subordinates and superiors, while the latter incorporates additional horizontal feedback from peers or other sources (Chappelow 2004).

Meanwhile, research is trying to catch up with these developments. Although leadership research has not been ignoring issues related to incongruent ratings, we see a tendency to discount managers' self-ratings as unreliable and a preference to replace them with other-ratings from subordinates. However, by treating divergences in ratings as measurement error that needs to be minimized, we miss an opportunity to study and better understand a phenomenon with real-life relevance: misconceptions of leadership behaviour and factors that may drive or mitigate the difference in self–other rating agreement (Fleenor et al. 2010).

Scholarship in the areas of business management, organizational behaviour, and psychology has begun to address this research gap (for an overview, see Fleenor et al. 2010 as well as Lee and Carpenter 2018). While public management research has made a great deal of progress with regard to the understanding of different leadership behaviours and their respective impact (Van Wart 2013; Kroll and Vogel 2014; Vogel and Masal 2015; Tummers and Knies 2016; Crosby and Bryson 2018), we have given little attention to the topic of self-other rater disagreement (Hassan and Rohrbaugh 2009 as well as Jacobsen and Andersen 2015 are notable exceptions).

With this article, we want to spotlight the issue of divergent self- and other-ratings of leadership and stimulate discussion among public management scholars. We think this is necessary because the rating gap may be, in fact, larger in the public than private sector. Rigid hierarchies, less flexible career paths, and separation due to educational degrees (Bach and Kessler 2007; van Dorp 2017) could lead to more distance among political appointees, managers at different echelons, and employees resulting in more rating disagreement. In addition, the perception gap may be more detrimental in the public sector as well. Due to the lack of financial incentives, less managerial resources, and more goal ambiguity (Hvidman and Andersen 2014; Pandey and Rainey 2007; Chen and Rainey 2014), it is often difficult to manage government organizations through "systems" or "by design," and instead, such organizations have to rely on functional human relations, effective leadership, and realistic perceptions of such leadership behaviour.

Incongruent leadership assessments can be caused by misperceptions on the part of the leaders as well as the team members. With regard to leaders, we argue that their leadership perceptions will be more in agreement with those of team members, the more the former are able and willing to reflect on their own behaviour. In particular, we find that leaders with a strong motivation to lead show more agreement with follower assessments, pointing towards the importance of people's interest in being or becoming leaders. We also find that the creation of managerial learning routines, involving leaders as well as team members, increases self–other agreement. On the part of the team members, we expect that certain personalities may be less prone to producing incongruent assessments than others. More specifically, we find that employees' agreeableness decreases rating incongruence, while openness increases it.

One of the implications of this research is that in addition to leaders' self-awareness, a concept widely studied and praised in the business literature (Fleenor et al. 2010), other-awareness matters too. Employees with different personalities, levels of job satisfaction, and of years of age seem to differ systematically in the evaluations of their leaders. Leaders, who are aware of this and take into account that different people have different leadership expectations, will also be more likely to meet these expectations, can customize leadership behaviour, and satisfy team members' needs. Our empirical analysis is based on data from three German public organizations from which the responses of 190 team members were matched with those of 51 team leaders.

A Model of Divergent Self- and Other-Ratings of Leadership

*Previous Research and an Integrative Approach*

A meta-analysis by Lee and Carpenter (2018) identifies 41 published articles on rater incongruence across industries and sectors. They find that self-assessment and other-ratings of leadership are only moderately correlated, and that leaders tend to over-estimate specific behaviours such as transformational leadership, servant leadership and ethical leadership. Studies that examine public sector datasets confirm these findings. School principals' self-assessments of instructional and transformational leadership vary greatly from the other-assessments provided by the teachers in the same schools (Ham, Duyar, and Gumus 2015; Park and Ham 2016; Wang, Wilhite, and Martino 2016; for contrasting findings, see Muterera et al. 2018).

Research in the realm of public management on this topic includes work by Hassan and Rohrbaugh (2009) who examine the 360-degree feedback ratings of 68 midlevel public managers. This study finds that self-ratings as well as ratings by supervisors, peers, and subordinates were incongruent; the gap between self- and other-ratings was bigger than that among several types of other-ratings; and that these incongruences varied in different performance domains. In a study on leadership behaviour, Jacobsen and Andersen (2015) show that only employee-based leadership ratings, as opposed to self-ratings, were positively related to performance improvements, pointing towards the difference between intended and perceived leadership practices.

Whereas the public management literature does not provide many insights on factors influencing the identified differences between self- and other ratings, work from other fields has documented important correlates. For example, managers in higher positions are more likely to overestimate their leadership behaviour, just like managers with lower educational attainment. Empathy mitigates self–other disagreement, whereas narcissism increases it. Managers in individualist cultures show less self–other agreement with regard to their leadership skills than managers in collectivist cultures (e.g., Atwater et al. 2009; Judge, LePine, and Rich 2006; Ostroff, Atwater, and Feinberg 2004; for a more detailed overview, see Fleenor et al. 2010).

While research on the topic is accumulating, we are still in search of holistic theories that help explain the gap between self- and other-ratings. We argue that incongruences in leadership ratings between managers and followers can be a function of factors occurring at different conceptual levels shown in figure 1. Hence, we cannot fully assess the importance of factors at one level if we do not, at the same time, account for factors at a different level, which is why we take an integrative approach to the study of rater disagreement. Our perspective is in line with previous calls for more integrative theory on leader-environment-follower interactions (Wofford 1982) that is inclusive of, among other things, social dynamics and cognitive elements (Avolio 2007). In that sense, leadership is constructed through the relationships connecting individuals rather than being an attribute of individuals (Balkundi and Kilduff 2006, 420).

Figure 1: Integrative Theoretical Framework
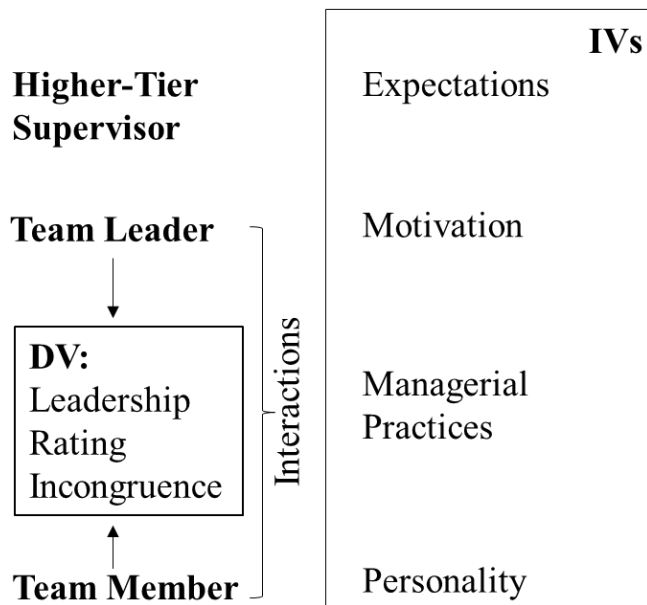
Note: DV = dependent variable; IVs = independent variables; while additional interactions within the model are possible (supervisor–team leader; supervisor–team member), our theoretical focus is on team leader–team member interactions because the dependent variable has been constructed at this level. Figure by Vogel & Kroll, 2019; available at https://doi.org/10.6084/m9.figshare.7619432 under a CC-BY 4.0 license.

Figure 1 shows our theoretical framework that integrates actors at three different levels. In line with previous work, we use both self-ratings by managers and other-ratings by followers and conceptualize disagreement as the lack of consensus between the ratings as our dependent variable (Atwater and Yammarino 1997; Fleenor et al. 2010). We call this "self–other disagreement," but also refer to it as "rater incongruence or divergence" or "perception gap." We think of leaders as people who are in charge of managing other people ("team leaders" and "team managers"), which is a much more inclusive definition than that of the organization's top management. We consider leadership as an essential part of managers' responsibilities (Mintzberg 2009) and do not differentiate between the terms leader and manager. We speak of followers, employees, subordinates, or team members when we refer to the people being led. The team leaders themselves have to report to superiors at a higher hierarchical level, whom we refer to as higher-tier supervisors.

Although it is rarely feasible to construct a model of self–other disagreement that incorporates multiple factors at each conceptual level, it was our goal to hypothesize and test one specific variable at each level that could be linked to rating incongruences. With regard to higher-tier supervisors, we think that their social expectations of team leaders may affect the latters' leadership behavior and related self-ratings (Merton 1957; Shivers-Blackwell 2004). An important factor for

team leaders themselves that may drive behavior and self-perceptions thereof is their own motivation (Chan and Drasgow 2001; Gagné and Deci 2005). To capture the role of social interactions between leaders and followers, we turn to managerial routines that are not directly related to leadership but offer opportunities for both sides to participate in managing the work unit (Levitt and March 1988; Schedler and Proeller 2010). At the follower level, we focus on team members' personality, which is a stable trait known to be influential in predicting workplace attitudes and perceptions (Costa and MacCrae 1992; Tett and Guterman 2000). In what follows, we explain all theoretical linkages in greater detail.

Supervisor- and Leader-Related Factors Driving Incongruence

A first factor that can explain differences in self–other disagreement are the signals leaders receive from their direct supervisors ("higher-tier supervisors"). Most managers also report to higher-tier superiors (department or division heads) themselves. We argue that studying the relationship between these managers (team leaders) and their supervisor helps to better understand leaders' behaviours and their self-perceptions thereof. In particular, we believe that supervisors' interest in managers' leadership generates normative pressure to care about leadership and to seek ways for improvement. Furthermore, interested supervisors may also provide important insights and feedback that will help managers to reflect on their own practices more objectively.

The normative pressure generated by a supervisors' interest can be explained by role theory (Merton 1957). By being a team leader, managers fulfil a social role, which is associated with certain expectations the managers themselves, as well as others, have about how a team leader should behave (Shivers-Blackwell 2004, 43). These role expectations are directly or indirectly communicated to the team leaders (Kahn et al. 1964) and have a significant impact on how leaders behave. In addition, higher-tier supervisors' expectations are especially influential because their role expectations can be transformed into role pressure. The more intense the role pressure perceived by a person, the more likely they will succumb to the resulting expectations. We assume that an interested supervisor is more likely to communicate their role expectations and create—implicitly or explicitly—pressure on managers to take on an active leadership role.

A second mechanism through which higher-tier supervisors affect managers' leadership perceptions is external feedback. Multisource feedback which includes feedback from "above" has been found to affect leaders' performance and yield more accurate self-assessments (Atkins and Wood 2002; Atwater, Roush, and Fischthal 1995; Day et al. 2014). If a supervisor is interested in how their subordinate managers lead, and if the supervisor addresses the issue of leadership with them, the subordinate managers gain a broader understanding of how their behaviour is perceived by others. Because these managers receive more feedback on their leadership behaviour, they are able to anchor their self-perceptions and better put their leadership practices into perspective, resulting in less perception incongruence.

> H₁: A higher-tier supervisor showing interest in managers' leadership will decrease the gap in leadership perceptions.

A second variable that may affect managers' self-assessments is their "motivation to lead," which is their motivation to assume leadership responsibility (Chan and Drasgow 2001). Such motivation consists of three dimensions that tap into different motives: affective-identity (taking on leadership responsibility due to self-inclination and preference), social-normative (taking on leadership responsibility due to compliance with social norms), and non-calculative motivation to lead (taking on leadership responsibility without calculating costs and benefits). Research on the impact of managers' motivation to lead has increased over the last 15 years. Such work showed that a high motivation to lead has a positive impact on leadership emergence, emotional intelligence, leadership behaviour, and leadership effectiveness (e.g., Hong, Catano, and Liao 2011; Van Iddekinge, Ferris, and Heffner 2009; Vogel 2016).

Our model accounts for the affective-identity dimension of managers' motivation to lead and its impact on their self-assessments. Leaders with a high affective-identity motivation find intrinsic enjoyment in leading other people (Chan and Drasgow 2001, 482), as managers do not aspire to become leaders because of external rewards or expectations. We assume that managers who deeply care about leadership will also put more effort into leadership-related tasks (Vogel 2016) and are more interested in getting feedback from their team members. Since they have an intrinsic motivation to lead, they might be more responsive to the feedback given by their followers and, therefore, have a more realistic understanding of their employees' demands and perceptions. Hence, hypothesis 2 states that leaders with a high affective-identity motivation to lead show more congruent self-assessments than others.

> H₂: A manager's affective-identity motivation to lead will decrease the gap in leadership perceptions.

In addition to feedback from "above" (higher-tier supervisors) and managers' motivation to lead, we will now take a closer look at the impact of broader organizational learning routines that are not necessarily linked to leadership behaviour but still offer opportunities for leaders and followers to interact and exchange as well as discuss ideas in a less hierarchical setting. The learning literature emphasizes how organizations can benefit from individual experiences through the creation of learning routines and, more specifically, forums in which individuals from different hierarchical levels and departments within the organization share, discuss, and make sense of more and less successful strategies of problem-solving (Levitt and March 1988; Piening 2013). Along these lines, public management research has documented that it is often not enough to adopt sophisticated management systems, but that there is a need to establish learning routines which allow people to regularly reflect on information from these tools to facilitate its purposeful use (Moynihan 2008).

Sophisticated management practices, which are associated with the "modern" paradigm of public management, include management by objectives, process optimization, quality management, and performance management (Schedler and

Proeller 2010). We call these practices "managerial reflection routines" as they are supposed to help managers systematically analyse and improve work processes and results as well as provide decision-makers with a better information base. We consider managers to be users of managerial reflection routines if they regularly optimize processes, discuss and define goals with their team, reflect upon work outcomes with their team, adapt team practices to client expectations, and host regular events where they proactively discuss how to improve the team's performance.

Side-effects of such learning routines are that they establish channels for two-way communication and increase opportunities to exercise and experience leadership. Rothstein (1990) argues that observers who frequently interact with leaders have more detailed insights into the leaders' work and therefore report ratings that are less divergent from leaders' self-assessments (Lee and Carpenter 2018). Hence, we argue that reflection routines, even when built around management practices rather than specific leadership behaviours, may also lead to more self-reflection that, in turn, should decrease the perception gap between leaders and employees.

> H_3: The use of managerial reflection routines will decrease the gap in leadership perceptions.

Follower-Related Personality Factors Driving Incongruence

While rater incongruences in leadership assessments can be a function of managers' misperceptions, they can also be attributed to follower-related factors (e.g., Hansbrough, Lord, and Schyns 2015; Felfe and Heinitz 2010). Since little is known about the impact of such factors (Fleenor et al. 2010), we focus on employees' personality, which is one of the most basic characteristics that may help explain differences in other-ratings. We employ the five-factor model of personality ("big five"), which proposes that most variation in personality can be accounted for by the following five robust factors: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Costa and MacCrae 1992).

Although personality has been found to significantly affect self-ratings (Judge, LePine, and Rich 2006; Bell and Arthur 2008), we know much less about its impact on other-ratings, suggesting that this relationship requires further exploration (Fleenor et al. 2010). Trait activation theory (Tett and Guterman 2000) can help to understand this relationship by explaining why certain personality characteristics influence subordinates' leadership ratings. The theory states that humans react in specific situations to cues they find in this situation, which then activate a certain trait (Lievens et al. 2006). We argue that such cues, related to the context of evaluating a supervisor, activate some personality traits more than others do. In particular, we think that agreeableness and conscientiousness affect other-ratings, while we do not see sufficient empirical evidence to propose hypotheses for the other three personality traits.

People who are highly agreeable are helpful, trusting, and empathetic (Costa and MacCrae 1992), and they are predisposed to view others positively (Hansbrough, Lord, and Schyns 2015, 222). In a situation in which agreeable persons have to assess others, this trait is activated, which is why their ratings of others tend to be less critical and more positive. In a study of students who rated each other's performance, Bernardin, Cooke, and Villanova (2000) find empirical evidence for such a leniency bias. A similar effect was found by Bartels and Doverspike (1997) when conducting an assessment centre and by Cheng, Hui, and Cascio (2017) regarding the rating of employees in a real estate agency in which agreeable raters gave elevated ratings. All these findings are underlined by a meta-analysis of 28 studies conducted by Harari, Rudolph, and Laginess (2015) that finds the same leniency effect of raters' agreeableness on their performance ratings. Since agreeable employees tend to provide less critical leadership ratings, we expect their assessments to be closer to leaders' self-ratings.

> H_4: A follower's agreeableness will decrease the gap in leadership perceptions.

The second personality trait that may be activated due to the process of assessing a supervisor's leadership behaviour is subordinates' conscientiousness. Conscientious people tend to think carefully before they act, and they put much attention to details (Costa and MacCrae 1992; Hansbrough, Lord, and Schyns 2015). They assess others more carefully and systematically and base their judgment on precise observations instead of categorization-based cognitive short-cuts (Hansbrough, Lord, and Schyns 2015). This accuracy hypothesis is also supported by Tziner, Murphy, and Cleveland (2005, 94) who argue that "highly conscientious raters may be less easily swayed by the rating context than their less conscientious peers." Bernardin, Cooke, and Villanova (2000) support this notion by showing that conscientious students rate their peers' performance more accurately.

Despite being thorough and accurate in their ratings, research has also shown that conscientious people may provide less straightforward assessments in certain situations (Cheng, Hui, and Cascio 2017). According to Bernardin, Tyler, and Villanova (2009) and Bernardin et al. (2016), conscientious people do not want to make their peers feel bad, which is why at times they may sacrifice accuracy for benevolence. However, in real-life settings, conscientious raters tend to be more accurate. Harari, Rudolph, and Laginess (2015) find evidence for such a pattern in their meta-analysis in which they compare rater effects across field and laboratory settings. Since our study is set up in a field environment—subordinates rate their real-life supervisors regarding experienced leadership behaviour—we expect conscientious team members to provide thoughtful and accurate assessments of their superiors, which are unlikely to be swayed by personal biases and thus likely to differ from leaders' self-assessments.

> H_5: A follower's conscientiousness will increase the gap in leadership perceptions.

Research Design

We use data from two different sources (leader and employee surveys) and operationalize rater incongruence, our dependent variable, as the difference between leaders' self-assessment and follower assessment of leadership behaviour. In addition to the variables we theorized about in the section above as potential predictors of self–other disagreement, we also control for confounding factors which were identified in prior research (see Fleenor et al. 2010). These include gender, age, job satisfaction, leadership training, and experience as well as the full set of the "big five" personality traits.

To account for the fact that drivers of perception incongruence are measured at different levels (leader-related versus follower-related variables) and that employees are nested in teams that work under the same leader (increasing the possibility of correlated errors), we use hierarchical linear modelling. The following sections explain the data, measurements, and limitations in greater detail.

Data

The data for our analysis were collected from three German public organizations between May and October 2014. We aimed for selecting organizations, which directly provide specific public services to citizens or businesses and at the same time vary in tasks, size, and jurisdiction (state versus local level).[1] The idea behind creating variation in these selection variables was to make sure that potential findings hold across contexts and ensure their external validity. Our data collection strategy resembles the cluster sampling approach: only a few organizations were targeted, but a full census of all employees was conducted.[2]

In the selected organizations we addressed all "street-level managers" (i.e., line managers who do not lead other managers) and all their followers. The authors sent online-based questionnaires including all items to measure managers' leadership behaviour as well as the independent variables to 112 managers and 1,364 followers. The response rates were 57.1% for managers and 34.0% for followers. Based on this response, 232 manager–follower dyads were created. A nonresponse bias analysis comparing early respondents with late respondents (Kypri, Stephenson, and Langley 2004; Groves 2006) did not reveal any indication of nonresponse bias. To be able to group responses from the same team, participants typed in a team-specific code at the beginning of the questionnaire (Vogel 2018). Participation in the survey was voluntary, it was assured that the raw data will not be shared with the organizations, and all participants will remain anonymous.

45 % of the managers and 66 % of the followers are female. 37.9 % of the followers are 46 to 55 years old, while 21.1 % are 26 to 35, and 25.3 % are 36 to 45 years old. Naturally, the managers are a bit older. 43.1 % are 46 to 55 years old, while 29.4 % are 36 to 45 and 21.6 % are older than 55 years. 60.8 % of the managers have ten

---

[1] Due to an anonymity agreement, we are not at liberty to disclose any additional information about the three organizations.
[2] In one of the organizations, it was not possible to include all teams. In this case, the sample consists of one complete department plus one sub-unit of each of the five other departments.

or fewer years of leadership experience, and a majority of 78.4 % had participated in leadership developmental programs.

Measuring Incongruence in Leadership Ratings

Leadership behaviour was assessed by using items from the Managerial Practices Survey (MPS, version G 17-4, see appendix 1) (Yukl 2012). The MPS was designed to capture the 15 leadership behaviours set out in the taxonomy of effective leadership behaviour (Yukl 2012). The behaviours are grouped into four dimensions: task-oriented, relations-oriented, change-oriented and external leadership behaviours. We used one item from the MPS for each leadership behaviour (which makes 3–4 items per dimension) and asked the leaders to assess their own behaviour based on these items. In addition, followers were asked to assess their leader's behaviour using the same items.[3]

We employed a second-order confirmatory factor analysis (CFA) to create one factor score based on the four leadership dimensions and 15 leadership behaviours. In line with Yukl's (2012) conceptualization, we labelled this factor score "effective leadership behaviour" (see appendix 2). The model fit indices indicate a good fit ($Chi2$(86) = 179.21, p < .001, *CFI* = .96, *TLI* = .96, *RMSEA* = .047, p(*RMSEA* <= .047) = .73, *SRMR* = .026), all paths are significant at p < .01, and the standardized factor loadings range between .71 and .96. The full result of the CFA is displayed in appendix 2.

Based on the results of the CFA, we operationalized our dependent variable: incongruence in leadership ratings. This incongruence is indicated by the absolute difference between a leader's self-perception of their leadership behaviour and the perception of each of their followers. In contrast to other studies on self–other agreement in leadership (e.g., Atwater et al. 2009; Braddy et al. 2014), we kept the perception of every follower as a single observation and decided not to aggregate followers' perceptions at the team level. If we aggregated followers' perceptions to a group mean, it would be impossible to fully take advantage of the nested structure of the data set and to consider follower characteristics as factors explaining divergent ratings. While our decision to keep followers' perceptions separate and calculate a single incongruence value for every dyad in the data set was mainly a conceptual one, we also employed several statistical measures to examine the appropriateness of data aggregation. The findings are mixed and justify our approach: $r_{WG(j)}$ = .86 (based on 95 % confidence intervals this number should be .92 or higher to be significant), *ICC(1)* = .44, and *ICC(2)* = .75.

---

[3] We focus on the taxonomy of effective leadership behaviour and the Managerial Practices Survey instead of other common operationalizations of leadership—like transformational leadership—because we wanted to capture a wide range of leadership behaviours. As the study of self–other agreement of leadership ratings is only at its beginning in the public sector, we belief it is important to gain a broad overview before focusing on more detailed aspects of leadership. This is in line with the development in empirical leadership research more broadly, where the distinction between task-oriented and relations-oriented leadership behaviours has only been emphasized relatively recently (Judge, Piccolo, and Ilies 2004; Humphrey 2002; Lee and Carpenter 2018).

Measuring Independent and Control Variables

The independent variables have been operationalized at two levels: the leader-related variables by surveying the leaders and the follower-related variables by asking the followers. All items—if not otherwise indicated—have been measured using a five-point Likert scale (1 = do not agree at all; 5 = totally agree). The wording of all the items can be found in Appendix 1. The measures for our controls (leadership experience, leadership training, gender, and age) are straightforward. We capture supervisor interest with two items and managers' affective motivation to lead as well as job satisfaction with one item (Chan and Drasgow 2001; Van de Ven and Ferry 1980; Hackman and Oldham 1974).

Our measures of reflection routines pick up on "modern" managerial practices, including process analysis, goal setting, and performance management (outcome and customer orientation as well as performance improvement efforts). However, all items do not only measure the existence of these practices but also whether they have been *regularly* used to *involve* managers and team members to *discuss* and *reflect* on the insights generated by them. Since these items are somewhat prone to social desirability bias, we asked the team members to assess them, instead of the team leaders. All reflection routines show high internal consistency (α = .93) and, according to principal factor analysis, load on a single factor.

We measure followers' personality employing the German 10-item short version of the big five inventory (Rammstedt and John 2007). This questionnaire uses two items for each of the five personality factors. In addition, we followed the advice of Rammstedt and John (2007) and included an additional item for agreeableness. A rotated solution of a principal factor analysis shows that the 11 items load on the five factors they were theoretically associated with (see appendix 1).

Limitations

Some limitations of our study are worth noting before we get to the results and their discussion. Although we have been able to reduce common method bias by using data from different sources (managers and followers) and constructing the dependent variable based on data from both sources, the study is prone to the common weaknesses of cross-sectional survey research, especially the limited ability to make causal claims. Another limitation is the drop in observations because we can only include dyads for which we can match followers and managers, resulting in a loss of information for all individuals without an appropriate match.

Our dependent variable does not discriminate between managers' over and under-estimation of their leadership behaviour. This is because the opposite of an incongruent assessment is a congruent assessment, no matter whether the incongruence has a positive or a negative sign. More normatively, we believe that any misperception is something organizations and managers themselves should avoid. To test whether the statistical results are sensitive to our coding decisions, we reran our models and this time only included those dyads where supervisors overestimated their own leadership behaviour. However, the results we present in this article remain the same and are unaffected by these different choices. We also want to emphasize that it is not the purpose of the study to determine whose assessment (leader versus follower) is more accurate. Rather, our point is that

divergences (no matter what their origin is) are problematic and will likely result in organizational dysfunctions.

Results

Table 1 provides an overview of the descriptives of the used variables and the correlations among them. We can see that the leadership rating incongruence ranges between 0.03 and 3.12, with the mean being 0.87 and the standard deviation 0.70. One-quarter of the observations show incongruences of 0.31 or less and one-quarter of more than 1.20 (see appendix 3). 148 of 190 observations (78 %) are over-estimations, which means the leaders themselves reported higher scores than their followers. The theoretical maximum of the incongruence is 3.58, and it is noteworthy that the empirically observed maximum of 3.12 is relatively close to the theoretical maximum, suggesting serious differences between ratings of leaders and followers.

To assess what drives incongruent leadership ratings, we use hierarchical linear modelling. Each leader–follower dyad was treated as a single observation in the analysis. Biased estimators were avoided by considering the hierarchical structure of the data. This means that the model accounts for the fact that some leaders are considered multiple times if more than one follower responded. Therefore, the analysis considers two different observational levels. Level 1 consists of the followers' responses, whereas level 2 is based on the responses by the leaders. In addition, we use agency-level fixed effects to control for organizational differences.

Table 1: Descriptives and Correlations

Leader-Level	Mean	S.D.	Range	(1)	(2)	(3)	(4)	(5)	(6)
(1) Motivation to lead	3.22	0.83	1–5	1.00					
(2) Supervisor's interest	3.26	1.01	1–5	0.09	1.00				
(3) Leadership experience	2.63	1.67	1–7	0.07	-0.05	1.00			
(4) Leadership training	0.78	0.42	0–1	0.08	-0.05	0.20	1.00		
(5) Job satisfaction	4.04	0.72	2–5	0.02	0.23	0.13	0.03	1.00	
(6) Gender (female)	0.45	0.50	0–1	-0.29*	0.02	-0.18	-0.10	0.17	1.00
(7) Age	3.80	0.85	2–5	-0.05	-0.15	0.45*	0.16	0.11	-0.16

Follower-Level	Mean	S.D.	Range	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(8) Leadership Rating Incongruence	0.87	0.70	0.03–3.12	1.00								
(9) Managerial reflection routines	0.09	0.93	-1.83–1.84	-0.42*	1.00							
(10) Extraversion	-0.01	0.73	-1.91–1.76	0.14*	-0.12	1.00						
(11) Openness	-0.07	0.66	-1.67–1.69	0.14*	0.02	0.54*	1.00					
(12) Conscientiousness	-0.12	0.67	-2.36–1.29	0.02	0.16*	0.26*	0.27*	1.00				
(13) Agreeableness	-0.05	0.64	-1.72–1.73	-0.12	0.23*	-0.15*	0.46*	0.15*	1.00			
(14) Neuroticism	0.03	0.63	-1.86–1.62	-0.03	-0.19*	-0.52*	-0.35*	-0.38*	-0.21*	1.00		
(15) Job satisfaction	3.85	0.90	1–5	-0.14	0.34*	0.17*	0.14	0.32*	0.08	-0.21*	1.00	
(16) Gender (female)	0.66	0.47	0–1	-0.07	-0.05	0.04	-0.02	0.24*	0.03	-0.01	0.05	1.00
(17) Age	3.42	1.02	1–5	0.16*	-0.16*	0.04	-0.01	-0.04	-0.10	-0.01	-0.07	-0.03

* $p < .05$

In the analysis, a random intercept model with level 1 and 2 predictors is used (Garson 2013). That is, we assume that the effects for different leaders have different intercepts for each of the leaders. The result of the hierarchical linear model is displayed in table 2. When comparing the full model (2) to the empty baseline model (1), several indicators suggest a good fit: The significance of the deviance test for the likelihood ratios indicates that the full model explains significantly more variance in the dependent variable than the baseline model. In fact, the full model reduces the unexplained variance by 41 % (R^2, S&B). As expected, the Akaike Information Criterion (*AIC*) is lower for the full model compared to the baseline, while this is not the case for the Bayesian Information Criterion (*BIC*).

Table 2 first displays the effects of the variables measured at the leader level (level 2), while the follower level (level 1) variables are shown in the second part of the table. The first hypothesis does not find support: higher-tier supervisors' interest in the leadership behaviour of their subordinated managers (hypothesis 1) does not reduce the leadership perception gap. In contrast, managers who enjoy leading others and find pleasure in taking on leadership responsibilities show significantly lesser differences between their self-perception and the perception of their followers (*b* = -0.16, *SE* = 0.087, *p* = .067). Hypothesis 3—which states that leaders' use of managerial reflection routines will reduce rater incongruence—can be confirmed (*b* = -0.29, *SE* = 0.072, *p* < .001).

With regard to team members' personality, we find evidence for the hypothesized effect of agreeableness (hypothesis 4). Followers who are more agreeable show less of a perception gap (*b* = -0.18, *SE* = 0.083, *p* = .027). In contrast with our expectations formulated in hypothesis 5, conscientiousness has no significant effect on rater incongruence. Interestingly, however, we see that followers who are more open to new experiences have reported significantly higher levels of self–other disagreement in leadership ratings (*b* = 0.22, *SE* = 0.092, *p* = .019). As expected, extraversion and neuroticism are not associated with rating incongruences.

Among the control variables, job satisfaction, gender, and age show significant coefficients. The gender effect seems to be the most interesting one, as it suggests that female leaders are less likely to have incongruent leadership ratings.

Table 2: Multilevel Regressions Analysis of Leadership Rating Incongruence

	(1) Baseline		(2) Final Model	
	b	SE	b	SE
Leader level				
Supervisor's interest (H1)			-0.085	(0.066)
Motivation to lead (H2)			-0.160*	(0.087)
Leadership experience			0.037	(0.042)
Leadership training			-0.179	(0.135)
Job satisfaction			0.169**	(0.077)
Gender (female)			-0.392**	(0.186)
Age			0.051	(0.059)
Follower level				
Managerial reflection routines (H3)			-0.286***	(0.072)
Agreeableness (H4)			-0.090	(0.087)
Conscientiousness (H5)			0.215**	(0.092)
Extraversion			0.049	(0.050)
Openness to experience			-0.184**	(0.083)
Neuroticism			-0.054	(0.076)
Job satisfaction			-0.067*	(0.036)
Gender (female)			-0.094	(0.079)
Age			0.107***	(0.040)
Constant	0.886***	(0.081)	0.845**	(0.403)
Var. level 2 residuals	0.227***	(0.099)	0.097***	(0.043)
Var. level 1 residuals	0.260***	(0.045)	0.188***	(0.021)
n (followers)	190		190	
n (leaders)	51		51	
AIC	355.53		313.04	
BIC	365.27		381.23	
log likelihood[a]	-174.77		-135.52***	

Note: Agency-level fixed effects included but not reported; robust standard errors (Huber-White sandwich estimator); estimator: maximum likelihood; [a] indicates the significance of the deviance test; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Some of our findings confirm intuition, others are unexpected, but all of them tend to be novel. That is, the literature on the factors we theorize about is sparse, which is why our hypothesis tests make a theoretical contribution, particularly since they are based on a robust multilevel analysis. Managers' motivation to lead is significant despite the fact that we control for two skill-related variables (training and experience); reflection matters even if centred on management routines instead of leadership behaviour; and agreeableness and openness show effects, but conscientiousness does not. We also find it noteworthy that our study provides evidence for the significance of the leadership perception gap in public organizations since the standardized mean difference between self- and other-ratings in our public-sector sample is 0.7, while it is close to zero in Lee and Carpenter's (2018) sector-spanning meta-analysis. The relevance and meaning of the hypothesized and unexpected findings require more explanation, which we offer in the following section.

Discussion

In contrast with what was hypothesized, we do not find evidence that managers provide more congruent self-assessments if their supervisors show great interest in their leadership behaviour. Although prior research documents that feedback generally reduces bias in self-assessments (e.g., Atwater, Roush, and Fischthal 1995), it is possible that feedback "from below" (employees) is more important than "from above" (higher-tier supervisors). Our empirical set-up may also constitute too demanding a test of the feedback-from-above hypothesis because we tested whether supervisor feedback reduces the gap in assessments between managers and employees (a 180-degree measure of rater incongruence), without being able to include actual assessments by higher-tier supervisors (which would require a 360-degree measure). It is also possible that supervisor interest affects how managers behave, but in ways that do not create more self-awareness. As explained in our theory section, such interest can be experienced as "role pressure" on the part of the managers, who may respond to high expectations from their supervisors by reporting inflated scores of their leadership behaviour, which would eventually increase the gap between self- and other-assessments (Shivers-Blackwell 2004).

The analysis, however, suggests that managers with a greater motivation to lead provide more accurate self-assessments of their leadership behaviour. This is particularly interesting, considering that scholars and practitioners alike have devoted more attention to "skill" (the ability to lead) as opposed to "will" (the desire to lead) when discussing leadership development efforts (e.g., Ingraham and Getha-Taylor 2004). Most developmental programs and selection processes focus on what abilities managers have, techniques they use to lead, or how they act in certain situations—succinctly their leadership "skills" (Seidle, Fernandez, and Perry 2016). Our findings, however, imply that there is another driver of managers' behaviour that should be assessed, developed, and cultivated: managers' urge to lead and take on responsibility—their "will" (see also Colquitt, LePine, and Noe 2000). While we suggest that the intrinsic motivation to lead needs to play a bigger role when selecting and promoting managers, this does not imply that organizations should focus their developmental efforts only on the most motivated managers (in fact, the most struggling managers may still need the most training) (Aguinis and Kraiger 2009).[4]

Our article also tests an extension of the generally accepted finding that feedback will reduce bias in leadership assessments (Bailey and Fletcher 2002; Day et al. 2014). While most research studied the impact of direct feedback about leadership behaviour, we examined the role of broader reflection routines, built around "modern" management practices that involve managers as well as employees. Our expectation was that, although these reflection routines are not

---

[4] Managers' motivation to lead should complement other selection criteria like merit or diversity, rather than replace them. Psychometric instruments to assess such a motivation could be used developmentally to increase self-awareness and help future leaders align interests, needs, and aspirations. They could also be employed to identify cases in which change is needed to make leadership roles more enjoyable, meaningful, or impactful.

primarily concerned with leadership behaviour but with organizational learning and improvement, such routines create additional opportunities for managers and employees to interact and exercise as well as experience leadership (Levitt and March 1988; Piening 2013) and hence reduce the leadership perception gap. We find support for this hypothesis and see that the use of such reflection routines makes leadership assessments of managers and employees more congruent. For scholars interested in leadership, our findings suggest it is not only the leadership-centred dialogue with employees that fosters self-awareness among managers, but that several types of leader–follower exchanges can have such an effect (see also Rothstein 1990). For management scholars, we can conclude that managerial practices related to instruments like process analysis or performance management can also have more or less intended impacts on leadership, and employees' perception thereof.

Although we did not hypothesize an effect of managers' gender on rater incongruence, it is worth discussing that we see such an effect in our analysis. We find that self–other disagreement is significantly lower if the assessed leader is female. Surprisingly, there is only limited research on gender differences in self–other assessments of leadership and therefore limited theorizing about such an effect (Fleenor et al. 2010). Ostroff, Atwater, and Feinberg (2004) suggest that women might be less likely to overestimate their own leadership behaviour, possibly because the difficulties they experienced when rising the ranks made them also more self-critical. Another explanation assumes that women are more likely to seek feedback and act on feedback by changing their behaviour and self-assessment (Roberts and Nolen-Hoeksema 1989), leading to more accurate self-assessments by female managers.

Our study is among the first to examine whether team members' personality affects the difference between leadership ratings by these members and self-ratings by their supervisors. We find that personality has some impact, although one hypothesis could not be confirmed. Employees, who are more agreeable, are more likely to provide lenient ratings similar to leaders' self-assessments. Employee openness leads to more incongruent ratings—a finding that deserves further explanation. One possibility is to account for the specifics of our sample. As reported above, for most leader–follower dyads an incongruent rating means that followers assigned leaders lower scores than the latter did in their self-assessment. With regard to openness, this could mean that employees who score high on this variable are also more critical of their leaders and expect them to be equally open-minded towards ideas, feelings, and new insights, resulting in lower leadership ratings and more incongruence. At the same time, we suggest caution when interpreting this finding. While we have statistical confidence in the reported coefficient, we will need additional research to sort out the exact causal mechanism behind the finding.

Conclusion

This article examined differences in leadership perceptions between managers and employees. We put forward an integrative model (e.g., Wofford 1982; Avolio 2007) that understands leadership (and the perceptions thereof) as being constructed through the social interactions among

actors at different levels of the organization. Hence, we argue that incongruent ratings are likely to be a function of factors located at different levels, including higher-tier supervisor's expectations; team leaders' motivation; managerial practices through which leaders and team members interact; and team members' personality. Using two surveys, administered in three German public organizations to collect data from 190 leader-follower dyads, we made the following observations.

First, managers' self-assessments of leadership become more aligned with assessments by their subordinates, the greater the managers' intrinsic motivation to lead is. This motivational effect was more relevant than their leadership experience, training, or feedback from their direct supervisors. Second, the establishment of managerial reflection routines reduces incongruence in leadership assessments. Although such routines were built around organizational learning and improvement efforts, we found "spillover effects," in that they offered additional opportunities for leaders and followers to interact, thereby reducing incongruence in leadership assessments. Third, we studied the extent to which followers' personality affects how they experience and rate their managers' leadership behaviour. While three out of five personality traits did not account for variation in rating differences, we found that agreeableness reduces rater disagreement whereas openness to experience reinforces it.

What can this article contribute to the study of leadership in public organizations? Understanding leadership behaviour is crucial, but often perceptions of behaviour are even more important than the behaviour itself. Variation in leadership perceptions can lead to differential organizational responses and consequences, including in- and out-group membership, the creation and deterioration of social capital, or absenteeism and job involvement. One main objective of this article was to direct attention to the study of self-other agreement in leadership research, a blind spot in the public management literature.

A second contribution is to provide more empirical insights into managers' self- and other-awareness. Incongruences in leader–follower ratings have been associated with little self-awareness, and we find indeed that self-awareness is driven by leadership-related "will" rather than "skill" factors, and that intrinsic drive matters more than external pressure. However, we also document that follower-related factors account for at least part of the rater incongruence. We, therefore, suggest that managers, in order to increase their odds to be successful, do not only have to be self-aware but also "other-aware." Put differently, if employees perceive the same leadership behaviour in different ways, then leaders may need to adjust behaviour from individual to individual, account for variation in leadership preferences, or—at a minimum—be aware that employees will respond differently to the same stimulus.

We see ample opportunities for further research. On the one hand, we acknowledge the need for private-public comparisons and replications of generic leadership studies on rater incongruence to better understand the extent to which we can expect convergent or divergent findings in public organizations. On the other hand, we suggest conducting research that involves variables unique to the public sector setting. For example, we could examine self- and other-ratings among political appointees and managers at different echelons, specifically, as we know that the accuracy of self-assessments varies across hierarchical positions (Gentry, Cullen, and Altman 2016; Sala 2003). Additionally, examining leadership ratings in the presence of high goal

ambiguity and the absence of strong financial rewards could provide important insights into self–other disagreement issues in the public sector. Further, research is needed that looks more closely at the negative consequences of incongruent leadership ratings and, hence, treats this phenomenon as the independent rather than the dependent variable.

We are in need of more research that simultaneously accounts for managers' as well as team members' personality and tests whether incongruent ratings are less likely to occur if managers and followers share similar personality profiles. We believe studying personality traits is important because these findings can provide important insights into the composition of teams and the adaptation of leadership styles when teams are dominated by certain personalities. Of course, our study should only be considered as one of the first steps in this direction, and we will need more research on the matter before being able to draw definitive conclusions.

References

Aguinis, Herman, and Kurt Kraiger. 2009. "Benefits of training and development for individuals and teams, organizations, and society." *Annual Review of Psychology* 60: 451–74.

Atkins, Paul W. B., and Robert E. Wood. 2002. "Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs." *Personnel Psychology* 55 (4): 871–904. doi:10.1111/j.1744-6570.2002.tb00133.x.

Atwater, Leanne E., Paul Roush, and Allison Fischthal. 1995. "The influence of upward feedback on self- and follower ratings of leadership." *Personnel Psychology* 48 (1): 35–59. doi:10.1111/j.1744-6570.1995.tb01745.x.

Atwater, Leanne E., David Waldman, Cheri Ostroff, Chet Robie, and Karin M. Johnson. 2005. "Self-Other Agreement: Comparing its Relationship with Performance in the U.S. and Europe." *Int J Selection & Assessment* 13 (1): 25–40. doi:10.1111/j.0965-075X.2005.00297.x.

Atwater, Leanne E., Mo Wang, James W. Smither, and John W. Fleenor. 2009. "Are Cultural Characteristics Associated with the Relationship Between Self and Others' Ratings of Leadership?" *Journal of Applied Psychology* 94 (4): 876–86. doi:10.1037/a0014561.

Atwater, Leanne E., and Francis J. Yammarino. 1992. "Does Self-Other Agreement on Leadership Perceptions Moderate the Validity of Leadership and Performance Predictions?" *Personnel Psychology* 45 (1): 141–64. doi:10.1111/j.1744-6570.1992.tb00848.x.

Atwater, Leanne E., and Francis J. Yammarino. 1997. "Self-other rating agreement: A review and model." In *Research in personnel and human resources management*. Vol. 15, edited by Leanne E. Atwater, Francis J. Yammarino, and Gerald R. Ferris, 121–74. Greenwich, Conn. JAI Press.

Avolio, Bruce J. 2007. "Promoting More Integrative Strategies for Leadership Theory-Building." *American Psychologist* 62 (1): 25–33. doi:10.1037/0003-066X.62.1.25.

Bach, Stephen, and Ian Kessler. 2007. "HRN and the New Public Management." In *The Oxford Handbook of Human Resource Management*, edited by Peter F. Boxall, John Purcell, and Patrick M. Wright, 469–88. Oxford: Oxford Univ. Press.

Bailey, Caroline, and Clive Fletcher. 2002. "The impact of multiple source feedback on management development: Findings from a longitudinal study." *J. Organiz. Behav.* 23 (7): 853–67. doi:10.1002/job.167.

Balkundi, Prasad, and Martin Kilduff. 2006. "The ties that lead: A social network approach to leadership." *The Leadership Quarterly* 17 (4): 419–39. doi:10.1016/j.leaqua.2006.01.001.

Bartels, Lynn K., and Dennis Doverspike. 1997. "Assessing the Assessor: The Relationship of Assessor Personality to Leniency in Assessment Center Ratings." *Journal of Social Behavior and Personality* 12 (5): 179.

Bass, Bernard M., and Francis J. Yammarino. 1991. "Congruence of Self and Others' Leadership Ratings of Naval Officers for Understanding Successful Performance." *Applied Psychology* 40 (4): 437–54. doi:10.1111/j.1464-0597.1991.tb01002.x.

Bell, Suzanne T., and Winfred Arthur. 2008. "Feedback acceptance in developmental assessment centers: The role of feedback message, participant personality, and affective response to the feedback session." *J. Organiz. Behav.* 29 (5): 681–703. doi:10.1002/job.525.

Bernardin, H. J., Donna K. Cooke, and Peter Villanova. 2000. "Conscientiousness and Agreeableness as Predictors of Rating Leniency." *Journal of Applied Psychology* 85 (2): 232–36. doi:10.1037/0021-9010.85.2.232.

Bernardin, H. J., Stephanie Thomason, M. Ronald Buckley, and Jeffrey S. Kane. 2016. "Rater Rating-Level Bias and Accuracy in Performance Appraisals: The Impact OF Rater Personality, Performance Management Competence, and Rater Accountability." *Hum. Resour. Manage.* 55 (2): 321–40. doi:10.1002/hrm.21678.

Bernardin, H. J., Catherine L. Tyler, and Peter Villanova. 2009. "Rating Level and Accuracy as a Function of Rater Personality." *Int J Selection & Assessment* 17 (3): 300–310. doi:10.1111/j.1468-2389.2009.00472.x.

Braddy, Phillip W., Janaki Gooty, John W. Fleenor, and Francis J. Yammarino. 2014. "Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self–other agreement." *The Leadership Quarterly* 25 (2): 373–90. doi:10.1016/j.leaqua.2013.10.001.

Chan, Kim-Yin, and Fritz Drasgow. 2001. "Toward a Theory of Individual Differences and Leadership: Understanding the Motivation to Lead." *Journal of Applied Psychology* 86 (3): 481–98. doi:10.1037//0021-9010.86.3.481.

Chappelow, Craig T. 2004. "360-Degree Feedback." In *The Center for Creative Leadership Handbook of Leadership Development*, edited by C. D. McCauley and E. van Velsor, 29–65. San Francisco, CA: Jossey-Bass.

Chen, Chung-An, and Hal G. Rainey. 2014. "Personnel Formalization and the Enhancement of Teamwork: a Public–private Comparison." *Public Management Review* 16 (7): 945–68. doi:10.1080/14719037.2013.770057.

Cheng, Kevin H. C., C. H. Hui, and Wayne F. Cascio. 2017. "Leniency Bias in Performance Ratings: the Big-Five Correlates." *Frontiers in psychology* 8: 521.

Colquitt, Jason A., Jeffrey A. LePine, and Raymond A. Noe. 2000. "Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research." *The Journal of applied psychology* 85 (5): 678–707. doi:10.1037//0021-9010.85.5.678.

Costa, P.T, and R.R MacCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual.* Odessa: Psychological Assessment Resources.

Crosby, Barbara C., and John M. Bryson. 2018. "Why leadership of public leadership research matters: And what to do about it." *Public Management Review* 20 (9): 1265–86. doi:10.1080/14719037.2017.1348731.

Day, David V., John W. Fleenor, Leanne E. Atwater, Rachel E. Sturm, and Rob A. McKee. 2014. "Advances in leader and leadership development: A review of 25years of research and theory." *The Leadership Quarterly* 25 (1): 63–82. doi:10.1016/j.leaqua.2013.11.004.

Felfe, Jörg, and Kathrin Heinitz. 2010. "The impact of consensus and agreement of leadership perceptions on commitment, Organizational Citizenship Behaviour, and customer satisfaction." *European Journal of Work and Organizational Psychology* 19 (3): 279–303. doi:10.1080/13594320802708070.

Fleenor, John W., James W. Smither, Leanne E. Atwater, Phillip W. Braddy, and Rachel E. Sturm. 2010. "Self–other Rating Agreement in Leadership: A Review." *Leadership Quarterly* 21 (6): 1005–34. doi:10.1016/j.leaqua.2010.10.006.

Fletcher, Clive. 1997. "Self Awareness - A Neglected Attribute in Selection and Assessment?" *Int J Selection & Assessment* 5 (3): 183–87. doi:10.1111/1468-2389.00058.

Gagné, Marylène, and Edward L. Deci. 2005. "Self-Determination Theory and Work Motivation." *J. Organiz. Behav* 26 (4): 331–62. doi:10.1002/job.322.

Garson, G. D. 2013. "Fundamentals of Hierarchical Linear and Multilevel Modeling." In *Hierarchical Linear Modeling: Guide and Applications*, edited by G. D. Garson, 3–25. Thousand Oaks, Calif. Sage.

Gentry, William A., Kristin L. Cullen, and David G. Altman. 2016. *The Irony of Integrity: A Study of the Character Strengths of Leaders.* Greensboro, NC: Center for Creative Leadership.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (5): 646–75. doi:10.1093/poq/nfl033.

Hackman, J. R., and Greg R. Oldham. 1974. *The Job Diagnostic Survey: an Instrument for the Diagnosis of Jobs and the Evaluation of Job Redesign Projects.* New Haven, Conn. Yale University, Department of Administrative Sciences. (Tech. Rep. No. 4).

Ham, Seung-Hwan, Ibrahim Duyar, and Sedat Gumus. 2015. "Agreement of self-other perceptions matters: Analyzing the effectiveness of principal leadership through multi-source assessment." *Australian Journal of Education* 59 (3): 225–46. doi:10.1177/0004944115603373.

Hansbrough, Tiffany K., Robert G. Lord, and Birgit Schyns. 2015. "Reconsidering the accuracy of follower leadership ratings." *The Leadership Quarterly* 26 (2): 220–37. doi:10.1016/j.leaqua.2014.11.006.

Harari, Michael B., Cort W. Rudolph, and Andrew J. Laginess. 2015. "Does rater personality matter? A meta-analysis of rater Big Five-performance rating relationships." *J Occup Organ Psychol* 88 (2): 387–414. doi:10.1111/joop.12086.

Hassan, Shahidul, and John Rohrbaugh. 2009. "Incongruity in 360-Degree Feedback Ratings and Competing Managerial Values: Evidence from a Public Agency Setting." *International Public Management Journal* 12 (4): 421–49. doi:10.1080/10967490903103375.

Hong, Ying, Victor M. Catano, and Hui Liao. 2011. "Leader emergence: The role of emotional intelligence and motivation to lead." *Leadership & Org Development J* 32 (4): 320–43. doi:10.1108/01437731111134625.

Humphrey, Ronald H. 2002. "The many faces of emotional leadership." *The Leadership Quarterly* 13 (5): 493–504. doi:10.1016/S1048-9843(02)00140-6.

Hunt, Brian, and Toni Ivergard. 2007. "Organizational climate and workplace efficiency." *Public Management Review* 9 (1): 27–47. doi:10.1080/14719030600853501.

Hvidman, Ulrik, and Simon C. Andersen. 2014. "Impact of Performance Management in Public and Private Organizations." *J Public Adm Res Theory* 24 (1): 35–58. doi:10.1093/jopart/mut019.

Ingraham, Patricia W., and Heather Getha-Taylor. 2004. "Leadership in the Public Sector: Models and Assumptions for Leadership Development in the Federal Government." *Review of Public Personnel Administration* 24 (2): 95–112. doi:10.1177/0734371X04263323.

Jacobsen, Christian B., and Lotte B. Andersen. 2015. "Is Leadership in the Eye of the Beholder? a Study of Intended and Perceived Leadership Practices and Organizational Performance." *Public Administration Review* 75 (6): 829–41. doi:10.1111/puar.12380.

Judge, Timothy A., Jeffery A. LePine, and Bruce L. Rich. 2006. "Loving Yourself Abundantly: Relationship of the Narcissistic Personality to Self- and Other Perceptions of Workplace Deviance, Leadership, and Task and Contextual Performance." *Journal of Applied Psychology* 91 (4): 762–76. doi:10.1037/0021-9010.91.4.762.

Judge, Timothy A., Ronald F. Piccolo, and Remus Ilies. 2004. "The Forgotten Ones? the Validity of Consideration and Initiating Structure in Leadership Research." *Journal of Applied Psychology* 89 (1): 36.

Kahn, Robert L., Donald M. Wolfe, Robert P. Quinn, J. D. Snoek, and Robert A. Rosenthal. 1964. *Organizational Stress: Studies in Role Conflict and Ambiguity.* New York: Wiley.

Kroll, Alexander, and Dominik Vogel. 2014. "The PSM-Leadership Fit: a Model of Performance Information Use." *Public Administration* 92 (4): 974–91. doi:10.1111/padm.12014.

Kutcher, Eugene, John Donovan, and Steven J. Lorenzet. 2010. "360-Degree Feedback." In *Handbook of Improving Performance in the Workplace: Volume 2: Selecting and Implementing Performance Interventions*, edited by Doug Leigh and Ryan Watkins, 221–50. San Francisco, Calif: Jossey-Bass.

Kypri, Kypros, Shaun Stephenson, and John Langley. 2004. "Assessment of Nonresponse Bias in an Internet Survey of Alcohol Use." *Alcoholism: Clinical & Experimental Research* 28 (4): 630–34. doi:10.1097/01.ALC.0000121654.99277.26.

Lee, Angela, and Nichelle C. Carpenter. 2018. "Seeing eye to eye: A meta-analysis of self-other agreement of leadership." *The Leadership Quarterly* 29 (2): 253–75. doi:10.1016/j.leaqua.2017.06.002.

Levitt, Barbara, and James G. March. 1988. "Organizational Learning." *Annual review of sociology* 14: 319–40.

Lievens, Filip, Christopher S. Chasteen, Eric A. Day, and Neil D. Christiansen. 2006. "Large-Scale Investigation of the Role of Trait Activation Theory for Understanding Assessment Center Convergent and Discriminant Validity." *The Journal of applied psychology* 91 (2): 247–58. doi:10.1037/0021-9010.91.2.247.

Merton, Robert K. 1957. *Social theory and social structure.* Glencoe, Ill. Free Press.

Mintzberg, Henry. 2009. *Managing.* 1st. San Francisco: Berrett-Koehler Publishers. http://www.worldcat.org/oclc/313653674.

Moynihan, Donald P. 2008. *The dynamics of performance management: Constructing information and reform.* Washington, D.C. Georgetown University Press.

Muterera, Jonathan, David Hemsworth, Anahita Baregheh, and Blanca R. Garcia-Rivera. 2018. "The Leader-Follower Dyad: The Link Between Leader and Follower Perceptions of Transformational Leadership and Its Impact on Job Satisfaction and Organizational Performance." *International Public Management Journal* 21 (1): 131–62. doi:10.1080/10967494.2015.1106993.

Ostroff, Cheri, Leanne E. Atwater, and Barbara J. Feinberg. 2004. "Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes." *Personnel Psychology* 57 (2): 333–75. doi:10.1111/j.1744-6570.2004.tb02494.x.

Pandey, Sanjay K., and Hal G. Rainey. 2007. "Public Managers' Perceptions of Organizational Goal Ambiguity: Analyzing Alternative Models." *International Public Management Journal* 9 (2): 85–112. doi:10.1080/10967490600766953.

Park, Joo-Ho, and Seung-Hwan Ham. 2016. "Whose perception of principal instructional leadership? Principal-teacher perceptual (dis)agreement and its influence on teacher collaboration." *Asia Pacific Journal of Education* 36 (3): 450–69. doi:10.1080/02188791.2014.961895.

Piening, Erk P. 2013. "Dynamic Capabilities in Public Organizations." *Public Management Review* 15 (2): 209–45. doi:10.1080/14719037.2012.708358.

Rammstedt, Beatrice, and Oliver P. John. 2007. "Measuring Personality in One Minute or Less: a 10-Item Short Version of the Big Five Inventory in English and German." *Journal of Research in Personality* 41 (1): 203–12. doi:10.1016/j.jrp.2006.02.001.

Roberts, Tomi-Ann, and Susan Nolen-Hoeksema. 1989. "Sex differences in reactions to evaluative feedback." *Sex Roles* 21 (11): 725–47. doi:10.1007/BF00289805.

Rothstein, Hannah R. 1990. "Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe." *J Appl Psychol* 75 (3): 322–27.

Sala, Fabio. 2003. "Executive Blind Spots: Discrepancies Between Self- and Other-Ratings." *Consulting Psychology Journal: Practice and Research* 55 (4): 222–29. doi:10.1037/1061-4087.55.4.222.

Schedler, Kuno, and Isabella Proeller. 2010. *Outcome-oriented public management: A responsibility-based approach to the new public management.* Charlotte, NC: Information Age Pub.

Seidle, Brett, Sergio Fernandez, and James L. Perry. 2016. "Do Leadership Training and Development Make a Difference in the Public Sector? A Panel Study." *Public Administration Review* 76 (4): 603–13. doi:10.1111/puar.12531.

Shivers-Blackwell, Sheryl L. 2004. "Using Role Theory to Examine Determinants of Transformational and Transactional Leader Behavior." *Journal of Leadership & Organizational Studies* 10 (3): 41–50. doi:10.1177/107179190401000304.

Tett, Robert P., and Hal A. Guterman. 2000. "Situation Trait Relevance, Trait Expression, and Cross-Situational Consistency: Testing a Principle of Trait Activation." *Journal of Research in Personality* 34 (4): 397–423. doi:10.1006/jrpe.2000.2292.

Tummers, Lars G., and Eva Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. doi:10.1111/padm.12224.

Tziner, Aharon, Kevin R. Murphy, and Jeanette N. Cleveland. 2005. "Contextual and Rater Factors Affecting Rating Behavior." *Group & Organization Management* 30 (1): 89–98. doi:10.1177/1059601104267920.

U.S. Office of Personnel Management. 2017. "OPM Leadership 360." https://www.opm.gov/policy-data-oversight/data-analysis-documentation/employee-surveys/buy-services/opm-leadership-360/.

Van de Ven, Andrew H., and Diane L. Ferry. 1980. *Measuring and Assessing Organizations.* New York: Wiley.

van Dorp, Erik-Jan. 2017. "Trapped in the hierarchy: The craft of Dutch city managers." *Public Management Review* 10 (4): 1–18. doi:10.1080/14719037.2017.1383783.

Van Iddekinge, Chad H., Gerald R. Ferris, and Tonia S. Heffner. 2009. "Test of a Multistage Model of Distal and Proximal Antecedents of Leader Performance." *Personnel Psychology* 62 (3): 463–95. doi:10.1111/j.1744-6570.2009.01145.x.

Van Wart, Montgomery. 2003. "Public-Sector Leadership Theory: An Assessment." *Public Administration Review* 63 (2): 214–28. http://www.jstor.org/stable/977592.

Van Wart, Montgomery. 2013. "Administrative Leadership Theory: a Reassessment After 10 Years." *Public Administration* 91 (3): 521–43. doi:10.1111/padm.12017.

Vogel, Dominik. 2016. *Führung im öffentlichen Sektor: Eine empirische Untersuchung von Einflussfaktoren auf das Führungsverhalten.* Potsdam: Universitätsverlag Potsdam. http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-89214.

Vogel, Dominik. 2018. "Matching survey responses with anonymity in environments with privacy concerns." *Intl Jnl Public Sec Management* 31 (7): 742–54. doi:10.1108/IJPSM-12-2017-0330.

Vogel, Rick, and Doris Masal. 2015. "Public Leadership: a Review of the Literature and Framework for Future Research." *Public Management Review* 17 (8): 1165–89. doi:10.1080/14719037.2014.895031.

Wang, N., S. Wilhite, and D. Martino. 2016. "Understanding the relationship between school leaders social and emotional competence and their transformational leadership: The importance of self-other agreement." *Educational Management Administration & Leadership* 44 (3): 467–90. doi:10.1177/1741143214558568.

Wofford, J. C. 1982. "An Integrative Theory of Leadership." *Journal of Management* 8 (1): 27–47. doi:10.1177/014920638200800102.

Yukl, Gary A. 2012. "Effective Leadership Behavior: What We Know and What Questions Need More Attention." *Academy of Management Perspectives* 26 (4): 66–85. doi:10.5465/amp.2012.0088.

Yukl, Gary A., Steve Wall, and Richard Lepsinger. 1990. "Preliminary Report on Validation of the Managerial Practices Survey." In *Measures of Leadership*, edited by Kenneth E. Clark and M. B. Clark, 223–37. West Orange, N.J. Leadership Library of America.
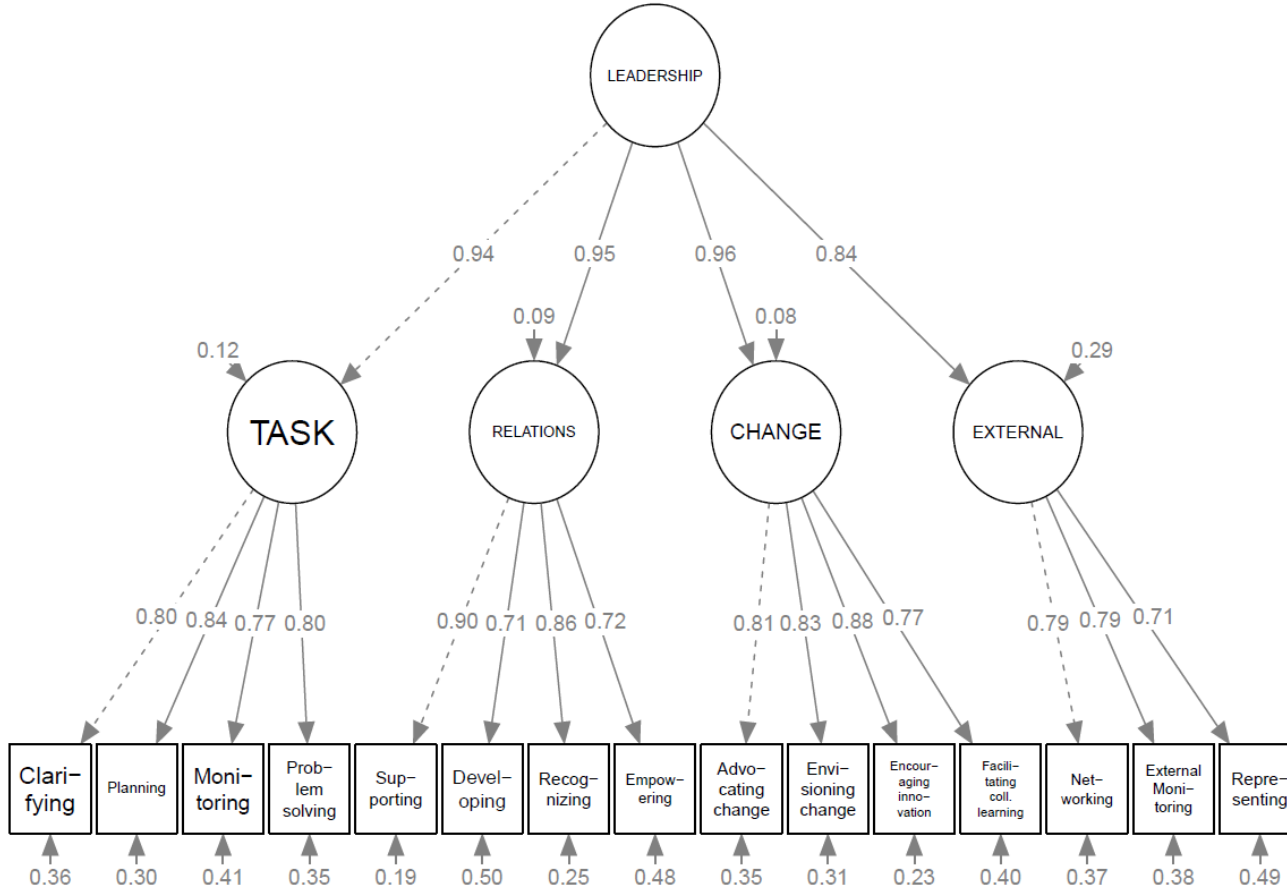
Appendix 1: Operationalizations

Variable	Operationalization
Leadership (Yukl, Wall, and Lepsinger 1990)	The Managerial Practices Survey (MPS) is not freely accessible. However, the wording of the items is quite close to a specific aspect of the definition of the corresponding behaviours (Yukl 2012, 84–85). Hence, the definitions and the item number of version G 17-4 of the MPS are displayed. • Clarifying: sets specific goals and deadlines for important aspects of the work (No. 4) • Planning: develops short-term plans for the work (No. 17) • Monitoring: checks on the progress and quality of the work (No. 33) • Problem solving: identifies work-related problems that can disrupt operations (No. 51) • Supporting: provides support and encouragement when there is a difficult or stressful task (No. 8) • Recognizing: provides recognition for member achievements and contributions to the organization (No. 22) • Developing: encourages members to take advantage of opportunities for skill development (No. 40) • Empowering: delegates responsibility and authority to members for important tasks and allows them to resolve work-related problems without prior approval (No. 53) • Envisioning Change: communicates a clear, appealing vision of what could be accomplished (No. 12) • Encouraging Innovation: encourages innovative thinking and new approaches for solving problems (No. 25) • Facilitating Collective Learning: no corresponding element in the definition (No. 41). Full definition: uses systematic procedures for learning how to improve work unit performance; helps members understand causes of work unit performance; encourages members to share new knowledge with each other. • Advocating Change: explains why a policy or procedure is no longer appropriate and should be changed (No. 57) • Representing: negotiates agreements and coordinates related activities with other parts of the organization or with outsiders (No. 32) • Networking: no corresponding element in the definition (No. 46). Full definition: attends meetings or events; joins professional associations or social clubs; uses social networks to build and maintain favorable relationships with peers, superiors, and outsiders who can provide useful information or assistance. • External Monitoring: analyzes information about events, trends, and changes in the external environment to identify threats, opportunities, and other implications for the work unit. (No. 16)

Supervisor's interest (α = .75) (based on Van de Ven and Ferry 1980)	• My supervisor is interested in how I lead • My supervisor regularly talks with me about how I lead my team (1 = do not agree at all; 5 = totally agree)
Motivation to lead (Chan and Drasgow 2001)	I usually want to be the leader in the groups that I work in (1 = do not agree at all; 5 = totally agree)
Leadership experience	For what number of years have you had a leadership position? 1 = 0–5 years 2 = 6–10 years 3 = 11–15 years 4 = 16–20 years 5 = 21–25 years 6 = 26–30 years 7 = more than 30 years
Leadership training	I participated in one or more leadership trainings after being appointed as a leader. 0 = no; 1 = yes
Job satisfaction (Hackman and Oldham 1974)	Generally speaking, I am very satisfied with my job (1 = do not agree at all; 5 = totally agree)
Gender	Are you … 0 = male; 1 = female
Age	How old are you? 1 = younger than 25 years 2 = 26–35 years 3 = 36–45 years 4 = 46–55 years 5 = older than 56 years
Managerial reflection routines (α = .93)	• My team and I regularly optimize our processes in order to minimize process duration and streamline interactions with other units. • My team and I regularly discuss our goals and define new goals. • My team and I regularly reflect on our work outcomes. • My team and I regularly reflect on our clients' expectations of our work and adapt practices accordingly. • My team and I have regular events, where we proactively discuss how to improve our performance. (1 = do not agree at all; 5 = totally agree)

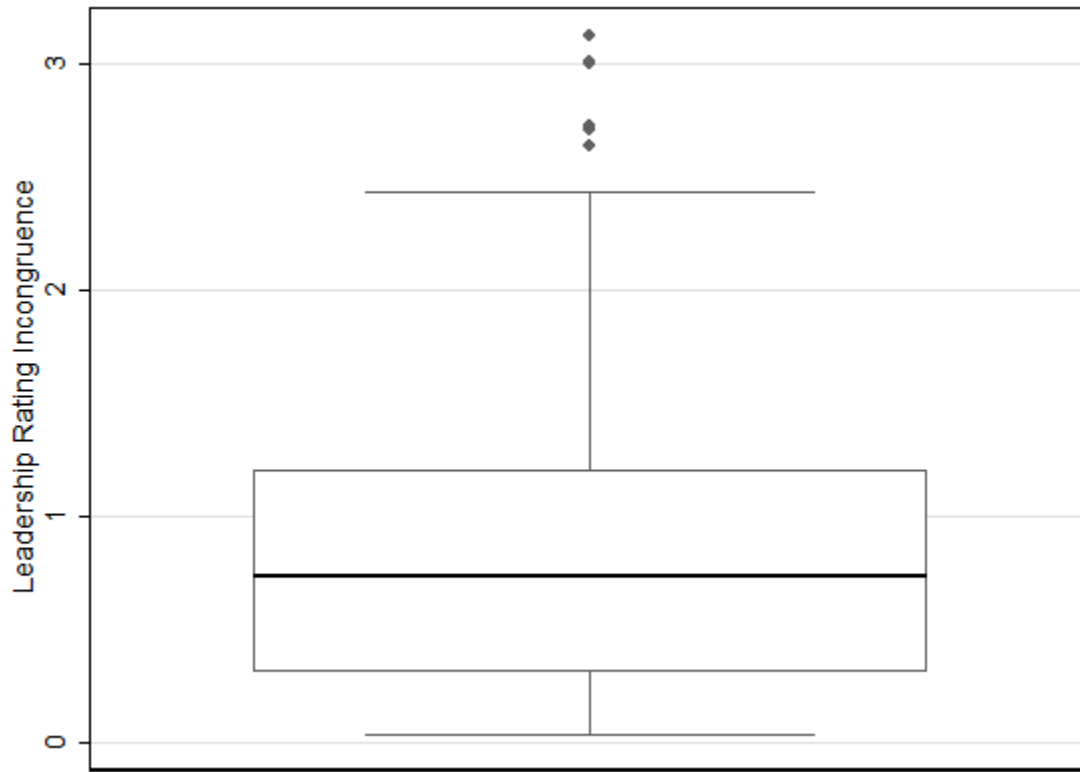| Big five personality traits (Rammstedt and John 2007)	I see myself as someone who …	Extra-version	Openness	Consci-en-tiousness	Neuroti-cism	Agreeable-ness
	… is reserved (R)	.59	.01	.07	-.00	-.10
	… is outgoing, sociable	.59	-.01	-.02	-.10	.23
	… is generally trusting	.11	.08	-.09	.14	.47
	… tends to find fault with others (R)	-.15	-.13	.13	-.15	.26
	… is considerate and kind to al-most everyone	-.06	.10	.30	-.01	.38
	… tends to be lazy (R)	.14	-.07	.48	.03	.05
	… does a thor-ough job	-.01	.14	.50	.04	.00
	… is relaxed, handles stress well (R)	-.05	-.06	.17	.46	-.02
	… gets nervous easily	-.21	-.01	-.13	.42	.19
	… has few artistic interests (R)	-.03	.44	-.03	-.05	.08
	… has an active imagination	.02	.55	.10	-.02	.03

Rotated factor loadings (principal factor, PROMAX rotation)
(1 = do not agree at all; 5 = totally agree)

Appendix 2: Confirmatory Factor Analysis

Appendix 3: Boxplot of Leadership Rating Incongruence