The Leadership Arena-Reputation-Identity (LARI) Model: Distinguishing shared and unique perspectives in multisource leadership ratings

Jasmine Vergauwe^{a*}, Joeri Hofmans^b, and Bart Wille^a

^aGhent University, Belgium ^bVrije Universiteit Brussel, Belgium

Accepted for publication in the Journal of Applied Psychology (January 30, 2022)

*Address correspondence to: dr. Jasmine Vergauwe, Department of Developmental, Personality, and Social Psychology, Ghent University. H. Dunantlaan 2, B-9000 Gent. Belgium. Jasmine.Vergauwe@ugent.be Tel.: +32 9 264 64 29

Funding: This work was supported by the Research Foundation – Flanders (FWO), grant number 1200919N.

Acknowledgements: We would like to thank Robert B. Kaiser (Kaiser Leadership Solutions) and Patrick Vermeren (PerCo) for enabling the use of their multisource data in this article (i.e., Sample 1 and 2 respectively). Further, we thank Louis Tay for his valuable and constructive feedback on an earlier version of the manuscript.

© 2022, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/apl0001012

Abstract

Multisource leadership ratings rely on the assumption that -in addition to the leader's selfevaluation- different rater groups (i.e., subordinates, peers and superiors) bring in unique perspectives and thus provide a more well-rounded analysis of the leader's behavior. However, the way in which multisource data are typically treated in research offers little information about the precise levels of overlap and uniqueness that are encapsulated in these different perspectives. Drawing on the Trait-Reputation-Identity (TRI) model, we propose a model that conceptualizes these shared and unique perspectives in terms of latent factors reflecting respectively (i) the consensus about the leader (i.e., the Leadership Arena), (ii) the impressions conveyed to others that are distinct from self-perceptions (i.e., the leader's Reputation), and (iii) the unique self-perceptions of the leader (i.e., the leader's Identity). This Leadership Arena-Reputation-Identity (LARI) model is formalized by means of bifactor modeling, which allows to statistically decompose the variance captured by multisource ratings. The LARI model was tested against five alternative models in two large multisource samples ($N_{1 \text{ leaders}} = 537$, N_{1} observers = 7,337; N₂ leaders = 1,255, N₂ observers = 15,777), each using different leadership instruments. In both samples, the LARI bifactor model outperformed the alternative models. A subsequent variance decomposition showed that each rater source indeed provides unique information about the target's behavior, although in varying degree. Across all leadership dimensions in both samples, superiors consistently provided the largest share of unique information among the three observer groups. Implications and future directions are discussed.

Keywords: multisource leadership ratings, 360-degree feedback, self-other agreement, unique leadership perceptions, bifactor modeling

The Leadership Arena-Reputation-Identity (LARI) Model: Distinguishing shared and unique perspectives in multisource leadership ratings

Across various disciplines of psychology, researchers and practitioners have shown great interest in how people's self-perceptions converge with perceptions by others (Connelly & Ones, 2010; Fleenor et al., 2010). In the leadership domain, in particular, it is typically argued that comparing self-ratings to other-ratings on key leadership dimensions highlights areas of agreement and disagreement (i.e., "blind spots"), which can be used to inform leadership training and development (e.g., Day et al., 2014). In addition, it allows tapping leaders' levels of selfawareness regarding their leadership behaviors and capabilities, which has also been linked to important performance outcomes (Atwater & Yammarino, 1992; Fleenor et al., 2010). Because of these reasons, multisource rating procedures and their feedback programs, often referred to as 360s, are widely used in leadership assessments across the globe (Slater & Coyle, 2014).

At the same time, however, research has questioned the use of 360s in organizational contexts (e.g., LeBreton et al., 2003; Mount et al., 1998; Scullen et al., 2000). Critics refer to the administrative burden and relatively high implementation costs associated with having multiple people rating one and the same leader. Central to this discussion is the question of whether and to what extent 360s indeed deliver what they promise. *Do they really allow uncovering different perspectives on a target leader's behavior*? Because only in the case that this is true, organizations can justify the implementation of burdensome and costly 360s as an investment for the future.

Multisource leadership ratings are based on the deep-rooted belief that, next to the leader themselves, individuals from different rater groups (i.e., supervisors, peers, subordinates) provide unique perspectives on a leader's behavior and performance (Borman, 1997). This is known as

the 'discrepancy hypothesis' and is purported to occur because of a variety of reasons, including leaders not acting similarly around all rater groups, and different rater groups having different opportunities to observe leaders (Borman, 1997; Fleenor et al., 2010; LeBreton et al., 2003; Lee & Carpenter, 2018; Tornow, 1993). For about 40 years already, leadership scholars have been testing this assumption by comparing self- and other-ratings and evaluating their levels of convergence (e.g., Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Lee & Carpenter, 2018; Mount, 1984). Consistent with the idea that these ratings represent a mixture of overlap and uniqueness, meta-analyses show moderate correlations between different observer sources (i.e., ranging between .22 and .32 for managerial jobs; Conway & Huffcutt, 1997), as well as moderate correlations between leader- and observer-ratings of leader behavior (i.e., in the .30 range across leadership dimensions, as well as across types of observers; Lee & Carpenter, 2018).

The relevance of multiple rater groups has also been extensively examined in research on the structure and reliability of multisource performance ratings (MSPRs) (e.g., Hoffman et al., 2010; Lance et al., 2008; Mount et al., 1998; Scullen et al., 2000; Viswesvaran et al., 2005). A large body of literature has relied on analytical approaches such as (second-order) factor analyses (e.g., Hoffman et al., 2010) and/or multitrait-multi*rater* analyses (e.g., Woehr et al., 2005) to identify the sources of variance underlying MSPRs. As the most recent example in this research tradition, Jackson et al. (2020) used Bayesian generalizability theory to get an estimate of the proportion of true score variance in multisource ratings associated with ratee-, rater-, source-, and dimension-related effects. In contrast to earlier studies that downplayed the importance of source effects (Mount et al., 1998; Scullen et al., 2000), Jackson et al. (2020) demonstrated that large portions of variance in MSPRs are attributable to source-related effects, cumulatively

explaining between 30.90 and 58.06% of the variance in ratings across samples. Such large source effects again underscore the idea that different rater sources provide unique perspectives on the leader.

Taken together, the literatures on self-other agreement and on the structure of MSPRs have a long history in applied psychology, and they seem to converge on the notion that different rater sources can indeed contribute both unique and shared information.Yet, very little is known about the exact nature of these different source effects and their size relative to each other. This raises the fundamental question: "How important are each of the four typical rater sources (i.e., self, peers, subordinates, superiors) in multisource leadership ratings?"

The Relevance of Scrutinizing Unique Rater Source Effects

Studying the existence and magnitude of these various rater source effects is important for theoretical reasons, while it might also provide crucial insights for leadership research and practice. First, from a theoretical point of view, it aligns well with the fundamental idea of the attribution theory of leadership, which assumes that leadership is determined, to a certain extent, by the perceptions of people surrounding the focal leader (e.g., Foti et al., 1982; Hogan, 1996; Lord et al., 1984, 2020). For example, research has shown that through socialization and earlier experiences with leaders, people develop cognitive structures or prototypes of leadership categories (i.e., implicit leadership theories), and that people rely on those cognitive structures or prototypes when evaluating a particular leader (Lord et al., 2020). Importantly, being *perceived* as a leader influences both self- and observers' assessments of leadership, and ultimately matters for individual- (e.g., career progress), team- (e.g., cohesion), and organizational outcomes (e.g., performance; Lord et al., 2020). Similarly, socio-analytic theory (Hogan, 1996) emphasizes the importance of people's reputations (i.e., observers' perceptions) on work outcomes such as

3

career success (Hogan & Blickle, 2018). Although a great deal has already been written about leadership being in the eye of the beholder (e.g., Lord et al., 2020; Nye, 2002), researchers seem to focus on these observer perceptions without taking into account potential overlap with leaders' self-perceptions. Separating observer perspectives from the leader's self-perception could potentially advance these attributional theories by shedding light on the exact extent to which leadership lies uniquely 'in the eye of the beholder'. Although part of these observer perceptions will be rater-specific, there will also be commonality, capturing the leader's general "reputation". A subsequent question is then whether and to which extent this unique observer perspective can be further decomposed in different source-specific reputation factors that lie uniquely in the eyes of peers, subordinates, and superiors, respectively.

Further, knowledge about the uniqueness of each rater source can also inform basic leadership research. As mentioned earlier, one of the central premises of 360-degree assessments is that different rater groups contribute unique information to the leader's overall assessment (Borman, 1997). However, the traditional treatment of multisource data does not allow for an empirical test of this central assumption. Most obviously, research aggregating the ratings across the different rater sources (e.g., Kaiser et al., 2015) represents a rudimentary way to account for the variability in rater perspectives. Although the resulting observer score might be 'rich' in the sense that all different perspectives contribute to this score, this approach neglects the possibility that the different sources diverge in their perceptions to an important extent (e.g., Jackson et al., 2020). A second, more common, approach in leadership research is to compare the different observer perspectives separately to the self-rating (e.g., Atwater et al., 2009), or to relate them separately to external variables of interest (e.g., Aramovich & Blankenship, 2020; Atwater & Brett, 2006). A crucial difficulty here is that, besides unique variation, each rater group also

shares variation with the other rater groups (Conway & Huffcutt, 1997). Because the exact composition of shared and unique variation between rater groups is unknown, it is unclear what the 'source-specific' effects actually reflect. Knowing these exact variance compositions does not only allow testing the central premise of 360s (i.e., the discrepancy hypothesis), it may also provide new insights regarding the way multisource ratings should be treated in research (e.g., aggregation methods), while revealing new perspectives on the construct of self-other agreement.

Finally, from a practical point of view, knowing which percentage of variance is captured by the self and the various observer groups gives an idea of the usefulness or even the necessity of including these different rater groups in multisource leadership ratings. Organizations can approach the use of these procedures in terms of a trade-off between collecting as much information as possible and keeping the implementation cost (and administrative burden) as low as possible. Knowing that the unique input provided by one or more rater groups is small or even negligible may help organizations in deciding which rater groups to include or not.

In sum, although source effects are "alive and well" (Hoffman et al., 2010; Jackson et al., 2020), much remains to be learned about the degree to which rater groups involved in 360s converge and diverge in their perspectives. To fully exploit the benefits of multisource leadership ratings, both theoretical and methodological innovations are needed. Theoretically, an advanced application of the Trait-Reputation-Identity (TRI) model (McAbee & Connelly, 2016) is proposed that can facilitate our conceptual understanding of these different rater source effects by separating consensus about the leader, unique self-perceptions of the leader, and unique observer-perceptions. Methodologically, we provide a tool to distinguish and quantify these shared and unique perspectives in multisource data. Before outlining our framework in greater

detail, we provide an overview of the different theoretical mechanisms that may explain why various rater groups may hold unique perspectives on a focal leader.

Theoretical Perspectives on Leader Self- and Observer Ratings

There are several accounts of why raters from different groups can hold unique perspectives on a target's behavior, and those accounts can broadly be categorized into those that refer to (i) role differences, (ii) informational differences, and (iii) motivational differences. First, different rater groups might be viewing the focal leader through different lenses due to evaluating the leader in varying roles (Atwater & Yammarino, 1997; Yammarino, 2003). In particular, superiors rate them as followers, peers rate them as coworkers, subordinates rate them as managers, and leaders may rate themselves as a blend of all of these roles (Lee & Carpenter, 2018). Each of these different roles may come with different ideas and beliefs about what constitutes "leadership", which could create differences between these perspectives (e.g., Hooijberg & Choi, 2000).

Second, particular rater groups can have more knowledge of specific leader behaviors because there is simply more information at their disposal to evaluate these particular behaviors (i.e., information availability) or because there is a higher chance of noticing particular behaviors (i.e., information detection) (Funder, 2012; Rothstein, 1990; Vazire, 2010). Self- and otherperceptions are, for instance, likely to differ in quantity and type of information available. The self has a major advantage over others because "no one else has access to more information" (p. 277; Paulhus & Vazire, 2007). By the same token, although –theoretically– the self can observe most of its own behaviors, it is unlikely to *detect* many of these behaviors. As oneself is not as salient in one's visual field as it is in others', some behaviors will not be detected. Moreover, even if people detect their own overt behavior, the self usually places more weight on thoughts

6

and feelings than on overt behaviors when forming self-perceptions. This effect is usually reversed when forming perceptions of others (Andersen, 1984). Against this background, the self-other knowledge asymmetry (SOKA) model (Vazire, 2010) expects that the leader themselves knows more than others about low-observable behaviors. Highly observable leader behaviors, on the other hand, are at least equally known to others as compared to the self (Vazire, 2010).

In a similar vein, informational differences can also exist between the different observer groups (i.e., superiors, peers, subordinates). This idea aligns with Funder's (1995, 2012) realistic accuracy model, which holds that observer accuracy depends on the extent to which behavioral cues are available to observation and the extent to which they are detected by an observer. In general, subordinates may have more knowledge of leadership behavior than would superiors, given that subordinates are the direct recipients of leader behaviors, whereas superiors are usually not (Hansbrough et al., 2015; Lee & Carpenter, 2018). However, some behaviors are easier to witness for particular (and not all) groups of observers, as leaders may not act similarly around all rater groups (Borman, 1997; Yukl, 2010). Hence, it is likely that some type of leader behaviors can only be witnessed in certain contexts. Subordinates may, for instance, have more knowledge of interpersonal leader behavior than would superiors, whereas superiors may have more unique knowledge of strategic or business-related leader behavior (Hiller et al., 2011).

Finally, social information processing theory (Salancik & Pfeffer, 1978) predicts that different motivational sources may drive leader- and the various observer ratings. Leaders themselves can be susceptible to self-enhancement bias and inflate their ratings in order to present themselves in a favorable light to others or to protect their own self-image (Atwater et al., 1998). Moreover, leaders may be well aware that the impression they make on others has implications for how others perceive and evaluate them. Therefore, leaders can be motivated to control the impressions others form of them, behaving in ways that create certain impressions in others' eyes (Leary & Kowalski, 1990). Observers, in turn, may also be vulnerable to biasing motives, as they may intentionally provide inflated ratings of leaders to avoid punishment from leaders or other forms of backlash (Lee & Carpenter, 2018). Not surprisingly, raters who pursue to (a) identify the target's weaknesses or (b) strengths, (c) provide fair ratings or (d) motivate the target, eventually provide different ratings of the same behavior. So raters pursuing different goals may actually give different ratings (Murphy et al., 2004). Hence, shared intentions among raters of the same group would also contribute to unique perspectives on the target's behavior (e.g., subordinates may want to emphasize the leader's strengths, whereas superiors may want to uncover developmental possibilities).

The Leadership Arena-Reputation-Identity (LARI) Model

The objective of the current paper is to propose a framework for conceptualizing and quantifying the different rater source effects in multisource leadership ratings. To this end, we turn to recent developments in personality psychology, a field that has amassed a substantial body of knowledge about how strong consensus is across raters and how self- versus other-ratings of personality uniquely contribute to the prediction of relevant outcomes at work and beyond (e.g., Connelly & Ones, 2010; Connolly et al., 2007).

Recently, McAbee and Connelly (2016) developed a formal framework for studying multi-rater personality data, teasing apart the different ways in which a person's personality is constructed by oneself and by others. The basis of their model is the Johari window (Luft & Ingham, 1955), which maps information known versus unknown to the self and observers in a two-by-two grid (see left part of Figure 1). The "Arena" reflects information that is shared

between the self and others (i.e., known to the self and observers). The "Facade" captures selfknowledge that is not shared with others (i.e., known to the self but unknown to observers). The "Blind-Spot" represents aspects of a target that others see, but the target is unaware of or does not endorse (i.e., known to observers but unknown to the self). Finally, the Johari window acknowledges that some information remains "Unknown" to (or not perceived by) both the self and observers (Luft & Ingham, 1955). McAbee and Connelly (2016) relabeled the quadrants of the Johari window as Trait (Arena), Reputation (Blind-Spot), and Identity (Facade) because these terms aligned better with long traditions of personality research. The "Trait" label reflects a historical emphasis on corroborating traits through consensual validation (Campbell & Fiske, 1959). In contrast, the distinction between "Reputation" and "Identity" has roots in symbolic interactionism (Blumer, 1986) and socioanalytic theory (Hogan, 1996). In short, the Trait-Reputation-Identity (TRI) Model (see right part of Figure 1) was a breakthrough in personality science by separating the variability in multisource personality ratings into (i) consensus about the person (i.e., Trait), (ii) unique self-perceptions (i.e., Identity), and (iii) impressions conveyed to others that are distinct from self-perceptions (i.e., Reputation).

-----Insert Figure 1 about here-----

As recognized by McAbee and Connelly (2016), much of the logic behind the TRI model can be directly transferred to the context of multisource leadership ratings. There are certain aspects of a leader's behavior that everyone (i.e., self and all others) agrees upon. There are also certain aspects that will be unique to the perception of the leader or to the perception of others. Yet, a truly unique feature of *multisource leadership ratings* is that the others or 'observers' are not interchangeable. They instead cluster in conceptually meaningful subgroups, such as (but not limited to) work peers, supervisors, and subordinates. As explained above, distinguishing these different observer groups is important because differences in roles, information, and motivations will likely result in higher homogeneity between people belonging to the same rater groups and more divergence between individuals belonging to different rater groups (e.g., Borman 1997; Lee & Carpenter, 2018). Hence, applying the TRI model to multisource leadership ratings requires a significant extension to the model that allows each of these observer groups to form their own unique reputation of a target leader.

One particular extension that has been proposed by McAbee and Connelly (2016) adds multiple rater contexts to the Reputation factor, acknowledging that personality manifestations can be different in different environments (e.g., at home, at work, among friends; see Figure 4 on p. 583). Although rater *contexts* represent rater groups in reference to multisource leadership ratings, our proposed LARI model provides a similar extension by separating the variance in multisource leadership ratings into six distinguishable latent factors: The Arena, the shared Reputation, the Reputation by peers, by subordinates, and by supervisors, and the Identity (see Figure 2). A formal description of these LARI factors along with an overview of elements contributing to the respective factors is provided in Table 1. The Arena factor captures information about the target's leadership behavior that is shared between the self and all others. Going back to the original Johari window, we prefer the term "Arena" over "Trait" (cf. TRI) as the former reflects better what this means in a leadership context. Specifically, this factor captures the features of the target's behavior on which all raters -including the leader themselves- agree. This arena is the place where private leader identity and public leadership reputation overlap.

A second factor is called general *Reputation* and refers to information that is shared by (all) observer groups but is unique from the perspective provided by the leader themselves. This external reputation can reflect (i) information not available to the self, (ii) information intentionally not shared by the self (but picked up by all observers), or (iii) systematic bias shared across all observer groups (e.g., physical appearance stereotypes or a leniency bias). Importantly, this general reputation factor reflects only what is shared across the different observer groups, whereas three source-specific reputation factors capture perceptions that are unique to the perspectives of raters from the same source. This feature of the LARI model is based on the idea that a leader's subordinates, peers, and supervisors each have unique interactions with and expectations of the leader (Lee & Carpenter, 2018).

Specifically, the Reputation by *subordinates* reflects what is unique in the perspective provided by subordinates, and not shared with the views provided by superiors, peers, or the leader themselves. Considering that much of a leader's behavior involves and is directed towards subordinates (Hansbrough et al., 2015), and subordinates have many opportunities to witness their leader in a leadership role (Conway et al., 2001), subordinate-ratings are probably the most common source of observer-rated leadership. Moreover, compared to the leader's peers and superiors, subordinates should have a unique perspective on various leader behaviors including their motivational style and individualized consideration (Hiller et al., 2011). *Peers*, relative to subordinates, likely have less opportunities to observe leader behavior (i.e., in part because they are themselves leaders). On the other hand, peers may also have the chance to observe a range of leader behaviors that other individuals in the organization rarely see (Braddy et al., 2014), such as behavior related to alignment, positioning, and boundary spanning (Hiller et al., 2011).

relative to subordinates and peers (Pollack & Pollack, 1996). However, it is likely that superiors monitor different aspects of a leader's functioning compared to the other two sources. Supervisors might be more motivated to notice leader behaviors aimed at realizing the company's strategic goals. Also, because of their hierarchical position, superiors may have different information at their disposal to evaluate or describe a focal leader. The performance of a target leader's unit relative to that of other units might, for instance, be better judged by superiors in the higher echelons.

Finally, the leader *Identity* factor represents self-perceptions of leadership that are not shared with others. Leadership research on self-other agreement has referred to the notion of "bias" in self-ratings to explain self-other discrepancies. This bias represents inaccurate self-perceptions that may arise for several reasons (Fleenor et al., 2010), including the limited opportunities to receive feedback from followers, thus limiting (dis)confirmation, and the lack of motivation to use followers' perceptions as relevant feedback on their behavior. In addition to these explanations that refer to "bias" or "inaccuracies", the uniqueness of leaders' self-perceptions may also reflect behaviors or attitudes that are simply low in observability (cf. SOKA model; Vazire, 2010), such as private goals and strivings that are not fully expressed. Such "hidden knowledge" thus captures relevant variance in leader behavior that observers are simply never exposed to (McAbee & Connelly, 2016). Indeed, coworkers probably witness only a portion of the leader's full repertoire of behavior (e.g., Allen et al., 2000), and are more likely to only witness and remember the results of the leader's behaviors (DeNisi et al., 1984).

Research Questions

Drawing on various theoretical perspectives on self-other and other-other (dis)agreement in leadership perceptions, it becomes clear how each of the informant groups in multisource leadership ratings may bring in perspectives that are partly distinctive and partly overlapping. The proposed LARI model scrutinizes the different rater source effects by disentangling these shared and unique perspectives for various leadership dimensions. As a first research question (RQ1), the viability of this model will be explored.

RQ1: Does the presented LARI model, decomposing the shared and unique perspectives of the four typical rater sources, fit multisource leadership data well?

In a second research question (RQ2), we will focus on the relative size of the different LARI factors. Grounded in the realistic accuracy model (Funder, 2012), the relative size of the unique observer perceptions of *subordinates* is generally expected to be larger, compared to the unique perceptions of peers and superiors, as subordinates would have more opportunities to observe and interact with focal leaders (e.g., Conway et al., 2001; Hansbrough et al., 2015; Hiller et al., 2001; Lee & Carpenter, 2018). Importantly, however, in the current study, a number of leadership dimensions will be taken into account, each focusing on substantially different aspects of leadership. A relevant related question then is whether the relative size of the different LARI factors depends on the specific leadership dimension being assessed (RQ3). On the one hand, different theories would expect this to be the case. Building on the SOKA model (Vazire, 2010), a crucial factor in this regard is the observability of those leadership dimensions. For highly observable dimensions, the shared perceptions (i.e., among everyone (Arena) and among observers (general Reputation)) are expected to be relatively large, whereas the unique selfperceptions of the leader (Identity) are expected to be relatively small. For dimensions that are more difficult to observe, a reversed pattern could be expected. Further, as leaders do not act similarly around members of different rater groups (Borman, 1997; Yukl, 2010), a larger relative size of the Reputation by subordinates might be particularly likely for interpersonal leadership

dimensions, whereas peers and superiors could be more knowledgeable about strategic or business-related leadership dimensions (Hiller et al., 2011). These expectations are further in line with the theoretical perspective offered by Guion (1965), which holds that source variation may occur because sources systematically differ in the dimensions they use to evaluate people. In contrast to these theoretical arguments, however, Jackson et al. (2020) recently provided evidence that different source perspectives do not depend on dimension-based evaluations. Instead, their evidence suggested that different rater sources "essentially bypass specific dimensions altogether and form source-dependent overall impressions of ratees" (p. 325). As such, the final two research questions are:

RQ2: What is the relative size of the different LARI factors? And does this generalize across samples and instruments?

RQ3: Is the relative size of the different LARI factors dependent on the specific leadership dimension being assessed?

Method

We analyzed two large multisource datasets, each of which was operated by a different consultancy firm specialized in leadership assessment, and each charting a different set of leadership dimensions. A first international sample rated *four* leader behaviors (using the Leadership Versatility Index (LVI); Kaiser et al., 2010), and a second –mainly Belgian– sample rated *eight* leadership styles (using the Circumplex Leadership Scan (CLS); Redeker et al., 2014). As further detailed below, the CLS exclusively maps interpersonal leadership dimensions, whereas the LVI covers both interpersonal and organizational leadership dimensions. Moreover, the CLS and the LVI also differ in terms of the rating scale format that is used to measure these leadership dimensions. Whereas the CLS uses a more traditional frequency-type Likert scale

ranging between 0 (*never*) and 4 (*always*), the LVI uses a less conventional "too little/too much" (TLTM) scale ranging from -4 (*much too little*), over 0 (*the right amount*) to +4 (*much too much*). Due to variance from the "too much" range (from 0 to +4), TLTM ratings of leadership provide incremental validity over Likert ratings in predicting performance (Vergauwe et al., 2017). Hence, the TLTM scale's ability to capture behavioral excess can be assumed to provide 'broader' assessment output with regard to the underlying leadership dimension. In sum, a number of differences can be observed between the two study samples, including the (inter)nationality of the targets, the type of leadership dimensions, and the rating scale format. However, cross-validating the LARI model using these different samples and leadership instruments is critical to this study as it will allow to strengthen our conclusions regarding the viability of the presented LARI framework. All research conducted in this study was approved by the Ethical Committee of the Faculty of Psychology and Educational Sciences of Ghent University (Study title: Shedding new light on 360° leadership assessment: A multi-rater framework for studying leadership; Reference number: 2018/54).

Transparency and Openness

Existing data from two consultancy firms were re-analyzed. We describe the study samples based on the available information in the Sample sections below. We also describe all data exclusions and all measures in the study, and we adhered to the Journal of Applied Psychology methodological checklist. Data were analyzed using Mplus 8.4 (Muthén & Muthén, 1998-2017). The analysis code (Mplus syntaxes) for testing all models (RQ1) and for the variance decomposition of the LARI factors (RQ2) is available on the Open Science Framework (OSF) via <u>this link</u>, as are the detailed results of these analyses in both study samples. Data, as well as the measures described in the method section, are not available due to their proprietary nature. The multisource data of Sample 1 *partially* overlap with two other publications, as they represent an extension of the multisource data in Kaiser et al. (2015) and Study 3 of Vergauwe et al. (2018). Specifically, more target leaders went through a development center, so the sample size is almost twice as large. The study design, research questions, and the analyses were not preregistered.

Sample 1

Participants. A first multisource dataset was obtained from a U.S.-based consultancy firm involved in leadership assessments across the globe. Next to self-ratings of the leaders (N = 537), an average of 14 raters (min. 3; max. 55) rated each leader in terms of leader behaviors, including at least one subordinate, one peer, and one superior. A total of 7,337 observers, including 1,142 supervisors, 2,695 peers, and 3,500 subordinates provided ratings in the context of a development center. Target leaders were on average 45.59 years old (SD = 7.97) and 72.5% were male. The geographic region of employment was highly diverse. Seventy-five percent of the leaders worked in North America, 17% in Western Europe, 3% in Africa, 3% in East Asia, and a smaller percentage was active in Latin America (0.4%), India (0.4%), Caribbean (0.4%), Australia (0.2%) and the Pacific Islands (0.2%). In terms of organizational level, the targets operated as supervisor (21.4%), middle manager (12.1%), functional head (28.1%), C-level executive (6.5%), or general manager (16.6%), while 15.3% indicated 'other' in this regard. The leaders were active in a variety of industries (e.g., IT, banking, aerospace, construction).

Measure. Target leaders, as well as the observers, provided ratings on the 48 items of the original English version of the Leadership Versatility Index (LVI), tapping into four leader behaviors: forceful, enabling, strategic, and operational (Kaiser et al., 2010). Whereas forceful and enabling leadership represent interpersonal leadership dimensions, strategic and operational

leadership represent organizational leadership dimensions. Each of the four leader behaviors includes three subscales. Forceful leadership is defined as assuming authority and expecting a lot from other people by (a) taking charge, (b) declaring themselves, and (c) pushing for performance. Enabling leadership concerns creating conditions for others to contribute through (a) empowerment, (b) participation, and (c) support. Strategic leadership is defined as positioning the team for the future by (a) setting direction, (b) stimulating growth, and (c) supporting innovation. Operational leadership, finally, refers to guiding the team to execute nearterm goals by (a) specifying the details of implementation ('execution'), (b) focusing resources ('efficiency'), and (c) managing in a process-oriented way ('order'). Each of the 12 subscales were measured by means of four items that were rated on a 9-point scale ranging between -4 (*much too little*), 0 (*the right amount*), and +4 (*much too much*) (Kaiser et al., 2010). Table 2 shows that both the self-ratings of the LVI scales as well as the observer ratings showed a high level of internal consistency (α 's between .69 and .84 for the self-ratings; and between [.72 - .89], [.75 - .90], and [.70 - .88] for the supervisor-, peer-, and subordinate ratings respectively).

To obtain model indicators for each of the four leader behaviors, subscale scores were aggregated across all individuals belonging to a specific rater group, such that 12 aggregated subscale scores were obtained for each of the three observer groups (e.g., a score on 'taking charge' for the subordinate-, peer-, and supervisor group). The $r_{wg(j)}$ interrater agreement coefficient (IRA; James et al., 1984) was computed for each subscale within superior, peer, and subordinate groups. The results in the Appendix (Table A) indicate that the level of similarity within the different rater groups was sufficiently high to support aggregation (LeBreton & Senter, 2008). All descriptive statistics, correlations, and internal consistencies of the variables in Sample 1 are reported in Table 2.

-----Insert Table 2 about here-----

Sample 2

Participants. A second multisource dataset was obtained from a Belgian HR consultancy firm. Leaders (N = 1,863) were assessed for developmental purposes, and a subset of 1,255 leaders that were rated by all three observer groups was retained for this study. In about 40% of the cases, demographical information is missing, as these were optional items. To have an idea of the sample characteristics, however, an estimate was made based on the complete cases. The mean age of the target leaders was 44.17 (SD = 8.80), and 75.1% was male. The targets' country of origin was largely Belgium (89.3%) and the Netherlands (9%), and a small percentage (1.7%) originated from other countries (i.e., 0.5% from France, 0.5% from South-Africa and Niger, and 0.7% from the United Arab Emirates, Spain, Croatia, Turkey and Italy). Their current organizational level was described as non-management (22.7%), front-line management (supervisor) (24.8%), middle management (manages managers) (29.7%), senior management (14.7%), executive (reports to CEO) (5.6%), and CEO (2.6%). In terms of industries, the sample was highly heterogeneous (e.g., 30.2% manufacturing, 10% information and communication, 9.1% human health/social work). An average of 12 raters (min. 3; max. 39) rated each leader, including at least one subordinate, one peer, and one superior. In total, 15,777 observers participated in this study (i.e., 2,175 supervisors, 5,068 peers, and 8,534 subordinates).

Measure. Both the target leaders and the observers provided ratings on the Circumplex Leadership Scan (CLS; Redeker et al., 2014), comprising 116 items that tap into (positive and negative) leadership behavior. The large majority of the participants (90.2% targets; 89.5% observers) provided ratings on the validated Dutch version of the CLS (Redeker et al., 2014), whereas a smaller percentage provided ratings in French (8.8% targets; 9.1% observers) or in

English (1% targets; 1.4% observers). Redeker et al. (2014) report that the original Dutch items "were translated in both English and French by professional linguists" (p. 438). Contact with one of the authors further clarified that two professional linguists translated the CLS using translation-back translation procedures (Brislin, 1970), until quasi complete agreement was reached. Table B in the Appendix shows the results of CFA models testing measurement invariance across languages (Dutch, French, and English), showing full metric and partial scalar measurement invariance across languages for each of the CLS leadership styles. Two dimensions span the circular ordering of the circumplex: 'communion' (affiliation; the horizontal axis), and 'agency' (control/dominance; the vertical axis). Eight leadership styles were assessed, each representing a different octant in this model: Coaching (e.g., "makes positive comments"), inspirational ("has strong character"), directive ("puts employees in their place"), authoritarian ("is bossy"), distrustful ("is suspicious"), withdrawn ("is isolated"), yielding ("wants to please everybody"), and participative leadership ("accepts other approaches"). Items were rated on a Likert scale ranging from 0 (*never*) to 4 (*always*). All descriptive statistics, correlations, and internal consistencies of the variables in Sample 2 are reported in Table 3. As can be seen in Table 3, both the self-ratings of the CLS scales as well as the observer ratings showed a high level of internal consistency (α 's between .74 and .87 for the self-ratings; and between [.80 -.91], [.80 - .82], and [.75 - .93] for the supervisor-, peer-, and subordinate ratings respectively).

Within each observer group, 116 aggregated item scores were calculated (e.g., CLS_item1 was aggregated across the leader's subordinates, across the leader's peers, and across the leader's supervisors) to use as (model) indicators. The $r_{wg(j)}$ IRA (James et al., 1984) was computed for each item within superior, peer, and subordinate groups. The results in the

Appendix (Table C) indicate that the level of similarity within rater sources is sufficiently high to support aggregation (LeBreton & Senter, 2008).

-----Insert Table 3 about here-----

Statistical Analyses

To address RQ1 (i.e., Does the LARI model fit multisource leadership data well?), we tested the LARI model using an analytical approach similar to the one proposed by McAbee and Connelly (2016). In essence, the statistical counterpart of our LARI model is a bifactor model (Reise, 2012). In bifactor modeling, the factor indicators simultaneously load on general and specific factors (which themselves are uncorrelated). Such a model reflects the idea that variation in the indicators is believed to be caused by these different general and specific sources. In the LARI bifactor model (see Figure 3, Panel A), self-ratings load on the Arena and Identity factor, while other-ratings load on the Arena, the general Reputation and a source-specific Reputation factor. This model adheres to the idea that self-ratings comprise a mixture of perceptions shared by everyone (i.e., Arena), self-perceptions not shared with others (i.e., Identity), and measurement error. Other-ratings, in turn, are a blend of perceptions shared by everyone (i.e., Arena), perceptions shared by all external observers (i.e., general Reputation), perceptions unique to the specific group one belongs to $(e.g., Reputation by peers)^1$, and measurement error. When testing the LARI bifactor model, we also allowed for residual correlations between identical indicators across observer groups (see also McAbee & Connelly, 2016; Olsen & Kenny 2006). These residual correlations represent consensus in aspects of those indicators not captured by the Arena and reputation factors.

¹ Note that when a single rater from a given group is used to represent a source-specific factor—rather than aggregated ratings across raters of that group—it is not possible to separate source-specific variance from rater-idiosyncrasy.

Apart from testing the LARI bifactor model (Panel A), we also compared it to a set of five alternative models, including a LARI bifactor model without a general Reputation factor (Panel B), a LARI bifactor model without source-specific Reputation factors (Panel C), a higherorder factor model (Panel D), a higher-order model including a general Reputation factor (Panel E), and a correlated-factors model (Panel F; see Figure 3 for an overview). In the LARI bifactor model, the general Reputation factor captures commonality in the other-ratings that is not shared with self-ratings. Hence, finding that model fit is not improved in the LARI bifactor model relative to the model without a general Reputation factor (see Panel B) would suggest that there is little communality in the perceptions uniquely shared by the external observers. Finding that model fit is not improved relative to the model without source-specific factors (Panel C), on the other hand, would suggest that the variance that is shared among raters from the same source is negligible. As such, support for this model would question the existence of source effects in multisource leadership ratings. The higher-order factor model (see Panel D) represents a model in which source-specific factors define a higher-order factor. Hence, finding comparable fit for the higher-order factor model than for the LARI bifactor model would suggest that the sourcespecific factors (i.e., the Reputation and Identity factors in the LARI model) and the higher-order factor (i.e., the Arena factor in the LARI model) are isomorphic (McAbee & Connelly, 2016). Compared to a higher-order factor model, a higher-order model including a general Reputation factor (see Panel E), in which the three external rater sources load on a general Reputation factor, is closer to the proposed LARI model, as it also separates Identity (self) from Reputation (all external observers). Finally, the correlated-factors model (see Panel F) tests the existence of source-specific factors without assuming a specific higher-order factorial structure in the sourcespecific factors. Hence, finding that model fit is not improved in the LARI bifactor model

relative to the correlated-factors model would suggest the existence of source-specific perceptions without requiring the specific LARI patterns of loadings.

-----Insert Figure 3 about here-----

To address RQ2 (i.e., What is the relative size of the different LARI factors?), the LARI bifactor model was scrutinized in terms of explained variance by each of the LARI factors. To this end, an extension of the Explained Common Variance (ECV) statistic (see Reise, 2012) was used. Specifically, we calculated the proportion of variance explained by each of the LARI factors as the sum of the squared factor loadings for the respective factor divided by the sum of the factor loadings across all LARI factors² (see also McAbee & Connelly, 2016). Note that this approach disregards residual variances (and correlated residuals), which means that ECV captures the percentage of the explained variance (and not the percentage of the total variance) that is accounted for by the different LARI factors.

Finally, to address RQ3 (i.e., Is the relative size of the different LARI factors dependent on the leadership dimension?), these variance decompositions among the LARI factors will be compared between leadership dimensions. In the interest of transparency, these comparisons will be described as exploratory.

All models were tested in Mplus 8.4 (Muthén & Muthén, 1998-2017) using Bayesian estimation and relying on the default Mplus priors³. To evaluate model fit, we evaluated the Comparative Fit Index (CFI) and the Tucker–Lewis Index (TLI), for which values \geq . 90 suggest

² Because in the bifactor model, for identification purposes the factor variances are fixed to one rather than fixing the loading of a marker item, the factor loadings in the LARI model are standardized with respect to the latent variables. ³ Working within a Bayesian framework, we did not report degrees of freedom (note that they are not part of the Mplus output when working with Bayesian estimation). However, the degrees of freedom can easily be calculated from the information that is reported. For the LARI model for forceful behavior (see Table 4), for instance, we have 75 estimated (or free) parameters (i.e., 33 loadings, 18 residual covariances between like items, 12 residual variances, and 12 item intercepts) and we have 90 data points (i.e., 78 variances and covariances + 12 item means/intercepts). This means that this model has 15 degrees of freedom.

an adequate model fit. Moreover, we use the Root Mean Square of Error of Approximation (RMSEA), with values of \leq .10 pointing to an acceptable fit, \leq .08 to an approximate fit, and \leq .05 to a good model fit (Chen et al., 2008). For model comparison purposes, we used the Deviance Information Criterion (DIC), which is the Bayesian alternative of the Akaike Information Criterion (AIC). Similar to AIC, DIC is as a measure that balances model fit and model complexity, which is why it allows comparing non-nested models. Lower values on DIC refer to a better fit-complexity ratio and thus (relatively) better models (Spiegelhalter et al., 2014). Finally, although the same analytical models were tested on the data of both study samples, one important difference between the two measures is that the LVI (Sample 1) is a hierarchically structured leadership instrument in which the four leadership dimensions (e.g., enabling) each have three underlying leadership facets (e.g., empowerment, participation, support), whereas the CLS (Sample 2) is not hierarchically structured. Therefore, the one difference between the analytical models that were tested on the two samples is that the four leader behaviors in Sample 1 were modeled by their respective *subscales*, whereas the eight leadership styles in Sample 2 were modeled by their respective *items*. In other words, in Sample 1, subscale scores were used as factor indicators, whereas the factor indicators were item scores in Sample 2.

Results

Sample 1

Testing the LARI model. Table 4 summarizes the model fit of all tested models⁴. First, the LARI model without source-specific Reputation factors does not fit the data well (e.g., CFI = .783, .858, .713, and .819; RMSEA = .191, .123, .193 and .107 for forceful, enabling, strategic,

⁴ The detailed results for all tested models, in both Sample 1 and Sample 2, can be found on OSF via this link.

and operational respectively). Although the other alternative models fit the data well, they are all outperformed by the LARI model for each of the four leader behaviors. This can be seen from the fact that the LARI model has the best absolute fit values (e.g., CFI = .993, 1, 1, and .990; RMSEA = .047, .00, .00, .023 for forceful, enabling, strategic, and operational respectively), but more importantly that it also has the lowest DIC value (DIC = 8927.91, 7101.20, 5937.18, and 6089.06 for forceful, enabling, strategic, and operational respectively). Hence, it appears that, among the tested models, the LARI model represents the data best.

-----Insert Table 4 about here-----

Variance explained by LARI factors. Variance decomposition of the LARI factors is presented for each of the four leader behaviors (see Figure 4). Across these behaviors, even for the smallest percentage of explained variance, the 95% credibility interval did not include zero (i.e., 5.1%; 95% CI [0.4 - 9.1] for Reputation by peers in forceful behavior). As systematic variation is captured by each LARI factor, each LARI factor turns out to be relevant.

For each of the four LVI leader behaviors, the leadership Arena explained the largest proportion of the variance in leader behavior, ranging between 35% (for strategic) and 52% (for enabling). Therefore, the general consensus about the leader, or the information on the targets' leadership behavior that is shared between the self and *all* others, represents the largest source of information. Although the Arena consistently explains the largest proportion of the variance, it is noticeably larger for the LVI's interpersonal leadership dimensions (forceful (46%); enabling (52%)) compared to the organizational dimensions (strategic (35%); operational (41%)). Except for forceful behavior, the second-largest share originates from the leader Identity, which are the unique self-perceptions of the targets. Here, the Identity seems to be larger for the organizational leadership dimensions (31%)– compared to the

interpersonal dimensions (forceful (13%); enabling (20%)). Next to the relevance of the general Reputation factor, which represents the shared perceptions across all observer rater groups, Figure 4 also shows the relevance of the three source-specific Reputation factors that include perceptions that are uniquely shared by raters from the same source. Among these three observer groups, superiors consistently provide the largest amount of (unique) information across the four leader behaviors (i.e., 8 to 12%), whereas peers provide the smallest amount of (unique) information (i.e., 5 to 8%).

When adding up the explained variance of the Arena and the general Reputation factor, an average percentage of 53.7% is obtained across the four leader behaviors (%= 63.2, 58.2, 45.3, and 48.1 for forceful, enabling, strategic, and operational respectively). The remaining (average of) 46.3% is represented by unique views of the four rater sources. As such, the shared views on leader behavior (i.e., shared among all raters, and shared among observers) reveal about half of the story, while the other half is captured by the unique source-specific perspectives.

Sample 2

Testing the LARI model. Table 5 summarizes the model fit of all tested models in Sample 2. Across leadership styles, the LARI model outperformed the other models, with superior absolute fit indices (CFI, TLI, RMSEA) as well as the lowest DIC values. Next in rank is the LARI model without general Reputation factor, followed by the two hierarchical models and the correlated factors model. Consistent with Sample 1, the LARI model without sourcespecific Reputation factors fitted the data worst. For most of the leadership styles (5/8), the LARI model provided a good fit to the data, while the other models did not (i.e., for inspirational, authoritarian, distrustful, withdrawn, and participative). For example, CFI for inspirational leadership was .904 for the LARI model, .878 for the LARI model without Reputation, .832 for both the higher-order models and the correlated-factors model, and .530 for the LARI model without source-specific Reputation factors. For coaching leadership (1/8), the alternative models (except for the one without source-specific factors) also fit the data well, but the LARI model outperformed the other models with higher absolute fit values and a lower DIC. For the remaining two leadership styles, directive and yielding, none of the models fit the data well. However, the LARI model did relatively better than the other models (e.g., CFI = .842 and .739 for the LARI and the higher-order model of directive leadership, respectively).

-----Insert Table 5 about here-----

Variance explained by LARI factors. Variance decomposition of the LARI factors is presented for each of the eight leadership styles (see Figure 5). Consistent with our findings in the first sample, even for the smallest percentage of explained variance, the 95% credibility interval did not include zero (i.e., 1.5%; 95% CI [1.0 - 4.3] for Reputation by subordinates in withdrawn behavior), supporting the idea that all LARI factors capture systematic variance.

Compared to Sample 1, the leadership Arena's share of explained variance is smaller (between 12 and 48%), whereas the general Reputation's share is relatively larger (between 5 and 29%). So a larger share of "blind spots" –on which all observers agree– is captured in the general Reputation factors of the leadership styles, compared to the four central leader behaviors in Sample 1. Moreover, more *variability* was captured by the general Reputation factor, and the Reputation factors by peers and subordinates in specific. In case Reputation by peers was small (e.g., distrustful and participative), Reputation by subordinates was relatively larger (and the other way around). Yet again, the shared perspectives on leadership (i.e., shared among all raters and among observers) reveal about half of the story, while the other half is captured by unique

source-specific perspectives. Specifically, an average percentage of 44.09% is obtained across the eight dimensions by adding up the explained variance of the Arena and the general Reputation factor (% = 35.8 for coaching, 35.6 for inspirational, 51 for directive, 53.7 for authoritarian, 39.2 for distrustful, 38.5 for withdrawn, 54.2 for yielding, and 44.7 for participative). The remaining (average of) 55.91% is represented by unique perspectives of the four rater sources. When evaluating the findings per leadership style, particularly the large Arena (48%) for the authoritarian leadership style flags up.

The unique self-perceptions of the targets (leader Identity) account for 14 to 21% of the explained variance in leadership behavior. Compared to Sample 1, in which these percentages ranged between 13 and 31%, there seems to be less variability with regard to the explained variance of this Identity factor in Sample 2. Among the three observer groups, superiors consistently provide the largest share of (unique) information across the eight leadership styles (i.e., 14 to 27%); a finding that is consistent with the findings in Sample 1. The smallest share of (unique) information originates from peers in 2/8 styles (distrustful and participative) and from subordinates for the remaining six styles.

-----Insert Figure 5 about here-----

Discussion

Multisource leadership assessments continue to be a popular method for obtaining a wellrounded analysis of leaders' strengths and weaknesses (e.g., Day et al., 2014; Slater & Coyle, 2014). In spite of the obvious advantages that multisource perspectives have to offer in terms of describing and understanding a focal leader, important questions remain as to the existence and the magnitude of the various source effects in traditional 360s (i.e., self, superiors, peers, subordinates). Although source effects, in general, have shown to be alive and well (e.g., Hoffman et al., 2010; Jackson et al., 2020; Lance et al., 2008), it has proven to be highly challenging to disentangle the shared and unique perspectives in multisource data. By investigating each rater source separately, one neglects the possibility that the different rater sources' perceptions show overlap (e.g., Conway & Huffcutt, 1997). Alternatively, by combining all ratings into a single score, one neglects the possibility that the different sources diverge in their perceptions to an important extent (e.g., Jackson et al. 2020). The consequence is that, at present, it remains unknown to what extent each rater source offers truly unique information about a leader (cf. the 'discrepancy hypothesis' in 360-research, e.g., Borman, 1997).

Against this background, we argued that exploiting the full potential of multisource leadership ratings requires a shift in theory and methodology. We proposed the LARI model as a framework that can be used to conceptualize the shared and unique perspectives in multisource ratings. Methodologically, we formalized the LARI model using bifactor modeling. The viability of the LARI model was tested against five alternative models in two samples, each using different leadership instruments. In general, the LARI models showed a superior fit to the 360 data relative to the alternative models, even when taking into account model parsimony. Further, variance decompositions of the LARI factors showed that systematic variation is captured by each of the LARI factors. Especially the non-trivial variance explained by the Reputation factors further supports the viability of the LARI model. In addition, our findings suggest that the shared perspectives on leadership (i.e., Arena + general Reputation) reveal about half of the story, while the other half is captured by unique source-specific perspectives.

Finally, we found tentative evidence for the idea that the relative size of the different LARI factors depends on the specific leadership dimension being rated. Consistent with the SOKA model (Vazire, 2010), the observability of the leadership dimension in question might account for the size of the unique self-perceptions of the leader (Identity) relative to the size of the shared perceptions on leadership (i.e., Arena + general Reputation). In particular, whereas the LVI (cf. Sample 1) tapped into leader *behaviors*, the CLS (cf. Sample 2) covered (interpersonal) leadership *styles*. Assuming that concrete behaviors—compared to styles—have a higher likelihood of being observed (by others), the finding that the explained variance of the shared perspectives on leadership (Arena + general Reputation) was on average 53.7% in Sample 1 and 44.1% in Sample 2, is consistent with the expectations of the SOKA model. Yet, from an overarching perspective, strategic and operational leader behaviors (rated in Sample 1) were the only non-interpersonal leadership dimensions in this study. Especially the large Identity factor for the strategic behavior dimension might be explained by the fact that this type of behavior may include cognitions, goals and aspirations that are not necessarily shared nor visible to others (e.g., "Thinks strategically" or "Anticipates the need to change direction--looks ahead"). In contrast, the Arena is particularly large for highly observable leadership dimensions (e.g., enabling (Sample 1); authoritarian (Sample 2)).

Although a few exceptions could be observed (such as the relatively large % of subordinates' unique perceptions for distrustful and participative leadership in Sample 2), the pattern of results generally does not support the idea of source × dimension interactions (cf. Guion, 1965; Hiller et al., 2011), such that subordinates would be particularly knowledgeable of interpersonal dimensions, whereas superiors and peers would be more knowledgeable of strategic or business-related leadership dimensions. Overall, and consistent with Jackson et al.'s (2020) findings, no meaningful differences were observed in this regard. For instance, subordinates' unique perceptions on interpersonal dimensions like forceful, enabling, coaching, inspirational, and authoritarian leadership were generally smaller –and not larger– compared to

the unique perceptions of superiors. More notably, among the three observer groups, superiors consistently provided the largest amount of (unique) information across leadership dimensions and study samples.

Implications

Our findings have important implications pertaining to leadership (i) theory, (ii) research, and (iii) practice. As a first key implication, our study results support the viability of the conceptual framework we introduced. Inspired by recent developments in personality psychology (McAbee & Connelly, 2016), we conceptualized the shared and unique perspectives in multisource leadership ratings into the Leadership Arena-Reputation-Identity model. While the Arena factor is complementary to existing concepts in the leadership domain (e.g., self-other agreement), the remaining LARI factors, tapping into the unique self- and observer-perceptions, shed new light on multisource leadership ratings.

Apart from the fact that the LARI model outperformed all of the alternative models, it is also noteworthy that the alternative model without source-specific factors (Panel C in Figure 3) fitted the data worst. Combining these findings lends credence to the existence of important source effects in the data (cf. Jackson et al., 2020). Further supporting this idea, the LARI model and the associated variance decompositions showed that each of the LARI factors captured *systematic* variation. As a set, these findings support the core assumption of 360s that each rater source (i.e., self, superiors, peers, subordinates) provides unique information with regard to the target's behavior (cf. the discrepancy hypothesis). These findings align with the ecological perspective on multisource ratings (Lance et al., 2008). In contrast to traditional psychometric theory that treats source effects as systematic bias, the ecological perspective emphasizes the

essential accuracy of perception-based knowledge, recognizing different perceptions as distinct views on a focal leader that may be equally valid (Lance et al., 2008; Landy & Farr, 1980).

As a third theoretical implication, disentangling the various rater perspectives sheds light on what constitutes "leadership". On the one hand, our results are consistent with the idea that leadership is indeed an attribution based on the leader's reputation, and therefore lies "in the eye of the beholder" (cf. Howell & Shamir, 2005; Lord et al., 2020; Nye, 2002). On the other hand, our results are also congruent with the idea that leadership is characteristic to the leader themselves. This is reflected in the Arena of the LARI model and accords to what classic trait psychologist would refer to as "true trait variance", or judgements which everyone agrees upon (Funder, 1995). Thus, our findings suggest that leadership is not "merely an attribution" that people make about an individual (e.g., Haslam et al., 2001; Meindl, 1995). Instead, both the trait theory of leadership (e.g., Kirkpatrick & Locke, 1991) and the attribution theory of leadership (e.g., Green & Mitchell, 1979; Martinko et al., 2007) align well with these results.

For research purposes, studies on self-other agreement in leadership could use the LARI model as a new way to examine the exact composition of shared and unique variation between rater groups. There has been a long tradition of focusing on the level of agreement between the leader and others (cf. Fleenor et al., 2010; Lee & Carpenter, 2018). Although different data-analytic techniques in self-other agreement research (e.g., categories of agreement, polynomial regression, within and between analysis; WABA) reveal different insights, none of the existing techniques allows to present a complete picture of what is *shared* versus *unique* across the multisource perspectives. The LARI (bifactor) model could therefore deliver new insights regarding self-other agreement in leadership, that may accompany the widely accepted methods (see Fleenor et al., 2010 for an overview). Researchers might consider the use of the Arena

factor, including the perspectives on which everyone agrees, as an indicator of self-other agreement that could be related to external variables of interest (e.g., leaders' performance, subordinates' job satisfaction). In addition, the remaining LARI factors offer insights on the shared perspectives among all observers, unique from the leader's self-report (general Reputation), and whether (truly) unique information is captured in the perspectives of the four traditional rater sources.

Further, our results contribute to the ongoing debate on whether self-ratings should be compared to multiple rater groups seperatly or to all rater groups combined. Both correlationbased (e.g., Pearson correlations, intraclass correlations (ICCs)), and non-correlation-based (e.g., $r_{wg(i)}$ interrater agreement coefficient) techniques have been used to estimate levels of rater agreement within and between rater groups (Fleenor et al., 2010). In this context, LeBreton et al. (2003) argued that multisource ratings are generally restricted in range (e.g., due to central tendency and leniency bias), which in turn attenuates correlations within and between rater groups. They further argued that, because of these range restrictions, past studies using correlation-based statistics erroneously concluded that rating similarity between rater groups is low. By applying correlations, ICCs, and r_{wg} to multisource ratings, LeBreton et al. (2003) found support for a 'restriction of variance hypothesis' as the correlations and ICC's were relatively small, whereas the r_{wg} revealed high levels of agreement both within and between rater groups. In light of these findings, the meaningfulness of the tradition in self-other agreement research of separately comparing self-ratings to the ratings of superiors, peers, and subordinates has been questioned (Fleenor et al., 2010). Drawing on our findings, however, we recommend scholars not to aggregate observer ratings of different rater groups, given that next to the perspective they all share (i.e., captured in the Arena and the general Reputation factor), they each have valuable

32

unique views on leadership. Therefore, an "overall" observer score obscures perspectives that are not shared, i.e., perspectives that might as well be opposites. This recommendation is also in line with recent insights on the magnitude of source-related effects in MSPRs (Jackson et al., 2020).

Finally, from a practical point of view, knowing how much of the explained variance is due to self- and the various observer groups might inform us about their usefulness in multisource leadership ratings. As leader behavior is mostly directed towards subordinates, the most common source of observer-rated leadership is subordinate-ratings (Hiller et al., 2011). However, our results reveal that subordinates are usually not the rater group providing the largest share of unique information. In both our samples, superiors were. Although information availability might be lower for superiors (Funder, 2012; Pollack & Pollack, 1996), our findings support the notion that, perceived from their role, superiors evaluate a leader's functioning in a unique (and thus different) way compared to peers and subordinates (cf. Atwater & Yammarino, 1997). Interestingly, this was not only the case for leadership dimensions for which superiors might be especially motivated to keep track (e.g., strategic behaviors), as would be predicted from social information processing theory (Salancik & Pfeffer, 1978). Instead, this particular finding transcended specific leadership dimensions or styles in both study samples. One potential explanation refers to superior's age and experience in evaluating others, and their vertical access to effectively evaluate the focal leader (Harris & Kuhnert, 2008). Organizations can approach the use of our procedures (cf. Analysis code on OSF) in terms of a trade-off between gathering as much information as possible and keeping the implementation cost (and administrative burden) as low as possible. For instance, knowing that superiors' unique input is consistently larger compared to peers and subordinates, and that the latter rater groups provide more overlapping

information regarding leadership behavior, organizations may want to reconsider the use of peer evaluations in their leadership assessments when cost-saving actions are in place.

Limitations and Future Avenues

Although the LARI model reveals a wealth of information, we did not examine relationships between the LARI factors and external variables of interest. In other words, we did not yet test the unique predictive validity of the LARI factors. Therefore, an interesting future avenue could be to test whether the unique and shared perspectives as offered by multisource data provide unique predictive power. For instance, from an attributional perspective on leadership, one can expect that the general Reputation factor adds significantly to the prediction of promotion decisions, above and beyond the Arena. In this regard, the LARI model could be particularly useful to further advance attributional theories on leadership. Socioanalytic theory (Hogan, 1996), for instance, generally uses people's reputation to predict important life outcomes (e.g., performance, career success), because the best predictor of future behavior is past behavior, and because reputations reflect a person's past behavior (Hogan & Blickle, 2018). The LARI model could be used to test this core assumption of socioanalytic theory (Hogan, 1996). One could further delve into this matter by testing whether unique perceptions (e.g., from superiors) are particularly relevant in this regard. Further, a substantial body of research has shown that specific leader behaviors relate to leader effectiveness (e.g., Judge et al., 2004; Vergauwe et al., 2017). Consistent with research on self-other agreement (Fleenor et al., 2010), a positive relationship could be expected between the Arena and leader effectiveness. In contrast, the leader Identity might negatively predict effectiveness, as self-enhancing tendencies (that are part of the Identity) have been linked to lower performance (Connelly & Hülsheger, 2012). Note that objective effectiveness criteria are probably preferred in this context, given that subjective

indicators like multisource ratings of leader effectiveness can also be decomposed in shared and unique rater perspectives on the criterion side.

Although we considered cross-validating the LARI model using different samples and leadership instruments to be critical to this study, allowing us to strengthen our conclusions regarding the viability of the presented LARI framework, it remains an open question whether some of the inconsistencies between the results were due to these differences. The general Reputation factor was, for instance, generally smaller in Sample 1 compared to Sample 2. It is possible that the level of blind spots was lower in Sample 1 because (more overt) leader behaviors were measured compared to (more abstract) leadership styles in Sample 2, or because of the TLTM rating format, which allows raters –and thus the leaders themselves – to reflect on their behaviors on a broader (and maybe deeper) level (cf. Vergauwe et al., 2017). Future research can further explore the conditions that influence the relative size of the different LARI factors.

Finally, the LARI factors can be expected to be dynamic, in the sense that leader identities as well as the perceptions others have about a leaders can change over time (Lord et al., 2020). Indeed, both leader identities as well as observers' perceptions may develop through ongoing interactions at work (e.g., leader-follower interactions) and changing contexts (e.g., financial crises). Understanding the complex ways in which leaders' self-definitions and observers' perceptions develop, change, and are influenced by interactions and contexts, could provide unique insights on the drivers of leader behaviors and actions (Epitropaki et al., 2017). In this context, examining the LARI model using longitudinal multisource leadership data could provide a wealth of information.

Conclusion

In the present paper, we proposed the LARI model, a model that allows disentangling of the shared and unique perspectives in multisource leadership ratings. Cross-validating empirical tests of the LARI model across two large samples and across leadership instruments supports this novel framework's viability. Our study results are promising, and we hope that they enthuse other scholars to explore the further possibilities of the LARI framework.

References

- Allen, T. D., Barnard, S., Rush, M. C., & Russell, J. E. A. (2000). Ratings of organizational citizenship behavior: Does the source make a difference? *Human Resource Management Review*, 10, 97–114. doi: 10.1016/S1053-48229900041-8.
- Andersen, S. M. (1984). Self-knowledge and social inference: II. The diagnosticity of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology*, 46, 294–307. doi: 10.1037/0022-3514.46.2.294
- Aramovich, N. P., & Blankenship, J. R. (2020). The relative importance of participative versus decisive behavior in predicting stakeholders' perceptions of leader effectiveness. *The Leadership Quarterly*, 31(5), 101387. doi: 10.1016/j.leaqua.2020.101387
- Atwater, L. E., & Brett, J. F. (2006). 360-Degree feedback to leaders: Does it relate to changes in employee attitudes? *Group & Organization Management*, 31(5), 578–600. doi: 1059601106286887
- Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self–other agreement: Does it really matter? *Personnel Psychology*, *51*(3), 577–598. doi:10.1111/j.1744-6570.1998.tb00252.x
- Atwater, L. E., Wang, M., Smither, J. W., & Fleenor, J. W. (2009). Are cultural characteristics associated with the relationship between self and others' ratings of leadership? *Journal of Applied Psychology*, 94(4), 876–886. doi:10.1037/a0014561
- Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions?
 Personnel Psychology, 45, 141–164. doi:10.1111/j.1744-6570.1992.tb00848.x
- Atwater, L. E., & Yammarino, F. J. (1997). Self-other rating agreement: A review and model. In G. R. Ferris (Ed.), *Research in personnel and human resources management*, 15. (pp. 121–174). JAI Press.
- Blumer, H. (1986). *Symbolic interactionism: Perspective and method*. Berkeley, CA: University of California Press.
- Borman, W. C. (1997). 360 degree ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, 7(3), 299-316. doi: 10.1016/S1053-4822(97)90010-3

- Braddy, P. W., Gooty, J., Fleenor, J. W., & Yammarino, F. J. (2014). Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self–other agreement. *The Leadership Quarterly*, 25, 373–390. doi: 10.1016/j.leaqua.2013.10.001.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105. doi: 10.1037/h0046016
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research, 36*, 462–494. doi: 10.1177/0049124108314720
- Connelly, B. S., & Hülsheger, U. R. (2012). A narrower scope or a clearer lens for personality? Examining sources of observers' advantages over self-reports for predicting performance. *Journal of Personality*, 80, 603–631.
- Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122. doi: 10.1037/a0021212
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The Convergent Validity between Self and Observer Ratings of Personality: A meta-analytic review. *International Journal* of Selection and Assessment, 15(1), 110–117. doi:10.1111/j.1468-2389.2007.00371.x
- 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(4), 331-360. doi: 10.1207/s15327043hup1004_2
- Conway, J. M., Lombardo, K., & Sanders, K. C. (2001). A meta-analysis of incremental validity and nomological networks for subordinate and peer rating. *Human Performance*, 14(4), 267–303. doi: 10.1207/S15327043HUP1404_1
- Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., & McKee, R. A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly*, 25, 63-82. doi: 10.1016/j.leaqua.2013.11.004

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance

appraisal process: A model and research propositions. *Organizational Behavior and Human Performance*, *33*(3), 360–396. doi: 10.1016/0030-5073(84)90029-1

- Epitropaki, O., Kark, R., Mainemelis, C., & Lord, R. G. (2017). Leadership, followership and identity processes: a multilevel review. *The Leadership Quarterly*, *28*, 104–29. doi: 10.1016/j.leaqua.2016.10.003
- Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self-other rating agreement in leadership: A review. *The Leadership Quarterly*, 21, 1005-1034. doi:10.1016/j.leaqua.2010.10.006
- Foti, R. J., Fraser, S. L., & Lord, R. G. (1982). Effects of leadership labels and prototypes on perceptions of political leaders. *Journal of Applied Psychology*, 67(3), 326–333. doi: 10.1037/0021-9010.67.3.326
- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102(4), 652. doi: 10.1037/0033-295X.102.4.652
- Funder, D. C. (2012). Accurate personality judgment. Current Directions in Psychological Science, 21, 177–182. doi: 10.1177/0963721412445309
- Guion, R. (1965). Personnel testing. New York, NY: McGraw-Hill.
- Green, S.G., Mitchell, T.R. (1979). Attributional processes of leaders in leader–member interactions. Organizational Behavior and Human Performance, 23, 429-458. doi: 10.1016/0030-5073(79)90008-4
- Hansbrough, T. K., Lord, R. G., & 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
- Harris, L. S., & Kuhnert, K. W. (2008). Looking through the lens of leadership: A constructive developmental approach. *Leadership and Organization Development Journal*, 29, 47–67. doi: 10.1108/01437730810845298
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43–62. doi: 10.1111/j.1744-6570.1988.tb00631.x
- Haslam, S. A., Platow, M. J., Turner, J. C., Reynolds, K. J., McGarty, C., Oakes, P. J., Johnson,S., Ryan, M. K., & Veenstra, K. (2001). Social identity and the romance of leadership:

The importance of being seen to be 'doing it for us'. *Group Processes & Intergroup Relations*, 4(3), 191-205. doi: 10.1177/1368430201004003002

- Hiller, N. J., DeChurch, L. A., Murase, T., & Doty, D. (2011). Searching for outcomes of leadership: A 25-year review. *Journal of Management*, 37, 1137–1177. doi: org/10.1177/0149206310393520.
- Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*. 63(1), 119-151. doi: 10.1111/j.1744-6570.2009.01164.x
- Hogan, R. (1996). A socioanalytic interpretation of the Five-Factor Model. In J. S. Wiggins (Ed.), The Five-Factor model of personality (pp.163–179). New York, NY: Guilford Press.
- Hogan, R., & Blickle, G. (2018). Socioanalytic theory: Basic concepts, supporting evidence and practical implications. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: The science of personality and individual differences: The science of personality and individual differences* (pp. 110–129). Sage Reference. doi: 10.4135/9781526451163.n5
- Hooijberg, R., & Choi, J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. *The Leadership Quarterly*, *11*, 341–364. doi: 10.1016/S1048-9843(00)00044-8
- Howell, J. M., & Shamir, B. (2005). The role of followers in the charismatic leadership process:
 Relationships and their consequences. *The Academy of Management Review*, *30*, 96–112.
 doi: 10.5465/AMR.2005.15281435
- Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, 105(3), 312–329. doi: 10.1037/apl0000434
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85–98. doi: 10.1037/0021-9010.69.1.85
- Judge, Piccolo, & Ilies, (2004). The forgotten ones? The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology*, 89, 36-51. doi: 10.1037/0021-9010.89.1.36

- Kaiser, R. B., LeBreton, J. M., & Hogan, J. (2015). The dark side of personality and extreme leader behavior. *Applied Psychology: In International Review*, 64, 55-92. doi: 10.1111/apps.12024
- Kaiser, R. B., Overfield, D. V., & Kaplan, R. E. (2010). Leadership Versatility Index version 3.0 facilitator's guide. Greensboro, NC: Kaplan DeVries Inc.
- Kirkpatrick, S. & Locke, E. (1991). Leadership: Do traits matter? *Academy of Management Perspectives*, 5(2), 48-60. doi: 10.5465/ame.1991.4274679
- Lance, C. E., Hoffman, B. J., Gentry, W. A., Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review*, 18, 223–232. doi: 10.1016/j.hrmr.2008.03.002
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72–107. doi: 10.1037/0033-2909.87.1.72
- Leary, M. R., & Kowalski, R. M. (1990). Impression management: A, literature review and twocomponent model. *Psychological Bulletin, 107*, 34-47. doi: 10.1037/0033-2909.107.1.34
- LeBreton, J. M., Burgess, J. D., Kaiser, R. B., Atchley, E. K., James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6, 80–128. doi:10.1177/1094428102239427
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11, 815–852. doi: 10.1177/1094428106296642
- Lee, A., & Carpenter, N. C. (2018). Seeing eye to eye: A meta-analysis of self-other agreement of leadership. *The Leadership Quarterly*, 29(2), 253–275. doi: 10.1016/j.leaqua.2017.06.002
- Lord, R. G., Epitropaki, O., Foti, R. J., & Hansbrough, T. K. (2020). Implicit leadership theories, implicit follower theories, and dynamic processing of leadership information. *Annual Review of Organizational Psychology and Organizational Behavior*, 7, 49-74. doi: 10.1146/annurev-orgpsych-012119-045434
- Lord, R. G., Foti, R. J., & de Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior & Human Performance, 34*(3), 343–378. doi:10.1016/0030-5073(84)90043-6

- Luft, J., & Ingham, H. (1955). *The Johari window: A graphic model of interpersonal awareness*.Paper presented at the Proceedings of the Western Training Laboratory in GroupDevelopment, LA, CA: UCLA Extension Office.
- Martinko, M. J., Harvey, P., & Douglas, S. C. (2007). The role, function, and contribution of attribution theory to leadership: A review. *The Leadership Quarterly*, 18(6), 561–585. doi: 10.1016/j.leaqua.2007.09.004
- McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The Trait-Reputation-Identity model. *Psychological Review*, 123, 569-591. doi: 10.1037/rev0000035
- Meindl, J. R. (1995). The romance of leadership as a follower-centric theory: A social constructionist approach. *The Leadership Quarterly*, 329-341. doi: 10.1016/1048-9843(95)90012-8
- Mount, M. K. (1984). Supervisor, self and subordinate ratings of performance and satisfaction with supervision. *Journal of Management*, *10*(3), 305-320. doi: 10.1177/014920638401000304
- Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology*, 51(3), 557– 576. doi: 10.1111/j.1744-6570.1998.tb00251.x
- 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, 158-164. doi: 10.1037/0021-9010.89.1.158
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
- Nye, J. L. (2002). The eye of the follower Information processing effects on attributions regarding leaders of small groups. *Small Group Research*, 33(3), 337-360. doi:10.1177/10496402033003003
- Olsen, J. A., & Kenny, D. A. (2006). Structural equation modeling with interchangeable dyads. *Psychological Methods*, 11, 127. doi: 10.1037/1082-989X.11.2.127
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY: Guilford Press.

- Pollack, D. M., & Pollack, L. J. (1996). Using 360 degree feedback in performance appraisal. *Public Personnel Management*, 25, 507–528. doi: 10.1177/009102609602500410
- Redeker, M., de Vries, R. E., Rouckhout, D., Vermeren, P., & De Fruyt, F. (2014). Integrating leadership: The leadership circumplex. *European Journal of Work and Organizational Psychology*, 23, 435-455. Doi: 10.1080/1359432X.2012.738671
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. doi: 10.1080/00273171.2012.715555
- Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75(3), 322–327. doi: 10.1037/0021-9010.75.3.322
- Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly*, 23(2), 224–253. doi: 10.2307/2392563
- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956–970. doi: 10.1037/0021-9010.85.6.956
- Slater, R. & Coyle, A. (2014). The governing of the self/the self-governing self: Multirater/source feedback and practices 1940-2011. *Theory & Psychology*, 24(2), 233-255. doi: 10.1177/0959354313520087
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), 76(3), 485–493. doi:10.1111/rssb.12062
- Tornow, W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management*, *32*, 221–230. doi: 10.1002/hrm.3930320203
- Vazire, S. (2010). Who knows what about a person? The Self-Other Knowledge Asymmetry (SOKA) Model. *Journal of Personality and Social Psychology*, 98, 281-300. doi: 10.1037/a0017908
- Vergauwe, J., Wille, B., Hofmans, J., Kaiser, R. B., & De Fruyt, F. (2017). The "too little/too much" scale: A new rating format for detecting curvilinear effects. *Organizational Research Methods*, 20, 518-544. doi:10.1177/1094428117706534
- Vergauwe, J., Wille, B., Hofmans, J., Kaiser, R. B., & De Fruyt, F. (2018). The double-edged sword of leader charisma: Understanding the curvilinear relationship between charismatic

personality and leader effectiveness. *Journal of Personality and Social Psychology*, 114, 110-130. doi: 10.1037/pspp0000147

- Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, 90(1), 108–131. doi:10.1037/0021-9010.90.1.108
- Woehr, D. J., Sheehan, M. K., Bennett, W. (2005). Assessing measurement equivalence across ratings sources: A multitrait-multirater approach. *Journal of Applied Psychology*, 90, 592–600. doi: 10.1037/0021-9010.90.3.592
- Yammarino, F. J. (2003). Modern data analytic techniques for multisource feedback. *Organizational Research Methods*, *6*, 6–14. doi: 10.1177/1094428102239423

Yukl, G. A. (2010). Leadership in organizations. Upper Saddle River, NJ: Prentice Hall.

Table 1

Conceptual Meaning of the Six Factors of the Leadership Arena-Reputation-Identity (LARI) Model and Elements Contributing to the LARI Factors

LARI factor	Conceptual meaning	Elements contributing to the factor
Leadership Arena	Information about the leader's behavior that everyone agrees upon, including the leader themselves, supervisors, peers, and subordinates	 Overt leader behavior congruent with a leader's self-reports. The extent to which the leader conveys his/her desired reputation both to others and in his/her self-description The leader's level of self-awareness Observers' motivation and ability to provide accurate ratings of the leader The extent to which observers are exposed to the leader The extent to which the particular behaviors are observable
General Reputation	Information about the leader's behavior that is shared by all observer groups but is unique from the leader's self-report	 Overt leader behavior that the leader in unaware of ("e.g., "blind spots") Overt leader behavior that is intentionally not shared by the leader, but picked up by all observers (e.g., impression management by the leader) Perceptions of the leader communicated about in the broad social network of subordinates, peers and supervisors Systematic bias shared across observers (e.g., gender or physical appearance stereotypes, leniency bias, halo effect)
Reputation by subordinates	Information about the leader's behavior that is unique in the perspective provided by subordinates (and not shared with the views provided by supervisors, peers, or the leader themselves)	 Overt leader behavior visible only to subordinates (e.g., "managing down"), and not represented in leaders' self-reports Communication about the leader among subordinates Leniency bias among subordinates due to strategic responding
Reputation by peers	Information about the leader's behavior that is unique in the perspective provided by peers (and not shared with the views provided by supervisors, subordinates, or the leader themselves)	 Overt leader behavior visible only to peers, and not represented in leaders' self-reports Communication about the leader among peers Friendship bias
Reputation by supervisors	Information about the leader's behavior that is unique in the perspective provided by supervisors (and not shared with the views provided by subordinates, peers, or the leader themselves)	 Overt leader behavior visible only to supervisors (e.g., "managing up"), and not represented in leaders' self-reports Communication about the leader among supervisors Leniency bias

Leader Identity	Information about the leader's behavior that is unique in the perspective provided by the leader themselves (and not shared with the views provided by the three	•	Accurate leader behaviors or attitudes that are low in visibility and remain private to the self Inaccurate self-perceptions of a leader's behavior (e.g., self-
	observer groups)		enhancement)
		٠	Strategic/intentional self-presentation tactics that don't overlap with
			observers' perspectives

Table 2

Descriptive Statistics and Variable Intercorrelations: Sample 1 (N = 537 target leaders)

	Μ	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.
1.Sex ^a	-	-	-																	
2.Age	45.59	7.97	.00	-																
3.Forceful: self	.09	.61	03	.00	.82															
4.Forceful: superiors	14	.56	.04	06	.41**	.89														
5.Forceful: peers	03	.52	.03	.00	.39**	.61**	.90													
6.Forceful: subordinates	.01	.43	.05	01	.44**	.48**	.56**	.87												
7.Enabling: self	06	.46	01	02	38**	32**	40**	39**	.73											
8.Enabling: superiors	11	.37	.08	.02	33**	62**	51**	40**	.29**	.84										
9.Enabling: peers	22	.39	01	03	35**	48**	68**	47**	.39**	.56**	.87									
10.Enabling: subordinates	25	.33	.00	03	33**	34**	39**	61**	.36**	.46**	.50**	.87								
11.Strategic: self	16	.56	17**	10^{\dagger}	34**	.17**	.13*	.08	$.11^{*}$	14*	10^{+}	06	.84							
12.Strategic: superiors	34	.40	02	13*	$.14^{*}$.39**	.21**	$.18^{**}$	14*	13*	08	02	.30**	.86						
13.Strategic: peers	28	.36	07	14*	.13*	.28**	.33**	$.17^{**}$	08	09^{\dagger}	.03	03	.25**	.41**	.88					
14.Strategic: subordinates	19	.30	03	16**	.19**	.28**	.29**	.39**	11*	07	11†	.05	.27**	.41**	$.48^{**}$.88				
15.Operational: self	14	.45	.04	.02	$.10^{\dagger}$	06	05	.05	.08	.00	03	05	24**	22**	25**	19**	.69			
16.Operational: superiors	06	.30	.08	.02	03	.02	.02	.06	03	.04	05	05	23**	17**	22**	15**	.31**	.72		
17.Operational: peers	09	.28	.04	.06	02	05	.06	.06	02	.02	04	04	22**	18**	29**	17**	.32**	.43**	.75	
18.Operational: subordinates	13	.25	$.10^{\dagger}$.06	.03	05	02	.24**	10^{\dagger}	02	.03	04	29**	17**	20**	20**	.32**	.31**	.37**	.70

Note. Leader behaviors are rated on a scale ranging between -4 (much too little), 0 (the right amount), and +4 (much too much); Means, standard deviations and correlations are based on N = 537 target leaders; Bold values on the diagonal show the internal consistencies (Cronbach alphas) of the relevant variables, which are based on N = 537 for target leaders' self-reports, N = 1142 for superior-reports, N = 2695 for peer-reports, and N = 3500 for subordinate-reports; ^aSex is dummy coded such that 0 = male and 1 = female; [†]p < .05, ^{*}p < .01, ^{**}p < .001.

Table 3

Descriptive Statistics and Variable Intercorrelations: Sample 2 (N = 1255 target leaders)

	Μ	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1.Sex ^a	-	-	-															
2.Age	44.17	8.80	14*	-														
3. Coaching: self	3.11	.35	01	.17**	.86													
4.Coaching: superiors	2.93	.40	$.18^{**}$	06	.21**	.91												
5.Coaching: peers	2.89	.34	$.11^{*}$	02	$.18^{**}$.37**	.92											
6.Coaching: subordinates	2.95	.37	$.07^{\dagger}$	08^{\dagger}	.19**	.30**	.36**	.93										
7.Inspirational: self	2.87	.41	08^{+}	.23**	.64**	$.10^{**}$.03	$.07^{\dagger}$.87									
8.Inspirational: superiors	2.74	.46	.13**	08^{\dagger}	$.07^{*}$.66**	.19**	.16**	.23**	.91								
9.Inspirational: peers	2.80	.36	$.08^{\dagger}$.00	$.08^*$.24**	.65**	.21**	.21**	.44**	.90							
10.Inspirational: subordinates	2.80	.37	.05	03	$.10^{**}$.22**	$.20^{**}$.72**	.26**	.38**	.42**	.90						
11.Directive: self	2.19	.43	14**	.06	.16**	06^{\dagger}	12**	05	.53**	.13**	.11**	$.20^{**}$.74					
12.Directive: superiors	2.10	.45	$.08^{\dagger}$	03	03	$.11^{**}$	07^{\dagger}	02	.25**	.63**	.34**	.34**	.32**	.80				
13.Directive: peers	2.18	.37	.00	03	04	05	.00	07^{\dagger}	.24**	.34**	.59**	.32**	.35**	.56**	.80			
14.Directive: subordinates	2.12	.35	04	.02	07^{\dagger}	07^{\dagger}	13**	01	.25**	.26**	.29**	.53**	.43**	$.50^{**}$.58**	.78		
15.Authoritarian: self	1.39	.43	06	.03	38**	17**	26**	19**	.05	$.10^{**}$	$.07^{\dagger}$	$.10^{**}$	$.48^{**}$.32**	.34**	.41**	.81	
16.Authoritarian: superiors	1.36	.46	01	$.12^{*}$	15**	46**	27**	18**	.12**	.11**	.15**	.14**	.23**	.61**	.47**	.41**	.39**	.87
17.Authoritarian: peers	1.41	.42	06	.09†	12**	26**	56**	25**	.19**	.15**	.05	.15**	.30**	.46**	.66**	$.50^{**}$.42**	.57**
18.Authoritarian: subordinates	1.30	.41	05	$.12^{*}$	15**	21**	31**	59**	.16**	.14**	.11**	05	.31**	.41**	.47**	.66**	.46**	.49**
19.Distrustful: self	.96	.33	.06	09†	61**	16**	21**	19**	34**	.00	02	04	.15**	.15**	$.18^{**}$.21**	.65**	.23**
20.Distrustful: superiors	.98	.39	08^{\dagger}	.15**	09*	72**	30**	20**	.05	37**	08^{*}	05	.14**	.19**	.22**	.19**	.22**	$.70^{**}$
21.Distrustful: peers	1.03	.34	07^{\dagger}	.09†	10*	30**	-78**	29**	$.10^{**}$	04	38**	05	.15**	.25**	.31**	.28**	.29**	.41**
22.Distrustful: subordinates	.94	.35	05	.12**	14**	25**	34**	82**	$.06^{\dagger}$	03	08^{*}	45**	$.18^{**}$	$.20^{**}$.26**	.33**	.30**	.33**
23.Withdrawn: self	1.02	.38	$.08^{\dagger}$	14**	52**	13**	04	12**	74**	24**	21**	29**	42**	23**	25**	26**	$.06^{\dagger}$	08^{*}
24.Withdrawn: superiors	1.07	.45	14**	.13**	03	58**	18**	16**	17**	83**	40**	34**	14**	54**	31**	25**	11**	.01
25.Withdrawn: peers	1.03	.35	09†	.04	06^{\dagger}	24**	61**	22**	15**	40**	86**	39**	11**	28**	48**	26**	06^{\dagger}	08^{*}
26.Withdrawn: subordinates	.98	.33	05	.04	09*	23**	25**	72**	17**	32**	39**	86**	12**	24**	25**	37**	04	05
27.Yielding: self	1.63	.38	.13**	08^{\dagger}	09*	02	.09*	.03	48**	27**	22**	24**	40**	33**	35**	32**	26**	24**
28. Yielding: superiors	1.58	.43	.04	.11*	$.06^{\dagger}$	10**	$.07^{\dagger}$.02	21**	60**	35**	31**	23**	63**	46**	41**	29**	42**
29. Yielding: peers	1.51	.35	$.08^{\dagger}$.02	.04	.01	.01	.04	23**	34**	56**	33**	23**	43**	63**	46**	30**	37**
30. Yielding: subordinates	1.49	.28	.05	02	.06	.01	$.08^{*}$	$.06^{\dagger}$	25**	34**	35**	45**	26**	44**	48**	62**	32**	36**
31.Participative: self	2.72	.32	$.11^{*}$.01	.58**	.14**	.17**	.13**	.16**	08^{*}	09*	09*	20**	22**	25**	27**	52**	25**
32.Participative: superiors	2.59	.35	$.11^{*}$	04	$.18^{**}$.65**	.29**	20^{**}	07*	$.10^{**}$	06^{\dagger}	08*	22**	38**	35**	34**	34**	69**
33.Participative: peers	2.54	.31	.12*	05	.15**	.29**	.74**	.28**	14**	07†	.15**	09*	28**	35**	42**	42**	39**	45**
34.Participative: subordinates	2.65	.30	.09†	11*	$.18^{**}$.24**	.33**	.81**	09*	05	01	.33**	22**	26**	33**	39**	36**	35**

Table 3 (continued)

	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.	33.	34.
17.Authoritarian: peers	.89																	
18.Authoritarian: subordinates	.58**	.90																
19.Distrustful: self	.25**	.26**	.78															
20.Distrustful: superiors	.38**	.31**	.15**	.87														
21.Distrustful: peers	.79**	.42**	.22**	.38**	.90													
22.Distrustful: subordinates	.42**	.81**	.24**	$.28^{**}$	$.40^{**}$.91												
23.Withdrawn: self	15**	12**	.45**	.01	06^{\dagger}	.01	.82											
24.Withdrawn: superiors	12**	09*	01	.49**	$.09^{*}$	$.06^{\dagger}$.23**	.89										
25.Withdrawn: peers	.05	07^{\dagger}	.01	.13**	$.50^{**}$.12**	$.18^{**}$.42**	.89									
26.Withdrawn: subordinates	04	.21**	$.07^{\dagger}$.11**	.15**	.61**	.25**	.34**	.41**	.88								
27.Yielding: self	33**	28**	$.08^*$	10**	20**	14**	.59**	.25**	.17**	.16**	.78							
28. Yielding: superiors	40**	34**	12**	04	20**	18**	.22**	.64**	.32**	.23**	$.40^{**}$.84						
29. Yielding: peers	49**	38**	14**	16**	16**	20**	.22**	.32**	.56**	.26**	.42**	.56**	.83					
30. Yielding: subordinates	44**	50**	14**	15**	21**	21**	.25**	.29**	.30**	.43**	.42**	.51**	$.58^{**}$.75				
31.Participative: self	26**	28**	49**	14**	16**	16**	05	$.10^{**}$	$.08^{*}$	$.06^{\dagger}$.35**	.21**	.22**	.25**	.76			
32.Participative: superiors	46**	38**	21**	64**	35**	28**	$.06^{\dagger}$	07^{\dagger}	.03	.02	.23**	.42**	.32**	.32**	.29**	.84		
33.Participative: peers	77**	49**	24**	34**	71**	37**	.13**	$.07^{\dagger}$	13**	.01	.31**	.34**	.47**	.38**	.30**	.44**	.86	
34.Participative: subordinates	45**	81**	24**	25**	35**	82**	$.06^{\dagger}$.02	01	37**	.22**	.24**	.29**	.43**	.28**	.34**	.45**	.88

Note. Means, standard deviations and correlations are based on N = 1255 target leaders; Bold values on the diagonal show the internal consistencies (Cronbach alphas) of the relevant variables, which are based on N = 1255 for target leaders' self-reports, N = 2175 for superior-reports, N = 5068 for peer-reports, and N = 8534 for subordinate-reports; ^aSex is dummy coded such that 0 = male and 1 = female; [†]p < .05, ^{*}p < .01, ^{**}p < .001.

Table 4

Model Characteristics of the LARI Model compared to Alternative Models in Sample 1 (N = 537 target leaders)

Model	Free	2.5% PP	97.5%	PP p-	CFI	TLI	RMSEA	RM	SEA	DIC
	parameters	limit	PP limit	value				90%	o CI	
Forceful										
LARI model	75	-14.861	56.612	.133	.993	.976	.047	.017	.068	8927.914
LARI model w/o G-Reputation	66	1.377	71.734	.024	.989	.969	.053	.039	.067	8937.837
LARI model w/o S-Reputation	55	662.573	734.413	<.001	.783	.597	.191	.189	.194	9587.641
Higher-order model	58	35.702	103.277	<.001	.978	.956	.063	.056	.072	8962.204
Higher-order with G-Reputation	58	34.586	106.447	<.001	.978	.955	.064	.056	.073	8963.244
Correlated factors model	60	24.024	97.353	<.001	.980	.958	.062	.053	.071	8957.106
Enabling										
LARI model	75	-36.251	37.309	.518	1.000	1.000	.000	.000	.024	7101.204
LARI model w/o G-Reputation	66	-36.495	33.868	.512	1.000	1.000	.000	.000	.041	7150.892
LARI model w/o S-Reputation	55	250.829	318.703	<.001	.858	.736	.123	.119	.127	7426.435
Higher-order model	58	-23.658	45.871	.254	.994	.989	.025	.000	.041	7153.820
Higher-order with G-Reputation	58	-24.401	47.139	.244	.994	.988	.026	.000	.042	7154.433
Correlated factors model	60	-23.756	43.405	.259	.995	.988	.026	.000	.044	7156.796
Strategic										
LARI model	75	-38.185	37.565	.523	1.000	1.000	.000	.000	.049	5937.181
LARI model w/o G-Reputation	66	-19.593	51.519	.191	.993	.981	.036	.000	.054	5946.525
LARI model w/o S-Reputation	55	678.188	747.467	<.001	.713	.470	.193	.191	.195	6631.464
Higher-order model	58	-14.889	54.288	.129	.992	.984	.034	.016	.047	5941.319
Higher-order with G-Reputation	58	-17.832	53.178	.137	.992	.984	.034	.013	.047	5940.315
Correlated factors model	60	-17.347	52.242	.147	.993	.985	.033	.006	.048	5941.439

Table 4 (continued)

Model	Free	2.5% PP	97.5%	PP p-	CFI	TLI	RMSEA	RM	SEA	DIC
	parameters	limit	PP limit	value				90%	6 CI	
Operational										
LARI model	75	-23.199	48.038	.244	.990	.986	.023	.000	.038	6089.060
LARI model w/o G-Reputation	66	-23.422	46.621	.241	.989	.973	.031	.000	.048	6108.971
LARI model w/o S-Reputation	55	191.877	261.980	<.001	.819	.675	.107	.103	.112	6312.123
Higher-order model	58	-21.184	50.432	.210	.989	.978	.028	.000	.043	6102.529
Higher-order with G-Reputation	58	-23.147	47.444	.225	.990	.980	.027	.000	.043	6100.808
Correlated factors model	60	-22.082	47.688	.225	.989	.978	.028	.000	.045	6103.902

Note. w/o G-Reputation = without General Reputation; w/o S-Reputation = without Source-specific Reputation (i.e., Reputation by superiors, peers, and subordinates); PP p-value = Posterior Predictive P-Value; PP limits represent the 95% confidence intervals for the difference between the observed and the replicated Chi-Square values; N = 537 target leaders, N = 1142 superiors, N = 2695 peers, and N = 3500 subordinates.

Table 5

Model Characteristics of the LARI Model compared to Alternative Models in Sample 2 (N = 1255 target leaders)

Model	Free	2.5% PP	97.5% PP	PP p-	CFI	TLI	RMSEA	RM	SEA	DIC
	parameters	limit	limit	value				90%	6 CI	
Coaching										
LARI model	375	2292.883	2554.724	<.001	.939	.929	.036	.036	.036	79539.677
LARI model w/o G-Reputation	330	2972.577	3224.931	<.001	.923	.912	.040	.040	.040	80164.897
LARI model w/o S-Reputation	271	18408.982	18671.134	<.001	.541	.499	.096	.095	.096	95547.546
Higher-order model	274	3905.733	4169.916	<.001	.900	.890	.045	.045	.045	81045.831
Higher-order with G-Reputation	274	3905.486	4161.453	<.001	.900	.890	.045	.045	.045	81043.023
Correlated factors model	276	3902.707	4166.861	<.001	.900	.890	.045	.045	.045	81048.336
Inspirational										
LARI model	375	3671.524	3958.234	<.001	.904	.889	.045	.045	.045	99699.931
LARI model w/o G-Reputation	330	4732.049	4986.164	<.001	.878	.862	.050	.050	.050	100701.987
LARI model w/o S-Reputation	271	18714.598	18973.046	<.001	.530	.486	.096	.096	.096	114631.159
Higher-order model	274	6583.061	6847.375	<.001	.832	.816	.058	.057	.058	102508.039
Higher-order with G-Reputation	274	6585.827	6845.620	<.001	.832	.816	.058	.057	.058	102503.972
Correlated factors model	276	6583.455	6838.245	<.001	.832	.816	.058	.058	.058	102505.655
Directive										
LARI model	300	3720.810	3941.524	<.001	.842	.808	.058	.057	.058	104153.366
LARI model w/o G-Reputation	264	4883.256	5110.729	<.001	.794	.759	.064	.064	.065	105286.041
LARI model w/o S-Reputation	217	9816.332	10029.988	<.001	.593	.544	.089	.089	.089	110161.325
Higher-order model	220	6244.608	6465.375	<.001	.739	.707	.071	.071	.071	106599.955
Higher-order with G-Reputation	220	6246.422	6468.524	<.001	.739	.706	.071	.071	.071	106601.215
Correlated factors model	222	6221.549	6440.460	<.001	.740	.707	.071	.071	.071	106575.084

Table 5 (continued)

Model	Free	2.5% PP	97.5% PP	PP p-	CFI	TLI	RMSEA	RM	SEA	DIC
	parameters	limit	limit	value				90%	6 CI	
Authoritarian										
LARI model	375	2948.975	3209.815	<.001	.920	.907	.040	.040	.041	120625.284
LARI model w/o G-Reputation	330	3713.572	3965.244	<.001	.900	.887	.044	.044	.045	121343.903
LARI model w/o S-Reputation	271	13241.532	13502.219	<.001	.656	.624	.081	.081	.081	130817.487
Higher-order model	274	4086.224	4342.102	<.001	.891	.881	.046	.046	.046	121663.811
Higher-order with G-Reputation	274	4082.003	4339.747	<.001	.891	.881	.046	.045	.046	121659.990
Correlated factors model	276	4078.712	4335.855	<.001	.891	.881	.046	.046	.046	121658.642
Distrustful										
LARI model	375	1973.340	2236.164	<.001	.929	.917	.033	.033	.034	101502.184
LARI model w/o G-Reputation	330	2530.880	2760.219	<.001	.912	.900	.037	.037	.037	101984.550
LARI model w/o S-Reputation	271	13049.381	13310.091	<.001	.561	.521	.081	.081	.081	112479.941
Higher-order model	274	3228.835	3483.319	<.001	.887	.877	.041	.041	.041	102657.065
Higher-order with G-Reputation	274	3225.880	3479.258	<.001	.888	.877	.041	.041	.041	102652.720
Correlated factors model	276	3220.018	3477.773	<.001	.888	.877	.041	.041	.041	102654.718
Withdrawn										
LARI model	375	2242.166	2505.567	<.001	.922	.916	.034	.034	.034	107280.255
LARI model w/o G-Reputation	330	2779.826	3028.565	<.001	.905	.892	.039	.038	.039	107904.987
LARI model w/o S-Reputation	271	13381.326	13641.747	<.001	.561	.520	.082	.081	.082	118451.746
Higher-order model	274	3182.082	3448.127	<.001	.891	.881	.041	.040	.041	108256.610
Higher-order with G-Reputation	274	3182.322	3439.083	<.001	.891	.881	.041	.040	.041	108253.187
Correlated factors model	276	3172.859	3434.033	<.001	.892	.881	.041	.040	.041	108249.158

Table 5 (continued)

Model	Free	2.5% PP	97.5% PP	PP p-	CFI	TLI	RMSEA	RMSEA		DIC
	parameters	limit	limit	value				90%	6 CI	
Yielding										
LARI model	375	4776.711	5029.464	<.001	.823	.793	.051	.051	.051	122120.154
LARI model w/o G-Reputation	330	6184.928	6439.976	<.001	.772	.742	.057	.057	.057	123488.869
LARI model w/o S-Reputation	271	11491.757	11752.704	<.001	.582	.543	.076	.076	.076	128742.209
Higher-order model	274	7504.916	7759.778	<.001	.725	.699	.061	.061	.062	124752.455
Higher-order with G-Reputation	274	7499.794	7762.419	<.001	.725	.700	.061	.061	.062	124750.533
Correlated factors model	276	7502.024	7764.264	<.001	.725	.699	.062	.061	.062	124756.592
Participative										
LARI model	350	1828.411	2064.027	<.001	.923	.909	.035	.034	.035	86712.048
LARI model w/o G-Reputation	308	2491.768	2735.815	<.001	.896	.881	.040	.039	.040	87344.392
LARI model w/o S-Reputation	253	9924.605	10166.639	<.001	.604	.564	.076	.076	.076	94717.464
Higher-order model	256	2870.386	3111.919	<.001	.881	.869	.042	.041	.042	87670.015
Higher-order with G-Reputation	256	2872.318	3114.201	<.001	.881	.869	.042	.041	.042	87674.006
Correlated factors model	258	2868.996	3110.010	<.001	.881	.869	.042	.041	.042	87670.577

Note. w/o G-Reputation = without General Reputation; w/o S-Reputation = without Source-specific Reputation (i.e., Reputation by superiors, peers, and subordinates); PP p-value = Posterior Predictive P-Value; PP limits represent the 95% confidence intervals for the difference between the observed and the replicated Chi-Square values; N = 1255 target leaders, N = 2175 superiors, N = 5068 peers, and N = 8534 subordinates.



Figure 1. The Johari window (Luft & Ingham, 1955) with LARI-labels in parentheses (Left), and the Trait-Reputation-Identity (TRI) Model (Right), adapted from "A multi-rater framework for studying personality: The Trait-Reputation-Identity model" by S. T. McAbee and B. S. Connelly, 2016, Psychological Review, 123(5), p. 570. Copyright 2016 by the American Psychological Association.



Figure 2. The Leadership Arena-Reputation-Identity (LARI) model.

Note. s = self, b = subordinates, p = peers, r = superiors in the item subscripts.





Panel B: The LARI model without general Reputation.

y1_{self}

Figure 3. Panel A: The LARI model.



Panel C: The LARI model without source-specific Reputation factors.



Panel D: Higher-order factor model.





Panel E: Higher-order factor model with general Reputation.

Panel F: Correlated-factors model.



Figure 4. Proportion of the explained variance accounted for by the LARI factors in Sample 1 (Leadership Versatility Index; Kaiser et al., 2010).



Figure 5. Proportion of the explained variance accounted for by the LARI factors in Sample 2 (Circumplex Leadership Scan; Redeker et al., 2014).

Appendix

Table A

Inter-rater Agreement $(r_{wg(j)})$ for the Leader Behavior Subscales (Leadership Versatility Index) in

Sample 1

	Subordinates	Peers	Superiors
Forceful			
Takes charge	.84	.83	.88
Declares themselves	.86	.83	.89
Pushes performance	.84	.82	.89
Enabling			
Empowerment	.86	.85	.90
Participation	.88	.87	.92
Support	.87	.89	.93
Strategic			
Direction	.88	.87	.90
Growth	.86	.85	.90
Innovation	.92	.90	.94
Operational			
Execution	.87	.87	.93
Efficiency	.91	.90	.93
Order	.91	.90	.94
Mean	.88	.87	.91
SD	.03	.03	.02
Agreement	strong	strong	very strong
N for $k = min. 2$	530	523	297

Note. To account for central tendency bias (most ratings ranged between -2 and +2 on the scale ranging from -4 to +4), a triangular null distribution was used in the computation of the IRA (see LeBreton & Senter, 2008); Level of agreement according to LeBreton and Senter (2008).

Table B

CFA Models Testing Measurement Invariance across Languages (Dutch, French, and English) in

Sample 2

	X^2	df	р	CFI	RMSEA
Coaching					
1. Invariance of the factor structure (i.e, configural invariance)	6595.69	270	<.001	.933	.067
2. Invariance of the factor loadings (i.e., metric invariance)	6698.23	298	<.001	.932	.064
3. Invariance of the item intercepts (i.e., scalar invariance)	7346.15	326	<.001	.926	.064
Inspirational					
1. Invariance of the factor structure (i.e, configural invariance)	14669.35	270	<.001	.822	.101
2. Invariance of the factor loadings (i.e., metric invariance)	15089.23	298	<.001	.817	.097
3. Invariance of the item intercepts (i.e., scalar invariance)	16695.23	326	<.001	.797	.098
3a. Partial invariance of the item intercepts – item 15 _{Dutch} free	16275.72	325	<.001	.802	.097
3b. Partial invariance of the item intercepts – item 6 _{French} free	16009.13	324	<.001	.806	.096
3c. Partial invariance of the item intercepts – item 12_{French} free	15766.41	323	<.001	.809	.095
Directive					
1. Invariance of the factor structure (i.e, configural invariance)	13046.25	162	<.001	.667	.123
2. Invariance of the factor loadings (i.e., metric invariance)	13083.62	184	<.001	.667	.115
3. Invariance of the item intercepts (i.e., scalar invariance)	14835.28	206	<.001	.622	.116
3a. Partial invariance of the item intercepts – item 109 _{French} free	14373.42	205	<.001	.634	.115
3b. Partial invariance of the item intercepts – item 105 _{Dutch} free	14035.04	204	<.001	.643	.114
3c. Partial invariance of the item intercepts – item 107 _{Dutch} free	13761.34	203	<.001	.650	.113
3d. Partial invariance of the item intercepts – item 110 _{Dutch} free	13636.55	202	<.001	.653	.112
3e. Partial invariance of the item intercepts – item 112 _{French} free	13470.44	201	<.001	.658	.112
Authoritarian					
1. Invariance of the factor structure (i.e. configural invariance)	7763.66	270	<.001	.903	.073
2. Invariance of the factor loadings (i.e., metric invariance)	8051.17	298	<.001	.900	.070
3. Invariance of the item intercepts (i.e., scalar invariance)	9546.95	326	<.001	.881	.073
3a. Partial invariance of the item intercepts – item 90 _{French} free	9061.54	325	<001	.887	.071
3b. Partial invariance of the item intercepts – item 104_{French} free	8713.80	324	<.001	.892	.070
Distrustful					
1. Invariance of the factor structure (i.e. configural invariance)	5758.85	270	<.001	.919	.062
2. Invariance of the factor loadings (i.e., metric invariance)	5877.17	298	<.001	.917	.060
3. Invariance of the item intercepts (i.e., scalar invariance)	6940.15	326	<.001	.902	.062
3a. Partial invariance of the item intercepts – item 88_{Erench} free	6540.32	325	<.001	.908	.060
Withdrawn					
1. Invariance of the factor structure (i.e. configural invariance)	5435.28	270	<.001	.911	.060
2. Invariance of the factor loadings (i.e., metric invariance)	5641.99	298	<.001	.907	.058
3. Invariance of the item intercepts (i.e., scalar invariance)	6812.97	326	<.001	.888	.062
3a. Partial invariance of the item intercepts – item 63_{Dutch} free	6431.83	325	<.001	.894	.060
3b. Partial invariance of the item intercepts – item 62_{Erapch} free	6256.34	324	<.001	.897	.059
Vielding	0200101	021	4001	1077	1003
1. Invariance of the factor structure (i.e. configural invariance)	16638.51	270	<.001	.631	.107
2. Invariance of the factor loadings (i.e., metric invariance)	16716.29	298	<.001	.630	.102
3. Invariance of the item intercents (i.e., scalar invariance)	19447.19	326	<.001	.569	.106
3a Partial invariance of the item intercents $-$ item 57 _{Dute} free	18699.06	325	< 001	586	104
3b. Partial invariance of the item intercepts $-$ item 57 putch free	18003 40	324	< 001	.602	.102
3c. Partial invariance of the item intercepts – item 52 rener free	17531 43	323	< 001	.612	.101
3d Partial invariance of the item intercepts $-$ item 51 _{Dutch} free	17309 25	322	< 001	617	100
3e. Partial invariance of the item intercepts – item 5 ³ French free	17151 62	321	<.001	621	100

Table B (continued)

Participative					
1. Invariance of the factor structure (i.e, configural invariance)	5722.53	231	<.001	.905	.067
2. Invariance of the factor loadings (i.e., metric invariance)	5774.59	257	<.001	.905	.064
3. Invariance of the item intercepts (i.e., scalar invariance)	7425.98	283	<.001	.877	.069
3a. Partial invariance of the item intercepts – item 32 _{French} free	6944.23	282	<.001	.885	.067
3b. Partial invariance of the item intercepts – item 36 _{French} free	6540.03	281	<.001	.892	.065
3c. Partial invariance of the item intercepts – item 39 _{French} free	6381.99	280	<.001	.895	.064

Note. CFI refers to the Comparative Fit Index, and *RMSEA* to the Root Mean Square Error of Approximation. The analyses were performed using the MLR estimator in Mplus 8.4 (Muthén & Muthén, 1998-2017) on the observerratings (N = 15,777). Of those observers, 14,119 (89.5%) completed the questionnaire in Dutch, 1,439 (9.1%) in French, and 219 (1.4%) in English. For each leadership style of the Circumplex Leadership Scan (CLS; Redeker et al., 2014) the change in fit indices was acceptable after imposing equality constraints on the factor loadings (i.e., < .015 for Δ RMSEA and < .010 for Δ CFI), implying that full metric invariance across languages holds. Additional equality constraints on the item intercepts revealed that for Coaching full scalar invariance held, while for the other CLS leadership styles partial scalar invariance could be established. For some dimensions (i.e., directive and yielding), the fit indices for the configural invariance model suggest that a one-factor model does not fit the data well. Although such model fit would be unacceptable in many cases, it is less of an issue for this particular study because we are interested in how the proposed LARI model fits the data *relative* to alternative ways of modeling the multisource data.

Table C

Inter-rater Agreement $(r_{wg(j)})$ for the Leader Behavior Items (Circumplex Leadership Scan) in

Sample 2

	Subordinates	Peers	Superiors
Coaching			•
Coaching item 1	.72	.76	.81
Coaching item 2	.66	.74	.77
Coaching item 3	.67	.73	.76
Coaching item 4	.57	.69	.74
Coaching item 5	.69	.75	.77
Coaching item 6	.61	.68	.74
Coaching item 7	.62	.67	.73
Coaching item 8	.68	.73	.77
Coaching item 9	.68	.73	.77
Coaching item 10	.59	.68	.74
Coaching item 11	.64	.73	.73
Coaching item 12	.66	.72	.77
Coaching item 13	.62	.69	.73
Coaching item 14	.67	.75	.78
Coaching item 15	.66	.74	.78
Inspiring			
Inspiring item 1	.44	.63	.68
Inspiring item 2	.58	.71	.76
Inspiring item 3	.48	.67	.69
Inspiring item 4	.68	.72	.77
Inspiring item 5	.59	.69	.72
Inspiring item 6	.60	.69	.70
Inspiring item 7	.62	.72	.77
Inspiring item 8	.62	.68	.69
Inspiring item 9	.60	.67	.72
Inspiring item 10	.73	.72	.74
Inspiring item 11	.72	.73	.73
Inspiring item 12	.65	.69	.71
Inspiring item 13	.60	.66	.69
Inspiring item 14	.58	.72	.75
Inspiring item 15	.55	.63	.67
Directive			
Directive item 1	.61	.68	.71
Directive item 2	.53	.58	.66
Directive item 3	.58	.66	.71
Directive item 4	.50	.60	.66
Directive item 5	.46	.55	.63
Directive item 6	.55	.62	.67
Directive item 7	.64	.70	.72
Directive item 8	.64	.69	.70
Directive item 9	.60	.66	.69

Table C (continued)

Directive item 10	.34	.46	.54
Directive item 11	.47	.59	.63
Directive item 12	.58	.64	.68
Authoritarian			
Authoritarian item 1	.61	.63	.68
Authoritarian item 2	.59	.61	.70
Authoritarian item 3	.52	.59	.66
Authoritarian item 4	54	61	68
Authoritarian item 5	50	60	.00
Authoritarian item 6	62	.00	.07 73
Authoritarian item 7	.02	.70	.75
Authoritarian itom 8	.40	.50	.05
Authoritarian item 0	.49		.04
Authoritarian item 10	.34	.01	.04
Autoritarian item 10	.49	.38	.04
Authoritarian item 11	.46	.56	.63
Authoritarian item 12	.6/	.71	.75
Authoritarian item 13	.55	.59	.70
Authoritarian item 14	.52	.62	.67
Authoritarian item 15	.57	.63	.69
Distrustful			
Distrustful item 1	.62	.68	.72
Distrustful item 2	.60	.68	.76
Distrustful item 3	.58	.64	.70
Distrustful item 4	.67	.71	.76
Distrustful item 5	.56	.61	.70
Distrustful item 6	.71	.71	.75
Distrustful item 7	.71	.71	.79
Distrustful item 8	66	70	74
Distrustful item 9	68	.70	74
Distrustful item 10	.00 64	70	73
Distructful item 11	59	.70	.75
Distructful itom 12	.59	.05	.09
Distructful item 12	.52	.00	.71
Distrustiui item 13	.59	.00	.70
Distrustful item 14	.63	.68	./3
Distrustful item 15	.59	.67	.72
Withdrawn			
Withdrawn item 1	.60	.66	.71
Withdrawn item 2	.64	.69	.75
Withdrawn item 3	.64	.68	.74
Withdrawn item 4	.55	.64	.69
Withdrawn item 5	.66	.72	.74
Withdrawn item 6	.63	.69	.71
Withdrawn item 7	.57	.67	.73
Withdrawn item 8	.68	.71	.72
Withdrawn item 9	.60	.65	.68
Withdrawn item 10	.59	.64	.71
Withdrawn item 11	62	68	70
	.02	.00	.70

Table C (continued)

Withdrawn item 12	.64	.69	.71	
Withdrawn item 13	.65	.68	.70	
Withdrawn item 14	.53	.65	.67	
Withdrawn item 15	.66	.70	.76	
Yielding				
Yielding item 1	.76	.76	.76	
Yielding item 2	.57	.64	.66	
Yielding item 3	.47	.58	.61	
Yielding item 4	.57	.64	.66	
Yielding item 5	.56	.63	.68	
Yielding item 6	.66	.70	.73	
Yielding item 7	.45	.53	.59	
Yielding item 8	.52	.59	.65	
Yielding item 9	.63	.69	.69	
Yielding item 10	.67	.69	.69	
Yielding item 11	.57	.62	.66	
Yielding item 12	.49	.59	.63	
Yielding item 13	.67	.71	.76	
Yielding item 14	.61	.68	.70	
Yielding item 15	.50	.62	.66	
Participative				
Participative item 1	.67	.70	.73	
Participative item 2	.65	.68	.72	
Participative item 3	.72	.76	.80	
Participative item 4	.75	.80	.81	
Participative item 5	.65	.70	.73	
Participative item 6	.56	.61	.65	
Participative item 7	.79	.82	.83	
Participative item 8	.62	.66	.67	
Participative item 9	.64	.68	.69	
Participative item 10	.60	.68	.74	
Participative item 11	.70	.75	.78	
Participative item 12	.70	.74	.77	
Participative item 13	.64	.70	.74	
Participative item 14	.58	.68	.72	
Mean	.60	.67	.71	
SD	.08	.06	.05	
Agreement	moderate	moderate	strong	
N for $\mathbf{k} = \mathbf{min.} 2$	1198	1102	543	

Note. To control for a slight skew (most ratings ranged between 1 and 4 on the 0 to 4 scale), the slightly skewed random response null distribution was used in the computation of the IRA (see LeBreton & Senter, 2008); Level of agreement according to LeBreton and Senter (2008).