Consulting Psychology Journal

Evaluating a Novel Method of Scoring Multirater Assessments of Leader Competencies

R. Michael Furr, Stephen B. Jeong, John W. Fleenor, and Sirish Shrestha

Online First Publication, October 17, 2024. https://dx.doi.org/10.1037/cpb0000285

AMERICAN PSYCHOLOGICAL ASSOCIATION
Society of Consulting Psychology
The Catalyst Division
Consulting Psychology Journal

ISSN: 1065-9293
https://doi.org/10.1037/cpb0000285

EVALUATING A NOVEL METHOD OF SCORING MULTIRATER ASSESSMENTS OF LEADER COMPETENCIES

R. Michael Furr[1], Stephen B. Jeong[2], John W. Fleenor[2], and Sirish Shrestha[2]
[1] *Department of Psychology, Wake Forest University*
[2] *Center for Creative Leadership, Greensboro, North Carolina, United States*

In leadership development, 360 assessments are multirater instruments in which leaders rate themselves on leadership competencies and are assessed by several raters (e.g., peers, direct reports, bosses). Such assessment methods afford valuable opportunities for both leaders and practitioners alike; however, they also present thorny challenges, including those related to scoring. We propose a Latent Approximation Method (LAM) for scoring 360s that is based on recent modeling advances for multirater data and provide a preliminary empirical evaluation of that method. To an international sample of nearly 17,000 leaders and nearly 150,000 raters, we applied LAM scoring to a 360 measure of several leadership competencies. We examined associations among LAM scores, explored associations between LAM scores and standard averaged scores, and evaluated the incremental predictive utility that LAM scores bring (beyond standard scores) to the prediction of criterion variables. Replicating across subsamples, results revealed considerable overlap between LAM scores and standard scores. Moreover, LAM scores provided significant but small incremental predictive power over Standard Averaging Method scores. These results indicate limited value of LAM scoring for practitioners, and we discuss the implications of these findings for both research and applications using 360 assessments.

R. Michael Furr https://orcid.org/0000-0002-4476-1907

Data analysis scripts are available on GitHub at https://url.ccl.org/mrscoring. Study materials and data are proprietary to the Center for Creative Leadership and are not publicly available. There are no conflicts of interest to declare.

R. Michael Furr played a lead role in writing–original draft, an equal role in conceptualization, and a supporting role in formal analysis and writing–review and editing. Stephen B. Jeong played a supporting role in formal analysis, writing–original draft, and writing–review and editing and an equal role in conceptualization. John W. Fleenor played a supporting role in conceptualization and writing–original draft and an equal role in writing–review and editing. Sirish Shrestha played a lead role in formal analysis.

Correspondence concerning this article should be addressed to R. Michael Furr, Department of Psychology, Wake Forest University, Winston-Salem, NC 27109, United States. Email: FurrRM@wfu.edu

What's It Mean? Implications for Consulting Psychology

This work has implications for both leadership development coaches and researchers alike. For practitioners, it explores a new and potentially useful technique for gaining insights about their client. For researchers, it surfaces new questions toward advancing our understanding of 360 assessments and the interpretation of their results using a novel modeling methodology.

Keywords: multirater assessment, leadership, latent variables, assessment, rating data

Supplemental materials: https://doi.org/10.1037/cpb0000285.supp

In the field of leadership assessment, 360° feedback is a multirater measurement method—often used by leadership development practitioners, including coaches—in which a leader's characteristics (e.g., competencies) are rated by multiple individuals. Raters typically include the leader herself, her direct reports (DRs; i.e., subordinates), her peers, and her immediate supervisor, all of whom rate characteristics ostensibly related to her effectiveness as a leader. This rich method of measurement provides potentially invaluable information that may increase understanding of the psychosocial factors shaping leadership effectiveness. In applied assessment contexts, 360 assessments can help leadership coaches understand clients and guide emerging as well as mature leaders toward greater understanding of themselves, including awareness of the impact their behavior has on others and their organization. These utilities notwithstanding, there are challenges complicating the 360-measurement process and its interpretations. The purpose of the current research is to propose a novel method for scoring and interpreting 360 assessments and to provide a preliminary empirical evaluation of that method.

The Value of 360 Assessments

Three-sixty assessments present psychologists, management, and leadership development practitioners, as well as consultants, with unique value and opportunities, along with thorny challenges. For example, ratings from a leader herself, her supervisor, her peers, and her direct reports may substantially converge or diverge in either direction. Patterns of convergence and divergence across raters and across constructs can, in turn, provide valuable insights for both practitioners and researchers.

Since its inception in the late 1980s, 360 assessments have become one of the most notable innovations in the field of leadership development (Chappelow, 2004). In the following decades, they have been used for a variety of purposes in diverse settings and have gained increasing acceptance and prominence in organizations. According to a recent survey of 85 diverse organizations from 30 different industries, 72% reported plans to continue or increase the use of 360 assessments (Rose & Biringer, 2022).

In applied settings, 360 assessments are indispensable for helping leadership coaches shed light on a leader's strengths and development needs in a way that participants view as more reliable and valid, thus more readily accepted by leaders themselves. These advantages stem from the fact that 360s involve multiple raters—as opposed to one rater, such as one's boss—to enable a varied and comprehensive perspective of a leader's characteristics and contribution (i.e., performance) to their organization. The aggregation of feedback from multiple rater groups, each varying in their relationship to the leader, mitigates—to some extent—the effects of personal biases on the ratings (Fleenor et al., 2020). Compared with assessments relying solely on self-reports or those provided by supervisors, 360s bring a reduced level of subjectivity along with face validity, making the feedback more useful and credible to recipients. This enables organizations to make strategic personnel decisions that are more likely to be accepted (Church et al., 2019). Early practitioners from the Center for Creative Leadership (CCL)—for example, Robert Dorn and Robert Bailey—recognized the importance of not only using assessments for development, but also of the provision of feedback as essential to effective leadership development interventions (Bracken et al., 2001).

As a key part of the 360 process, leaders receive feedback reports detailing results of their assessment. With the assistance of a feedback coach or their supervisor, leaders examine their reports, focusing on both high (strengths) and low ratings (development needs), as well as the differences between their own and others' perceptions of their capabilities and effectiveness. Based on these insights, leaders create individualized development plans designed to help them modify specific behaviors to become more effective.

In addition to facilitating behavior change, 360 assessments enable organizations to promote important values and set expectations by communicating the kinds of behaviors expected from employees. To reinforce valued behaviors, an organization can include specific competencies on their 360 assessment that are important to success in that organization. For example, an organization can communicate the importance of Decisiveness by providing feedback on how well the leaders are performing on that specific competency.

In research settings, 360 assessments have been used extensively to examine relationships between individual differences and leadership effectiveness. This work collectively demonstrates the value that 360s hold in elucidating the relationships among perceived leader capabilities, performance measurement, and various factors moderating these relationships.

For example, using 360 assessments, Braddy et al. (2014) examined how the relationship between leader effectiveness and derailment potential—defined as the likelihood of a leader being demoted or involuntarily dismissed (Leslie & Braddy, 2015)—might vary by rater group. Using random coefficient modeling, relative weight analysis, and polynomial regression, they found that (a) self, direct report, peer, and supervisor ratings of leader effectiveness differed and were associated with derailment potential; (b) peer ratings were most predictive of derailment potential while self-ratings were least predictive; and (c) derailment potential was lowest when self-ratings were lower than other ratings and when self–other rating agreement converged on higher, rather than lower, ratings (Braddy et al., 2014). In another study involving 360 assessment data, Bono et al. (2017) investigated the impact of gender on supervisor ratings of leaders' derailment potential in upper levels of organizations. With a large sample of leaders ($n \approx 12{,}500$), they found evidence of gender bias where ineffective interpersonal behaviors were less prevalent among female managers; however, when present, these behaviors were more damaging to females compared to males (Bono et al., 2017).

The Challenge of Scoring 360 Assessments

The value that 360 assessments bring to practitioners and research psychologists is not without costs and challenges. They are expensive, time-consuming, and oftentimes difficult to obtain. Moreover, the potential for rater divergence to provide important insights also raises important psychometric questions—for example, how should 360 assessments be used and scored to optimize prediction? Should perspectives from specific rater groups (self, peer, boss, direct reports) be given more or less weight when it comes to evaluating performance or derailment potential? Is there additional information in 360 assessments that traditional scoring has not yet tapped? If so, what methods are available to help extract and use the potentially valuable information available in 360 ratings?

In research contexts, the use of multirater data, such as 360 assessments, has been approached in diverse ways. Researchers from areas including clinical psychology, industrial–organizational psychology, psychiatry, developmental psychology, personality psychology, and beyond have developed statistical models and suggestions for optimizing the use of multirater assessment information (Bauer et al., 2013; De Los Reyes et al., 2013; Hoffman et al., 2010; Kraemer et al., 2003; Martel et al., 2017; McAbee & Connelly, 2016). Such approaches are useful for researchers seeking to analyze multirater data in maximally informative ways.

Outside of research, practitioners face the thorny challenge of scoring an individual client's 360 assessments. There are at least three simple and intuitive solutions to this problem in an applied context. One is to aggregate the information from all raters to form an overall score for each scale (e.g., averaging a leader's self-ratings of Decisiveness with her supervisor's ratings of the leader's Decisiveness, with her peers' ratings of the leader's Decisiveness, etc.). A leader would thus obtain a single aggregated score on each scale, representing the individual's standing on the relevant construct in the eyes of the

"typical rater." The appeal of this approach is that it uses all relevant information and provides applied psychologists with just one set of scale scores for a given client. One obvious shortcoming of this approach is that it discards potentially valuable information that is unique to each rater group. This is important to the extent that average ratings provided by different rater groups can, and often do, vary significantly (Fleenor et al., 2010), thus placing undue limitations on the assessors' ability to interpret and draw firm conclusions from aggregated scores.

A second simple solution is to aggregate the information from all raters except the self, to form an overall "informant" or "all but self" score for each scale (e.g., averaging a supervisor's ratings, peer ratings, and subordinate rating of a leader's Decisiveness). A leader would thus obtain a single aggregated score on each scale, representing the individual's standing on the relevant construct in the eyes of the people with whom they work. Although this approach differentiates self-ratings from other ratings, it still discards information that may be unique to each of the other rater groups.

A third potential simple solution is to score responses from each rater group separately, again, using typical scoring procedures (e.g., averaging responses across items on a scale, for each rater group). Thus, an assessor would obtain, say, a Decisiveness score based on the leader's self-report score, a Decisiveness score based on her supervisor's ratings, a Decisiveness score based on an average of her peers' ratings, etc. Although this approach can preserve potentially unique and valuable information provided by each rater, it may obscure information that the various rater groups share or have in common.

In this article, we refer to these simple, intuitive approaches as the Standard Averaging Method (SAM) of scoring 360s. The general problem with these SAM approaches (either in aggregate form or separated by rater groups) is that they confound information that is common across rater groups with information that is unique to each rater group (e.g., self, peer, direct reports). For example, the simple average of one's self-ratings may reflect, to some unknown degree, a view that is *shared by all raters*. Simultaneously, it may also reflect, again to an unknown degree, a view that is *unique to the self*. Similarly, supervisor, peer, and direct reports' ratings may each reflect, to some unknown degree, a view that is shared by raters, while also reflecting, again to an unknown degree, a view that is unique to the individual rater. The notion that different groups are privy to both unique and common information about an individual is not new and aligns well with personality measurement research where asymmetry in the accuracy of personality judgments has been reported (Vazire, 2010). Hence, whether estimated as a single omnibus score or differentiated by rater group, standard aggregated scoring method (i.e., SAM) likely confounds what is common and unique to each rater.

In applied assessment contexts, such confounding might make it difficult for both practitioners and client leaders to tease apart the common and unique insights offered by different groups of raters. For example, practitioners may find it challenging to convey a given leader's level of Decisiveness (as perceived across all rater or rater groups) while at the same time conveying the uniquely high or low level of Decisiveness as perceived by each group of raters. The inability to clearly differentiate and acknowledge such information may compromise the value and potential impact of 360 assessments for client leaders.

In this article, we explore a potential solution to the limitations of the SAM approach by incorporating latent variable analysis that accounts for both unique and shared variances. This approach—termed Latent Approximation Method (LAM)—is outlined in the next section.

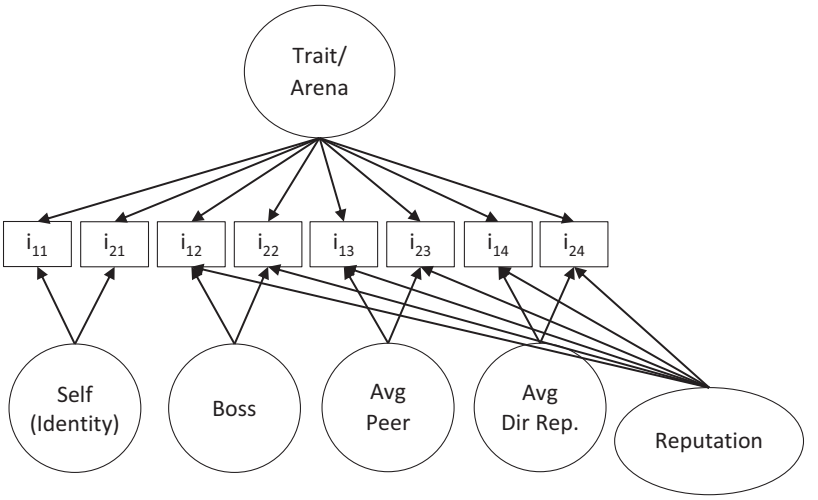
A Potential Solution to the Scoring Challenge: LAM

Recent methodological developments may provide a solution to the aforementioned limitation. Figure 1 partitions 360 ratings of a single characteristic into several latent variables and is based on the Trait, Reputation, Identity model (TRI; McAbee & Connelly, 2016) and the Leadership Arena–Reputation–Identity model (LARI; Vergauwe et al., 2022).

The *Arena* variable (called the *Trait* variable by McAbee & Connelly, 2016) represents variance shared by all raters. It reflects leaders' general levels or capability around a given construct or competency (e.g., Decisiveness) as perceived by the self, boss, peers, and direct reports. For example,

Figure 1

A Trait, Reputation, Identity Model/Leadership Arena–Reputation–Identity Model for 360 Assessments



Note. i = item; Avg peer = average peer; Avg Dir Rep. = average direct report.

it represents the implicit "agreed-upon" level of an individual's Decisiveness, as common across all raters. The *Reputation* variable represents variance shared by all raters except for the self. It reflects leaders' levels of a given construct as seen uniquely by others (i.e., as differing from how the leaders see themselves on the construct). The *Identity* variable represents variance unique to the leaders' self-perceptions (i.e., as differing from how they are seen by their colleagues). Other rater-specific variables represent how the leaders are seen uniquely by each rater group (e.g., boss, peers, direct reports) as separate from the others and as separate from the self. Thus, TRI/LARI models differentiate variance that is common across all ratings (i.e., Arena), variance that is common across some ratings (i.e., Reputation), and variance that is unique to each rater group. They thus provide rich and detailed insights into perceptions of a leader's psychological characteristics.

Although developed relatively recently, TRI/LARI models have already shaped how researchers analyze multirater data. Initial work partitioned variance in multirater ratings of personality traits and leader competencies. For example, McAbee and Connelly (2016, Demonstration 1) applied the TRI model to ratings of the Big Five personality traits. They found that the Trait factor accounted for more variance (from 44% to 69%) than the Reputation or Identity factors in ratings of conscientiousness, agreeableness, neuroticism, and extraversion. Similarly, Vergauwe et al. (2022) applied the LARI model to leadership 360 ratings from two instruments, finding again the Arena (or Trait) factor accounting for most variance in nearly all the scales that they examined. Interestingly, both studies revealed that the relative magnitudes of the factors' variances varied across constructs. This is consistent with previous findings that self–other agreement and interjudge consensus (as operationalized in terms of correlations across focal participants) vary across traits (e.g., Fleenor et al., 1996, 2010; Funder, 1995; Furr et al., 2007; Helzer et al., 2014; Vazire, 2010).

This initial work has been followed by research revealing correlations between TRI/LARI factors and important outcomes (Blickle et al., 2018; Connelly et al., 2022; Wihler et al., 2022). For example, Connelly et al. (2022) examined the association between job performance and multirater ratings of the Big Five personality constructs. They discovered that different TRI factors (from the Big Five) were associated with job performance, although findings varied across personality constructs and across different assessments of performance. Wihler et al. (2022) also applied the TRI model to multirater ratings of the Big Five and of job performance, finding that personality–performance correlations varied depending on the specific TRI-based factors involved (e.g., the Identity factor from ratings of

Conscientiousness was correlated with a corresponding "Façade" factor from ratings of job dedication). Thus, the field appears to have enthusiastically embraced the potential value of using TRI/LARI models to address important questions in a research context.

The TRI/LARI models have generated significant and increasing interest within a research context, but their use in an applied assessment context has not yet been explored. In an applied context, practitioners and coaches could potentially use information from these models to generate scores approximating an individual leader's standing on the six latent variables (or factors) in the model. That is, a LAM approach to scoring may provide practitioners with insights into constructs reflected in the TRI/LARI model of 360 assessments. Put differently, compared to the conventional SAM approach to scoring, a LAM approach might provide a more nuanced understanding of clients' strengths and areas of opportunity.

The development and implementation of LAM-based scoring for a given 360 assessment would require research-based preliminary steps and subsequent applied scoring algorithms. In two preliminary steps, researchers would use a large normative sample of leader respondents who have participated in a 360 assessment and would fit a TRI/LARI model (e.g., Figure 1) to the item-level assessment data. They would then extract item-level scoring parameters (e.g., factor scoring weights) that can be applied to (future) leaders' responses. Subsequently, practitioners can apply those scoring weights to item-level responses from a client leader's 360 assessment. This would subsequently result in six scores for a given leader, approximating the leader's standing on the six latent variables in the TRI/LARI model. These scores may then be interpreted and used separately from or in conjunction with SAM-based scores to further gain insight into a given leader's characteristics.

LAM-based scoring may provide novel insights into multiple aspects of the leader's psychological characteristics, as perceived at multiple levels and perspectives. Such insights might go well beyond those obtained via the aggregated, SAM-based scoring methods.

Evaluating the Potential Solution

LAM scores may or may not provide practitioners with insights not captured by simpler scoring methods. However, the value of this novel scoring method has not been empirically evaluated. Two key empirical questions can begin to reveal whether LAM scores, based on the TRI/LARI model, have practical value. First, to what degree do LAM scores provide information that differs meaningfully from SAM scores? Second, do LAM scores predict important outcomes (e.g., job performance) better than SAM scores? The present study addresses both questions.

These questions are important for at least two reasons. First, a novel LAM-based method of scoring 360s may present new avenues for understanding and managing 360 assessments in a way that enhances their value beyond conventional scoring and use. If LAM scores can reveal deeper or more refined insights into a leader's psychological and behavioral characteristics, then they may provide a greater opportunity to create more elegant and nuanced development solutions that optimize time, effort, and cost associated with developing leadership capabilities. This would, in turn, advance organizational performance in a host of ways. However, it is not yet clear whether LAM scores provide insights that go beyond SAM scores.

Second, understanding the potential value of LAM scores, instead of or along with SAM scores, is also important because adopting LAM-based scoring likely entails significant challenges in an applied context. Assessment systems are generally designed, administered, and validated through the use of SAM-based scores, making those scores more familiar to practitioners. Thus, shifting to (or adding) a novel set of LAM-based scores would likely entail considerable costs related to time, resources, and education. Existing SAM-based scoring mechanisms would need to be expanded, revised, or replaced; coaches would need training on the appropriate interpretation of a new set of scores; and even the feedback materials (e.g., score reports) would require substantial revision. Such costs are worthwhile only if the new—LAM-based—scoring method provides meaningful value over and above the existing—SAM-based—method. In sum, there is a need to empirically evaluate the potential value of a LAM-based scoring method for 360s.

The Present Study

We provide an empirical evaluation of a LAM-based scoring method for 360 assessments. Our primary empirical focus is on the two questions mentioned earlier: To what degree do LAM scores provide information that differs meaningfully from simple SAM scores? To what degree do LAM scores predict outcomes better than SAM scores? The answers to these questions will have implications for researchers and practitioners using 360s.

We used a two-phase process based on a sample of more than 16,000 leaders and nearly 150,000 total respondents. Using one random half of the data as a normative sample, we first fit a TRI/LARI model to a set of 360 assessments, extracted factor scoring weights, computed LAM-based scores for individual leaders, correlated those scores with outcomes (i.e., ratings of derailment and performance), and compared them to scores obtained via conventional aggregation methods. Using the other half of the data as a cross-validation sample, we applied the scoring weights (as obtained from the normative sample and as could be done by practitioners) and repeated the correlational analyses. The latter analysis was deemed crucial because it represents the potential value of the scoring procedure (derived from our normative sample) for assessment practitioners working with leaders in an applied context.

Method

Participants

Focal participants are 16,669 leaders who, as part of participation in a leadership development program, completed 360° feedback assessment through the CCL from 2018 to 2022. Each focal participant rated her/himself and was assessed by several raters, including subordinates, peers, and bosses. Participants selected raters and provided their email addresses. A total of 144,811 (boss = 16,255, peer = 65,514, direct reports = 63,042) raters assessed the focal participants. Participants and raters completed assessments online. Prior to participation, each participant signed CCL's Consent Form allowing research use of their data.

For our main analysis, we retained participants with ≥1 response on each of the three items on a Decisiveness scale (see below) from each of four rater groups—self-ratings, direct reports, boss, and peer. This process resulted in a subsample of $N = 13{,}867$. Sampling was based on data availability rather than on power estimates. However, our analyses were extremely well-powered to detect even small effects (e.g., power = .99 to detect $r = .05$). Because nearly all results were significant at extreme p levels, we do not provide exact p values. To evaluate replicability across multiple competencies (beyond Decisiveness), we report analyses of additional scales in an online Supplemental Material. Similar data retention methods were used for each scale; thus, sample size varies from scale to scale. However, samples for all scales ranged from $N = 8{,}845$ (Compassion and Sensitivity) to $N = 13{,}867$ (Decisiveness), excluding null values. To ensure participants' data privacy, responses to demographic questions were optional and varied by question. Among those providing demographic information, participants represented >120 countries across Africa, Asia, South America, Australia, Europe, and North America, with the largest proportion from the United States (69%). Participants were primarily male (65%) and Caucasian/White (75%). They were 43.2 years old on average ($SD = 8.1$) with 87% holding at least a bachelor's degree. They represented five managerial levels: first-level supervisors (4%), middle managers (27%), upper–middle managers (38%), executives (21%), and top leadership (e.g., chief financial officer, chief executive officer; 4%). They also represented a wide range of industries, including education, government, aerospace, manufacturing, banking, and pharmaceuticals.

Measures

Our main analyses focused on the Decisiveness competency scale as assessed via CCL's proprietary 360, multirater assessment—Benchmarks for Managers (i.e., "Benchmarks"). Benchmarks (Leslie & Braddy, 2015) assesses 16 leadership competencies, including Decisiveness. Benchmarks was based on CCL's "Key Events" research—initiated in the late 1980s, involving in-depth interviews with

79 successful executives from Fortune 100 organizations—that sought to understand transformational experiences or events that served as foundations to leadership success (Lindsey et al., 1987). Each competency scale contains three to 13 items, with reliabilities ranging from .74 to.93 (Leslie & Braddy, 2015). The Decisiveness scale, in particular, contains three items with coefficient alphas ranging from .71 (Self, $N = 24{,}096$) to .79 (All Other Raters, $N = 257{,}111$). Decisiveness is conceptualized as a preference for making "quick and approximate actions in many management situations" and captured using three items—"Does not hesitate when making decisions," "Does not become paralyzed or overwhelmed when facing action," and "Is action-oriented" (Leslie & Braddy, 2015).

To evaluate replicability across competencies, we report analyses of three additional competencies (Compassion and Sensitivity, Being a Quick Study, and Respect for Differences) in an online Supplemental Material. Our main analysis focuses on Decisiveness for three reasons. First, focusing on a single competency allows us to elaborate the logic and process of the LAM-based scoring procedure in depth for this initial introduction of the method. Second, Decisiveness is consistently rated—by focal participants as well as raters—as one of the most important of the Benchmarks competencies. Third, although all Benchmarks competencies are important in the context of leadership, Decisiveness may be of particular interest in the context of this study.

Benchmarks also includes content that we used as criterion variables. It includes a 10-item Derailment scale assessing "Difficulty Changing or Adapting" (e.g., "Does not use feedback to make necessary changes in his/her behaviors"). It also includes a general performance item: "How would you rate this person's performance in his/her present job?" Participants (the leaders themselves) and raters (direct reports, peers, and bosses) completed the assessment of participants' Decisiveness. Raters completed the assessment of participants' derailment and performance.

Results

All analyses were conducted via R programming language, with TRI/LARI models fit via the lavaan package (Rosseel, 2012). Analyses were essentially exploratory and were not preregistered. Syntax is available at https://url.ccl.org/mrscoring.

Obtaining TRI/LARI Factor Scores From the Normative Sample

Splitting the sample of participants into random halves, we obtained one subsample of 8,334 leaders. We retained only those leaders for whom data were available on all Decisiveness items from self-ratings, at least one direct report rating, at least one boss rating, and at least one peer rating. This produced a final normative sample of 6,993 leaders with complete data needed to estimate the TRI/LARI model and to obtain factor (i.e., LAM) scores.

We fit the TRI/LARI model (see example in Figure 1, illustrating two items per rater) based on procedures outlined by Vergauwe et al. (2022).[1] For each item, we averaged across all available ratings within each rater group (i.e., direct report, boss, peer). We thus calculated an average direct report rating, an average boss rating, and an average peer rating for each Decisiveness item. Table 1 presents means and standard deviations for the 12 variables. The TRI/LARI model fit well (root-mean-square error of approximation = .06, Tucker–Lewis index = .95, comparative fit index = .96).

To implement a LAM-based scoring method (based on the TRI/LARI model), we extracted factor scoring weights (see Table 1).[2] These weights allow practitioners to weight and sum item-level 360 data, enabling them to estimate individual leaders' scores (i.e., LAM) on each of the six TRI/LARI-based Decisiveness latent variables.

[1] Our model differed from the Vergauwe et al. (2022) model in that we did not include correlated residuals across items with identical content. We omitted these parameters for two reasons: (a) The model without those extra parameters fit the data extremely well, and (b) the model with those parameters encountered significant problems with convergence.

[2] Weights are based on the regression scoring method in lavaan.

Table 1

Item Means, Standard Deviations, and Factor Scoring Weights Derived From the Trait, Reputation, Identity Model/Leadership Arena–Reputation–Identity Model (Normative Sample)

		Factor scoring weight					
Item	*M* (*SD*)	Self	Direct report	Boss	Peer	Reputation	Trait/Arena
Self-i1	3.617 (0.807)	.248	−.017	−.019	−.018	.000	.023
Self-i2	3.920 (0.777)	.261	−.010	−.011	−.010	.000	.013
Self-i3	4.191 (0.683)	.183	−.024	−.026	−.025	.000	.031
DR-i1	4.067 (0.572)	−.042	.147	−.057	−.054	.004	.064
DR-i2	4.295 (0.529)	−.004	.237	−.025	−.026	.031	.006
DR-i3	4.357 (0.500)	−.046	.306	−.041	−.036	−.029	.070
Boss-i1	3.822 (0.780)	−.032	−.038	.142	−.040	.003	.048
Boss-i2	4.095 (0.775)	−.007	−.018	.262	−.019	.015	.010
Boss-i3	4.324 (0.711)	−.034	−.028	.165	−.028	−.018	.051
Peer-i1	3.928 (0.561)	−.073	−.088	−.097	.060	.006	.109
Peer-i2	4.143 (0.539)	−.015	−.046	−.048	.215	.045	.023
Peer-i3	4.275 (0.502)	−.068	−.050	−.059	.205	−.044	.102

Note. $N = 6{,}993$. i = item; DR = Direct Report.

To generate LAM scores for each participant in the normative sample, we mean-centered the ratings on each of the 12 items. We then weighted the (mean-centered) item-level ratings by the relevant factor scoring weight and summed across items.[3] Thus, for example, a participant's Trait LAM score is (see Equation 1):

$$\text{Trait}_p = \sum_{i=1}^{12} W_{it} C_{ip}, \tag{1}$$

where W_{it} is the factor scoring weight for item i on the Trait latent variable, and C_{ip} is the mean-centered rating of participant p on item i.

The scoring weights begin to reveal the meaning of LAM scores, and the pattern in Table 1 (for Decisiveness) is similar to patterns for the other competencies we examined. Specifically, for a given LAM score (e.g., self, boss), weights are positive for items rated by the source associated with the LAM score, but are negative for other items. For example, for Self-LAM scores, self-rated items are weighted positively, while items rated by others are weighted negatively. Based upon this pattern of weights, leaders receive the highest Self-LAM score when rating themselves highly and when rated *low* by other raters (and they receive the lowest Self-LAM score when rating themselves low and when rated *high* by others). Thus, LAM scores represent a leader's competency level as uniquely seen by each type of rater (e.g., Self-LAM scores are highest when the self is unique in providing high ratings and is lowest when the self is unique in providing low ratings). Reputation LAM scores differ from other LAM scores as they include near-zero weights for self-ratings and a blend of weak positive and negative weights from other raters. Although this pattern generally replicates across the other three competencies, the other competencies' Reputation weights are generally larger than those for Decisiveness.

Values above Table 2's diagonal are correlations among TRI/LARI estimated factor (i.e., LAM) scores. Although the TRI/LARI model is based on orthogonal latent variables, the *estimated* LAM scores are not strictly orthogonal. Correlations were generally low across the Self, Direct Report, Boss, Peer, and Reputation variables, with most ranging from −.22 to .15 ($p < .05$ for correlations $\geq |.03|$, based on $N = 6{,}993$). This indicates that those estimated LAM scores provide relatively unique perspectives on a

[3] Mean-centering ensures that the estimated factor scores obtained by this weighting algorithm are identical to those produced directly by lavaan.

Table 2
Correlations Among Decisiveness Scores: LAM-Based Scores and SAM-Based Scores

LAM-based score	1	2	3	4	5	6
1. Self	—	−.09	−.06	−.20	.03	.18
2. Direct report	−.10	—	−.18	−.21	.10	.29
3. Boss	−.07	−.18	—	−.22	.08	.28
4. Peer	−.18	−.22	−.21	—	.15	.44
5. Reputation	.03	.08	.08	.13	—	.05
6. Trait/Arena	.19	.28	.28	.44	.00	—
SAM-based score	**1**	**2**	**3**	**4**	**5**	**6**
1. Self	—	.21	.23	.23	.30	.65
2. Direct report	.21	—	.27	.37	.69	.64
3. Boss	.22	.26	—	.37	.79	.72
4. Peer	.24	.37	.36	—	.75	.69
5. All non-self-raters	.30	.69	.79	.74	—	.92
6. All raters	.65	.63	.72	.69	.92	—

Note. Correlations above the diagonal in each matrix are from the normative sample, and those below the diagonal are from the cross-validation sample. $N_{\text{normative}} = 6{,}993$; $N_{\text{cross-validation}} = 6{,}874$. Correlations $\geq |.03|$ are significant at $p < .05$. LAM = Latent Approximation Method; SAM = Standard Averaging Method.

participant's Decisiveness competency. The correlations with the estimated Trait/Arena scores are all positive and somewhat higher, ranging from .05 to .44. Across all LAM scores, the average correlation is low, at .04. As shown in the Supplemental Tables S1a, S7a, and S13a, similar findings emerge for three additional competencies—Compassion and Sensitivity, Being a Quick Study, and Respect for Differences.

Comparing LAM Scores to SAM Scores (Normative Sample)

An important question is whether the LAM scores provide information that differs from SAM scores. In applied assessment contexts, assessors usually rely on SAM-based scoring—that is, calculating several simple sums or average for a given construct. For example, a "Self-rated Decisiveness" mean score is calculated by averaging a leader's responses to the three Decisiveness items. Similar averages or sums are often calculated and reported for ratings from Direct Reports, Bosses, Peers, and All Other Raters (i.e., all raters excluding self-ratings). Using the same approach, we calculated SAM scores for participants in the normative sample.

Values above the diagonal in Table 2 are correlations among SAM scores. These correlations are strong as compared to those among the LAM scores. Correlations among SAM scores are positive, ranging from .21 to .92 (mean $r = .52$, as compared to mean $r = .04$ for LAM scores in 2a). The difference between the two sets of correlations reveals that the TRI/LARI model produces scores from which common variance has been substantially removed. That is, LAM scores may reflect more directly what is unique to each rater perspective, whereas SAM scores may confound what is unique and what is common across perspectives. Again, similar findings emerge for other competencies, as shown in Supplemental Tables S1b, S6b, and S13b.

Despite the de-confounding achieved via LAM procedures, LAM scores convey information highly similar to that provided by SAM-based scores. Table 3 presents correlations between the LAM-based and SAM-based scores, revealing substantial correlations between conceptually similar scores ($p < .05$ for all correlations $\geq |.03|$). For example, LAM-based self-rated Decisiveness scores are correlated at $r = 95$ with SAM-based self-rated Decisiveness items. Similar findings emerge for each rater group (Direct Reports, etc.). Interestingly, an exception to this is the relatively low correlation ($r = .20$) between the TRI/LARI Reputation score and the "All Other Rater" average. These patterns of overlap generally

Table 3
Correlations Between LAM-Based Scores and Standard Averaging Method-Based Scores

	SAM-based score					
LAM-based score	Self	Direct report	Boss	Peer	All non-self-rater	All rater
Normative sample						
Self	.95	.01	.03	−.03	.01	.40
Direct report	.00	.90	−.02	.02	.36	.29
Boss	.03	−.02	.90	.01	.48	.40
Peer	−.05	.02	.02	.86	.36	.27
Reputation	.03	.18	.10	.19	.20	.17
Trait/Arena	.47	.65	.66	.82	.94	.95
Cross-validation sample						
Self	.95	.01	.02	−.01	.01	.40
Direct report	.00	.90	−.02	.01	.35	.28
Boss	.02	−.02	.91	.01	.49	.40
Peer	−.03	.02	.01	.86	.35	.27
Reputation	.01	.15	.08	.15	.16	.13
Trait/Arena	.48	.65	.65	.82	.94	.95

Note. $N_{\text{normative}} = 6{,}993$; $N_{\text{cross-validation}} = 6{,}874$. Correlations $\geq |.03|$ are significant at $p < .05$. LAM = Latent Approximation Method.

emerged with other competencies as well (see Supplemental Tables S2, S8, and S14). However, Reputation scores overlap more strongly with the "All Other Rater" average ($r = .66-.81$) for the other competences than for Decisiveness, though still more weakly than the other LAM scores.

Predicting Leader Performance and Derailment (Normative Sample)

A LAM-based scoring method for 360 assessments is most likely to be integrated into both research and applied assessment if it provides valuable insights beyond those provided by SAM-based scoring methods. To examine this, we compared the two scoring methods' ability to predict job performance and derailment.

As noted earlier, each participant's performance and derailment were rated by direct reports, peers, and bosses. For each criterion variable (i.e., performance and derailment), ratings were positively correlated across rater groups (e.g., performance: r between direct reports and peers = .37, r between direct reports and bosses = .29, r between bosses and peers = .41). We thus aggregated to obtain a general performance criterion variable and a derailment criterion.

Both the LAM and SAM sets of scores are correlated with performance (as shown in Table 4); however, the pattern of associations differs across sets. The LAM-based score correlations range from −.12 to .58, with average absolute value of .27 and standard deviation of .21 (note that all correlations are significant at $p < .05$, due to the large sample size, $N = 6{,}993$). In contrast, the SAM-based scores are more strongly but uniformly correlated with performance, ranging from .07 to .62, with an average absolute value of .44 and standard deviation of .17. The difference is consistent with the fact that LAM scores are less confounded with each other (in terms of what is unique and common to each score). Thus, each LAM score is less strongly correlated with the criterion than a conceptually similar averaged score; however, each SAM-based score may provide little information that is uniquely useful (as differentiated from other averaged scores).

Evaluating the overall predictive power of each set of scores and the unique predictive information of each score, we examined a series of regression models. In the first model, we evaluated the predictive power of the LAM-based scores by entering the six LAM scores as predictors of the performance criterion. As show in Table 4's "Model 1" column, the LAM-based scores explained 43% of the variance in performance ($p < .05$). Although all individual predictors were statistically significant (again, with $N = 6{,}993$), the strongest were the direct reports, peers, bosses, and Trait/Arena scores. This indicates that Decisiveness ratings from each group of raters (i.e., direct reports) provided

Table 4

Associations Between Decisiveness Scores and Performance Criterion: Correlations and Regression

	Normative sample				Cross-validation sample			
		Standardized weights from regression model				Standardized weights from regression model		
Decisiveness score	*r* with perf.	1	2	3	*r* with perf.	1	2	3
TRI/LARI								
Self	−.12	−.04		−.93	−.12	−.04		−.74
Direct reports	.31	.37		$-.16^{NS}$	.29	.38		$.08^{NS}$
Boss	.24	.32		−.41	.24	.33		$-.19^{NS}$
Peer	.30	.34		−.40	.30	.36		$.03^{NS}$
Reputation	.07	−.06		−.42	.06	−.04		−.24
Trait/Arena	.58	.24		−2.83	.56	.21		−1.57
Averaged								
Self	.07		−.13	1.28	.07		−.11	.94
Direct reports	.48		.32	1.18	.47		.33	.65
Boss	.44		.27	1.45	.43		.21	.93
Peer	.49		.30	1.98	.48		.32	.98
All non-self	.62				.61			
All	.53				.51			
R^2		.43	.41	.43		.42	.40	.42

Note. $N_{normative}$ = 6,993; $N_{cross\text{-}validation}$ = 6,874. Correlations ≥ |.03| are significant at $p < .05$. perf. = performance; TRI = Trait, Reputation, Identity model; LARI = Leadership Arena–Reputation–Identity model; NS = not statistically significant.

predictive insights about performance. At the same time, shared perceptions of Decisiveness across all raters (including self) also provided significant predictive information.

In a second model, we evaluated the predictive power of SAM-based scores by using four SAM scores—self-ratings, direct report ratings, boss ratings, and peer ratings—to predict performance. All six SAM scores cannot be entered simultaneously, because the "All Other Rater" average and "All Raters" average are entirely multivariately redundant with the other four scores. As shown in Table 4's "Model 2" column, the SAM scores explained 41% of the variance in performance ($p < .05$), slightly less than (by 2%) the variance explained by the LAM-based scores. Again, all individual predictors were statistically significant. However, this model does not clearly reflect or distinguish both the predictive insights that are unique to each rater group *and* insight that is common across rater groups (because the "All Other Raters" score and the "All Raters" scores cannot be included in the same model as the specific rater group scores).

In a final model, we evaluated whether the LAM-based scores provided predictive information beyond that provided by the SAM-based scores. In this model (Model 3), we added the six LAM-based scores to the four SAM-based scores as predictors of the performance criterion. This additional set of predictors increased R^2 from .41 to .43 (change in R^2 $p < .05$), suggesting that the LAM-based scores provide a significant but small amount of predictive information beyond that provided by SAM-based scores.[4]

As shown in Supplemental Tables S3, S9, and S15, similar findings emerged for other competencies. Specifically, when predicting performance ratings, LAM scores provide R^2 increments of .06 (over SAM scores) for Compassion and Sensitivity, .01 for Quick Study, and .01 for Respect for Differences (all $p < .05$). One difference between Decisiveness and other competencies was that Reputation LAM scores predicted performance more strongly for the other competencies (r = .29−.55) than for

[4] Note, the regression weights in this model are available in Table 4 but are not of primary concern, as we would not anticipate entering all 10 predictors into a single model, in practice. Our focus in Model 3 is on the proportions of variance explained by the two sets, rather than the value of individual predictors.

Decisiveness ($r = .07$). This is likely because (as noted earlier) Reputation scoring weights were stronger for the other competencies than for Decisiveness; thus, Reputation scores reflect more meaningful variance for the other competencies.

To evaluate replicability across criteria, we conducted similar analyses predicting derailment (see Table 5). Results were generally consistent with those obtained for performance. As compared to SAM scores, LAM scores were less strongly but also less uniformly correlated with derailment. In addition, LAM scores, as a set, added incremental predictive utility beyond SAM scores. Several noteworthy differences appeared as well. First, LAM-based Reputation scores appear more strongly predictive of derailment than of performance. Second, LAM-based Trait/Arena scores appear more weakly related to derailment than to performance. Third, the incremental predictive utility of LAM scores was greater for derailment (approximately 5%) than for performance (2%).

Cross-Validation and Replication

For practitioners wishing to use LAM-based scores (based on the TRI/LARI model) in an applied context, the normative sample's factor scoring weights (Table 1) can be applied to responses for individual leaders. Ideally, the LAM-based scoring weights derived from the normative sample should generalize to new individuals seeking leadership coaching and development. Our next goal is to evaluate this generalizability—how do the LAM-based scoring weights function in a sample of completely new participants?

We now turn to the cross-validation subsample of participants ($N = 6{,}874$) who had complete ratings on the three Decisiveness items as rated by each of the four rater groups (i.e., for a total of 12 ratings per leader). When working with individual clients, practitioners may not be in a position to fit TRI/LARI models to a large database and extract LAM-based factor score estimates for an individual leader. Rather they would compute individuals' LAM-based scores by using existing factor scoring weights and applying those weights to the ratings of that individual. We approached the cross-validation sample in this way.

We first mean-centered the cross-validation participants' Decisiveness ratings, by subtracting each raw rating from the normative means in Table 1. We then weighted and summed those mean-centered

Table 5

Associations Between Decisiveness Scores and Derailment Criterion: Correlations and Regression

	Normative sample				Cross-validation sample			
		Standardized weights from regression model				Standardized weights from regression model		
Decisiveness score	*r* with derail	1	2	3	*r* with derail	1	2	3
TRI/LARI								
Self	.16	$.00^{NS}$		1.08	.16	$-.01^{NS}$		.88
Direct reports	−.26	−.45		$.01^{NS}$	−.28	−.48		$-.11^{NS}$
Boss	−.23	−.42		.22	−.25	−.45		$.19^{NS}$
Peer	−.31	−.50		.43	−.29	−.50		$.12^{NS}$
Reputation	−.20	−.05		.35	−.18	−.04		.26
Trait/Arena	−.44	$.03^{NS}$		3.42	−.44	.05		2.66
Averaged								
Self	.02		.19	−1.52	.01		.19	−1.24
Direct reports	−.40		−.27	−1.15	−.41		−.28	−.89
Boss	−.37		−.24	−1.37	−.38		−.25	−1.22
Peer	−.43		−.29	−2.38	−.41		−.27	−1.67
All non-self	−.53				−.53			
All	−.41				−.42			
R^2		.36	.32	.37		.37	.32	.38

Note. $N_{normative} = 6{,}993$; $N_{cross\text{-}validation} = 6{,}874$. Correlations $\geq |.03|$ are significant at $p < .05$. TRI = Trait, Reputation, Identity model; LARI = Leadership Arena–Reputation–Identity model; NS = not statistically significant.

scores using the factor scoring weights from the normative sample (again from Table 1). For example, we obtained an estimate of the LAM-based Self/Identity score for the first participant in the cross-validation sample as shown in Table 6. This participant has "raw" ratings ranging from 3.67 to 5.0 across the three Decisiveness items and across four rater groups. Using the normative sample means (see Table 1), this participant's item ratings were mean-centered (e.g., 4.000 − 3.617 = .383, for Self-rating of Item 1). These centered ratings were then weighted by the relevant factor scoring weights as derived from the normative sample (again, see Table 1 for the Self/Identity factor weights and the weights for all other factors). For example, the centered self-rating on the first item was multiplied by the scoring weight for that item on the LAM-based Self/Identity factor to produce a weighted rating − .383 × .248 = .095. These weighted ratings are summed across the 12 ratings to produce the participant's estimated score on the Self/Identity factor: $\text{Self/Identity}_p = \sum_{i=1}^{12} W_{it} C_{ip} = .115$.

We examined correlations among the LAM-based Decisiveness scores for the six factors (below the diagonal in Table 2), correlations among SAM-based scores (below Table 2's diagonal), and correlations among the sets of scores (Table 3). Across analyses, results in the cross-validation sample were highly consistent, and often identical, with those in the normative sample. Not surprisingly, LAM scores were relatively weakly correlated with each other (Table 2) but highly correlated with conceptually similar SAM-based scores (Table 3).

We next examined links between Decisiveness and the criteria of performance and derailment, by computing the same correlation and regression analyses as in the normative sample. As Tables 4 and 5 show, results are highly consistent with those in the normative sample. For example, LAM scores accounted for 42% of the variance in performance (compared to 43% in the normative sample), which was 2% more than was accounted for by SAM scores (it was also 2% more in the normative sample). In sum, the correlational properties of a LAM-based scoring procedure replicated well when applying our normative sample's means and scoring weights to the cross-validation sample. Finally, we examined cross-validation samples for the three additional competencies and found results very similar to those with Decisiveness. Results are in Supplemental Tables S4–S6, S10–S12, and S16–S18.

Discussion

When working with 360 assessments, both practitioners and researchers face opportunities and challenges. Although 360s provide opportunities for insights into leaders' characteristics and competencies, they also entail practical and methodological challenges. One methodological challenge involves

Table 6
Example of Calculating a LAM Score: The LAM-Based Self/Identity Score for the First Participant in the Cross-Validation Sample

Rater-item	Raw rating (R)	Norm mean (M)	Centered rating ($R - M = C$)	Factor scoring weight for Self/Identity (W)	Weighted rating ($W \times C$)
Self-i1	4.000	3.617	.383	.248	0.095
Self-i2	4.000	3.920	.080	.261	0.021
Self-i3	4.000	4.191	−.191	.183	−0.035
DR-i1	3.667	4.067	−.401	−.042	0.017
DR-i2	4.000	4.295	−.295	−.004	0.001
DR-i3	3.667	4.357	−.691	−.046	0.032
Boss-i1	4.000	3.822	.178	−.032	−0.006
Boss-i2	5.000	4.095	.905	−.007	−0.006
Boss-i3	5.000	4.324	.676	−.034	−0.023
Peer-i1	4.000	3.928	.072	−.073	−0.005
Peer-i2	4.429	4.143	.285	−.015	−0.004
Peer-i3	3.857	4.275	−.417	−.068	0.028
Sum (estimated factor score):					0.115

Note. LAM = Latent Approximation Method; i = item; DR = Direct Report.

scoring 360 assessments in a way that helps practitioners and researchers distinguish between what is common and unique across groups of raters. Recent advances in statistically modeling 360s might be leveraged to address this challenge and aid practitioners in their quest toward developing more elegant solutions to assessing and enhancing leadership capabilities. Specifically, the TRI model and the closely related LARI model have generated intense interest recently among researchers studying 360 assessments. These models can be used in a novel way to generate scores for individual leaders, potentially offering insights not accessible with standard scoring approaches. The purpose of the current article was to explain such a method of scoring 360s and to evaluate that method. We examined 360 ratings of four competencies, gathered from more than 16,000 leaders and nearly 150,000 raters, representing more than 120 countries from around the world. As an empirical evaluation of the utility of a LAM of scoring based on TRI/LARI models for multirater assessments, we empirically addressed two primary issues.

The first was the degree to which LAM scores might provide information distinct from that provided by simple SAM scores. Results were mixed. Although Reputation LAM scores differed somewhat from SAM scores, LAM scores were generally very strongly correlated with conceptually similar SAM scores (e.g., the average self-rating). This suggests that LAM scores generally provide little information beyond that provided by individual SAM scores. This indicates limited value for LAM scores.

The second issue was LAM scores' predictive utility—that is, whether LAM scores add incremental validity beyond that obtained via SAM scores. Our findings reveal that sets of LAM scores add significant (with our large sample sizes) but small incremental value beyond sets of SAM scores. Across all four competencies and both criterion variables, the six LAM scores together accounted for approximately 1%–6% more of the variance in performance than did the SAM scores. Thus, as a set, LAM scores predict at least two outcomes (i.e., performance and derailment) slightly better than SAM scores. This pattern was consistent for all four competencies to varying degrees, and it suggests that some (perhaps a small amount of) information may be available in a set of LAM scores that is not available in a set of SAM scores.

Given these somewhat equivocal results, we conducted additional qualitative idiographic comparisons of individual leaders' Decisiveness LAM and SAM scores. We attempted to identify meaningful patterns of divergence between the sets of scores at the level of individual leaders (i.e., gauging LAM scores' utility in a leadership development consulting context). We surmised that areas of divergence—if they can be explained and reproduced—could potentially highlight LAM scores' unique value tied to its ability to decompose unique from common variances. This insight could be used by practitioners and researchers alike to gain a better understanding of the relationships among averaged rater group scores across rater groups. Based on the relatively low correlation between Reputation (LAM) and All Informants (SAM) scores, we sorted the 6,000 leaders in the normative sample based on Reputation scores. We then compared the 20 leaders with the highest Reputation scores to the 20 with the lowest, attempting to determine whether this helped reveal any differences between LAM and SAM scores that were clear, meaningful, systematic, and potentially useful in a coaching context. Overall, the patterns of LAM and SAM scores were largely similar and consistent with our previous findings. In sum, our quantitative analyses, our qualitative examination, and our experience with leadership development consulting point to the conclusion that a LAM-based scoring procedure likely offers little applied advantage over more familiar SAM-based procedures. That said, our study raises questions with potential implications for leadership development and research.

Implications for Leadership Development

Although scores based on TRI/LARI models might offer little advantage over simpler (and more intuitive) SAM scores, TRI/LARI models may still hold value for consulting psychologists working in leadership development. We see at least three areas of potential value that deserve further attention. First, TRI/LARI models have been used to partition variance in leadership characteristics, in terms of variance attributable to the Trait/Arena factor, to the Identity factor, and so forth. To our knowledge, such decompositions have not yet been considered in the context of consulting. It is worth exploring whether information from such decompositions can be used in a consulting context.

Second, our study revealed that the Reputation (LAM) factor might differ from the other LAM factors, in terms of its relatively nonredundancy (i.e., somewhat lower correlation) with a corresponding SAM concept. The redundancy of Reputation LAM scores (with the "All Other Rater" SAM score) varied considerably across competencies, as did their association with criterion variables. This variation makes it difficult to draw general conclusions about the robustness, meaning, and utility of Reputation LAM scores. However, results indicate that Reputation LAM scores might provide somewhat novel information to consultants working in leadership development, even if no other LAM scores do. Future research can build on the insight provided by our findings to explore both the practical meaning of the Reputation factor and its potential use in a consulting context. For example, given that the Reputation factor is defined as residual variance unique to the rater sources that arises both (a) from errors in raters' perceptions (e.g., stereotypes) and (b) from information relevant to the leadership dimensions unavailable to the self (e.g., subtle facial or nonverbal responses that arise during meetings; McAbee & Connelly, 2016), how might this change as a function of time and as familiarity of the leader is enhanced? One might presume that as individuals within each rater group become increasingly familiar with a leader, the error in their perceptions as well as how they interpret the leader's nonverbal expressions might shift toward the leader's self-reported view of herself. Hence, familiarity could hold important implications in coaching contexts.

A third potential value of TRI/LARI models in leadership development is the creation and psychometric evaluation of new 360 assessments. For example, some assessment items might function quite differently when rated by a leader's subordinates than by other raters (e.g., bosses, peers). If so, then those items might need to be reevaluated, reweighted, reinterpreted, or even removed when designing or validating a 360 assessment. Such insights may become apparent only when the subordinates' ratings are modeled alongside ratings from those other raters, and TRI/LARI models provide a coherent, efficient, and informative method of doing so.

Research Implications

Although a LAM-based scoring procedure may offer little advantage over the more familiar SAM-based procedures, we believe TRI/LARI models continue to offer value to researchers. Indeed, we firmly believe that such models may be leveraged in valuable ways, including (a) decomposing variances in ratings attributable to different rater groups and (b) using 360 assessment ratings in a predictive context via a latent variable modeling framework. With respect to the first, one direction for research is to better understand how differences in variances emerge from (and thus reflect) fundamental issues of rater agreement or disagreement within and across raters. In other words, in what ways do the percentages of variance in ratings (from each rater group) relate to the amount of agreement between self and peers, or between different peers (but not the self), etc. Regarding the second, our current analysis is based on factor scores estimated (via the regression method) on the basis of parameters obtained from a fitted TRI/LARI model. It is possible that TRI/LARI factor information functions differently, in terms of associations with important outcomes, when all information and associations are derived within a single structural model.

The current work is the first to evaluate whether predictive information obtained from LAM scores differs from that obtained via standard SAM scores, and it reveals minimal differences. However, future research can explore this in greater depth, in at least two ways. First, does it depend on the specific competency or criterion? We examined two criteria (performance and derailment) and four competencies (Decisiveness, Compassion and Sensitivity, Quick Study, and Respect for Differences). While these competencies were selected based on the fact that competencies differ in their observability—especially to different rater groups—they may differ on other characteristics. Moreover, the degree to which these competencies are expressed (and ultimately observed) may depend on leaders' personality or evaluativeness of a given competency. For example, Compassion and Sensitivity may manifest more readily—across rater groups—for extraverted leaders than introverted leaders. The extent to which we can better understand the dynamics among competencies, leader personality, and ratings provided by different rater groups may help advance research surrounding 360 assessments. The same kind of dynamic likely exists for criterion variables. Are there other criteria—other than performance or derailment—that can enhance our understanding of the various factors that influence ratings?

Second, as noted earlier, our study indicated that the TRI/LARI Reputation factor might be more distinct from SAM scores than are the other TRI/LARI factors. Are there systematic reasons behind this, under what conditions is there meaningful variability in the Reputation factor, and what does that variability represent in a practical psychological sense? Our current efforts were based on inductive reasoning—that is, based on data. It may be worth taking a more deductive approach by stepping back and asking why the Reputation factor might capture variance distinct from SAM scores.

Limitations, Constraints on Generalizability, and Future Directions

Representing more than 120 countries from across the world, our sample was large and diverse in many ways. Along with the replication of the results across two randomly generated subsamples, this provides confidence in the generalizability of our findings. For reasons related to General Data Protection Regulations, demographic questions were optional; a large portion (42%) of the participants shared no information about their backgrounds. Of the data that were available, most (69%) participants were from the United States. Thus, our results may reflect that culture most strongly. However, it should be noted that CCL's Benchmarks for Managers data set (one tied to the present study's analyses) was thoroughly examined for item equivalence using a two-step procedure. First, translation comparisons (U.S. English, U.K. English, French Canada, French Europe, Dutch, German, Spanish, and Spanish Americas) were made to eliminate items that showed bias or favoritism toward one group over another. Second, item response theory-based differential functioning of items and tests framework (Raju et al., 1995) was used to identify and either revise or remove a total of 13 items (2.2%) from the item pool (Braddy, 2007). That said, for a thorough investigation into potential variability and viability of a TRI/LARI-based scoring method across cultures, future research may consider replicating our study with a more complete set of demographic data or utilize culture-specific data with results that can be compared and contrasted with our present study.

In this study, we focused on four diverse competencies and two criteria (performance and derailment); however, the utility of TRI/LARI-based scores might vary from one construct to another, based upon factors such as observability and desirability (Vazire, 2010). Future research should evaluate such possibilities by evaluating the utility of the TRI/LARI-based scoring of multirater assessments across an even wider range of constructs and criteria.

Lastly, in our data, information for the criterion variables was obtained from (some of) the same individuals who provided ratings of four separate competencies. That is, the direct reports, peers, and bosses provided both the ratings of the participants' perception of those four competencies and ratings of performance (and derailment). This raises two potential concerns. First, it may inflate the apparent predictive power of a given competency as reported by direct reports, peers, and bosses. When using ratings of single competencies to predict performance and derailment, effects were strong (e.g., $R^2 > .40$ for Decisiveness). These effects would likely be smaller for different types of criterion variables. Second, the fact that the criterion variables were based on peers', bosses', and direct reports' ratings might underrepresent the ability of self-rated competencies to predict outcomes, at least in comparison to those other sources of ratings. That said, the purpose of the present study was not to evaluate the predictive power of specific competencies in general, but rather to compare two scoring procedures—LAM versus SAM. Although the current assessment of each of the four competencies (e.g., Decisiveness) as predictive of performance may have biased findings toward certain predictors (i.e., those based on Peers, Direct Reports, or Bosses), it would not systematically bias findings in comparing the two scoring procedures. Nonetheless, future research should evaluate different criterion variables (e.g., more objective criteria such as promotion or firing).

Conclusions

Three-sixty assessments and related feedback play a central role in leadership development efforts. They provide leaders with valuable perspectives held by multiple rater groups, perspectives that can reveal information that leaders already know as well as those that they do not (i.e., blind spots). Applying a recently introduced methodology designed to disentangle shared perspectives from those

uniquely held by rater groups, the present study evaluated an alternative to the traditional method of scoring 360 assessments.

Our findings cast doubt on the value of LAM scores above and beyond what can be obtained through more traditional SAM scoring methods. While we found slight divergence in LAM and SAM scores, the sets of scores, for the most part, were highly similar to one another as indicated by their high correlations. This finding is important, given the growing attention and enthusiastic embrace that the TRI/LARI methodology has enjoyed. Current enthusiasm, notwithstanding, our results indicate that the sophistication of TRI/LARI models may not accrue to great benefit when scoring 360 assessments. Just as a 1970 Volkswagen Beetle can take a passenger from Point A to Point B in about as much time as that would be taken by a 2024 Tesla Model S, the more sophisticated method may not necessarily be superior to the traditional, tried-and-true method, at least when it comes to scoring and interpreting 360 assessments.

At the same time, our findings offer promising directions for future research and point to ways in which insights drawn from 360 assessments and the resulting feedback related to development can potentially be enriched through the use of TRI/LARI models. We hope this work encourages researchers to consider ways in which this novel scoring method can be used—both in and outside of leadership development—to further advance psychological measurement in both research and applied contexts.

References

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475–493. https://doi.org/10.1037/a0032475

Blickle, G., Schütte, N., & Wihler, A. (2018). Political will, work values, and objective career success: A novel approach—The Trait-Reputation-Identity Model. *Journal of Vocational Behavior*, *107*, 42–56. https://doi.org/10.1016/j.jvb.2018.03.002

Bono, J. E., Braddy, P. W., Liu, Y., Gilbert, E. K., Fleenor, J. W., Quast, L. N., & Center, B. A. (2017). Dropped on the way to the top: Gender and managerial derailment. *Personnel Psychology*, *70*(4), 729–768. https://doi.org/10.1111/peps.12184

Bracken, D., Timmreck, C. W., & Church, A. H. (Eds.). (2001). *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (1st ed.). Jossey-Bass.

Braddy, P. (2007). *A psychometric analysis of Benchmarks* [Unpublished research report]. Center for Creative Leadership.

Braddy, P., Gooty, J., Fleenor, J., & Yammarino, F. (2014). Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self–other agreement. *The Leadership Quarterly*, *25*(2), 373–390. https://doi.org/10.1016/j.leaqua.2013.10.001

Chappelow, C. T. (2004). 360-degree feedback. In C. D. McCauley & E. Van Velsor (Eds.), *The Center for Creative Leadership handbook of leadership development* (2nd ed., pp. 58–84). Jossey-Bass.

Church, A. H., Bracken, D. W., Fleenor, J. W., & Rose, D. S. (Eds.). (2019). *Handbook of strategic 360 feedback*. Oxford University Press. https://doi.org/10.1093/oso/9780190879860.001.0001

Connelly, B. S., McAbee, S. T., Oh, I.-S., Jung, Y., & Jung, C.-W. (2022). A multirater perspective on personality and performance: An empirical examination of the trait-reputation-identity model. *Journal of Applied Psychology*, *107*(8), 1352–1368. https://doi.org/10.1037/apl0000732

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*(1), 123–149. https://doi.org/10.1146/annurev-clinpsy-050212-185617

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *The Leadership Quarterly*, *7*(4), 487–506. https://doi.org/10.1016/S1048-9843(96)90003-X

Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. (2010). Self-other rating agreement in leadership: A review. *The Leadership Quarterly*, *21*(6), 1005–1034. https://doi.org/10.1016/j.leaqua.2010.10.006

Fleenor, J. W., Taylor, S., & Chappelow, C. (2020). *Leveraging the impact of 360-degree feedback* (2nd ed.). Barrett-Koehler.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Furr, R. M., Dougherty, D. M., Marsh, D. M., & Mathias, C. W. (2007). Personality judgment and personality pathology: Self-other agreement in adolescents with conduct disorder. *Journal of Personality*, *75*(3), 629–662. https://doi.org/10.1111/j.1467-6494.2007.00451.x

Helzer, E. G., Furr, R. M., Hawkins, A., Barranti, M., Blackie, L. E. R., & Fleeson, W. (2014). Agreement on the perception of moral character. *Personality and Social Psychology Bulletin*, *40*(12), 1698–1710. https://doi.org/10.1177/0146167214554957

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. https://doi.org/10.1111/j.1744-6570.2009.01164.x

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *The American Journal of Psychiatry*, *160*(9), 1566–1577. https://doi.org/10.1176/appi.ajp.160.9.1566

Leslie, J. B., & Braddy, P. W. (2015). *Benchmarks for managers technical manual*. Center for Creative Leadership.

Lindsey, E. H., Homes, V., & McCall, M. W., Jr. (1987). *Key events in executives' lives*. Center for Creative Leadership. https://doi.org/10.35613/ccl.1987.1089

Martel, M. M., Markon, K., & Smith, G. T. (2017). Research review: Multi-informant integration in child and adolescent psychopathology diagnosis. *The Journal of Child Psychology and Psychiatry*, *58*(2), 116–128. https://doi.org/10.1111/jcpp.12611

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*(5), 569–591. https://doi.org/10.1037/rev0000035

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353–368. https://doi.org/10.1177/014662169501900405

Rose, D. S., & Biringer, J. C. (2022). *Current practices in 360 feedback*. 3D Group.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Vazire, S. (2010). Who knows what about a person? The Self-Other Knowledge Asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Vergauwe, J., Hofmans, J., & Wille, B. (2022). The Leadership Arena-Reputation-Identity (LARI) model: Distinguishing shared and unique perspectives in multisource leadership ratings. *Journal of Applied Psychology*, *107*(12), 2243–2268. https://doi.org/10.1037/apl0001012

Wihler, A., Blickle, G., Ewen, C., Genau, H., Fritze, S., Völkl, L., Merkl, R., Missfeld, T., & Mützel, M. (2022). An integrative approach to more nuanced estimates of personality–job–performance relations. *Applied Psychology*, *72*(2), 588–624. https://doi.org/10.1111/apps.12391

Received August 2, 2023
Revision received July 22, 2024
Accepted July 31, 2024 ■