

Article

When Is Inference About the Relationship Between Self–Other Agreement and Leader–Member Exchange Warranted? Rating Discriminability as a Boundary Condition on Inference

Takashi Hongo ^{1,*}  and Dai Senoo ^{1,2}

¹ Department of Industrial Engineering and Economics, School of Engineering, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan; senoo.d.0967@m.isct.ac.jp

² Innovative-Life Society Visionary Initiative, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

* Correspondence: hongo.t.37e2@m.isct.ac.jp

Abstract

Response-surface analyses of leadership self–other agreement (SOA) and leader–member exchange (LMX) require three distinct questions to be kept apart: observed rating differentiation, measurement validity, and parameter-specific inferential support. We propose an applied diagnostic workflow built from established methods and illustrate it with 360-degree data from two Japanese organizations contrasting in rating differentiation (125 dyads; 39 supervisors). Additional analyses for Task Thinking (124 dyads) examined missing-data specifications, coefficient-specific information concentration, small-sample inference, and design-conditional recovery. Organization A’s negative incongruence-line curvature and the between-organization curvature contrast ($\Delta a_4 = -2.630$) remained supported across the examined specifications; the omnibus five-term interaction test was sensitive to missing-data handling. Under the specified generating models, Organization B’s observed design detected curvature matching Organization A’s estimate in 96.7% of replications and half-sized curvature in 45.0%; its distributional warning profile therefore did not imply low power for the larger curvature, and the causes of its nonsignificant estimate remain unresolved. The workflow separates descriptive warning signals from model-dependent estimation performance and from untested measurement explanations. It is proposed, not validated, and the directional asymmetry observed in Organization A is hypothesis-generating, because only eight dyads represented the self-overestimation side.

Keywords: rating discriminability; self–other agreement; leader–member exchange; response surface analysis; 360-degree feedback; performance appraisal; leniency bias; diagnostic workflow; cluster-robust inference



Received: 4 September 2026

Revised: 16 October 2026

Accepted: 19 October 2026

Published: 23 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Organizations read managerial qualities and development implications from 360-degree feedback data, and mean scores and rankings often ground promotion, placement, and development-investment decisions. Such interpretations rest on a premise that is rarely examined: that subordinates’ other-ratings exhibit systematic between-supervisor differentiation—rating discriminability. Where they do not, inferences that depend on differences between supervisors rest on weaker observational support. This premise concerns between-supervisor inference specifically, and, as this study shows, weak differentiation on distributional indicators does not by itself determine how much information a design holds for a particular surface

parameter. We treat discriminability as a boundary condition on inference—a condition under which inference from the observed ratings is better or worse supported, not an empirically established necessary or sufficient condition.

Self–other agreement (SOA) has accumulated for more than 30 years as an index of the accuracy of leaders’ self-awareness, captured through the gap between self-ratings and other-ratings (Atwater & Yammarino, 1992; Fleenor et al., 2010), and is now analyzed with polynomial regression and response surface analysis (RSA) rather than difference scores (Edwards & Parry, 1993; Edwards, 2002; Shanock et al., 2010). The measurement literature, meanwhile, holds that most variance in multisource ratings reflects rater idiosyncrasy rather than ratees (Scullen et al., 2000), that variance restriction distorts relationships among rating sources (LeBreton et al., 2003; Sackett et al., 2007), and that target variance is required for agreement to be measurable at all (Kenny, 1994; Allik et al., 2010). Some SOA–outcome studies use ICC(1) and rWG(j) to justify aggregation (Aarons et al., 2017; Nielsen et al., 2022), but whether the other-rating variance entering an SOA surface differentiates the supervisors about whom inference is made, and how weak differentiation bears on surface-parameter recovery, remain underarticulated. In organizations administering 360-degree feedback for the first time, leniency and ceiling concentration may be more likely (London & Smither, 2002), so that SOA may partly reflect unpracticed rating style—although, as shown below, distributional compression does not by itself determine a design’s power for a given parameter.

Direct tests of SOA on leadership behavior as a predictor of leader–member exchange (LMX; Graen & Uhl-Bien, 1995) within the RSA framework are, to our knowledge, limited to three studies: Barbuto et al. (2011, 2012) and Ertürk et al. (2018). None reports the surface-parameter estimates with uncertainty: Barbuto et al. (2012) terminated the analysis at a first-order gate requiring both predictors to be significant, without cluster-robust correction for 333 dyads nested in 83 leaders; Ertürk et al. (2018) stopped at on-axis classification and visual inspection across 36 organizations and report no rating-differentiation indicators (full audit in Table S15); the leader–follower congruence literature on the similarity of self-rated traits (e.g., Zhang et al., 2012; AlReshaid et al., 2025) concerns a different object than agreement between self- and other-ratings of the same supervisor. The pattern Barbuto et al. report—LMX variance carried by the other-rating alone—is not by itself diagnostic of discriminability: under one identical first-order model it is reproduced in one of our organizations and not in the other, and whether that difference reflects discriminability, a different relationship, or other sample features is the kind of question this study makes explicit rather than settles from the pattern alone.

We therefore ask three questions. RQ1: What should be examined, and in what order, before an RSA-based SOA–LMX surface is interpreted—in particular, what do observable indices (ceiling rate, variance, and the between-supervisor intraclass correlation) describe about subordinate ratings, and what can and cannot be judged from them about measurement validity and about the design’s support for specific surface parameters? RQ2: What premises are required on the data side and the procedure side for RSA-based inference, and how do description, interpretation, and estimation change under alternative preprocessing, centering, small-sample inference, and assumed curvature and error structures? RQ3 (exploratory): What organizational explanations—for example, first-time versus long-term operation of 360-degree feedback—can be generated for future testing from an observed between-organization contrast? We answer them by applying one analytical procedure to 360-degree and LMX data from two Japanese firms whose subordinate ratings differ greatly in observed differentiation.

Two disclosures frame this study. The hypotheses that were originally set—that SOA congruence would predict LMX in common across both organizations—were largely un-

ported, and the pooled model with an organization dummy could not be maintained once a formal test rejected surface homogeneity (Section 3.3); we report this transparently and position the study as hypothesis-generating through transparent post hoc analysis (thanking; Hollenbeck & Wright, 2017; Woo et al., 2017; Fisher & Aguinis, 2017), avoiding HARKing (Kerr, 1998). Second, the coefficient-level design-support diagnostics, the multiple-imputation sensitivity analyses, and the design-conditional simulation (Sections 4.2, 4.4 and 4.5) were added after peer review; the reframing of RQ1 and of the first contribution reflects what those analyses showed and was not prespecified.

The contributions are threefold and differ in evidential strength. The primary contribution is methodological: an applied diagnostic workflow, built from established methods rather than a new estimator, which keeps apart observed rating differentiation, measurement validity, and parameter-specific inferential support, and links each kind of evidence to the claims it can justify (Sections 2.1 and 5.1; Table 1). Its diagnostic performance and transferability are not validated here. The empirical illustration provides a second, narrower contribution: the two samples' Task Thinking surfaces differ under a formal interaction test, and the focal curvature contrast remained supported across the examined missing-data and small-sample specifications, whereas the omnibus five-term test was specification-sensitive (Section 4.4); a design-conditional simulation shows that the weaker-differentiation design had high power, under the specified generating models, for the curvature of the size estimated in the other sample (Section 4.5). Third, the fitted midpoint-referenced cross-section in Organization A suggested a local directional asymmetry; with eight dyads from six supervisors on the self-overestimation side, this remains hypothesis-generating rather than a general claim about the consequences of self-overrating. For administrative research the stakes are practical: leadership research on large Japanese firms, including in this journal, has relied on single-source employee surveys (Saruhashi et al., 2025), and the three indices can be computed by any human resources function from data it already holds, as a trigger for review rather than a certification of appraisal data.

Table 1. Diagnostic quantities used in this study: what each describes and what it cannot judge.

Quantity	Question	Describes	Does Not Judge	Reported in
Ceiling rate, variance, ICC(Z) of Z	1 Observed differentiation	Compression, spread, and between-supervisor share of subordinate ratings	Validity at any level; support for any parameter	Table 2
Discrepancy-direction coverage; distinct self-rating values	1/3	Observed support for each incongruence direction and for X	Detectability of a4 as such	Sections 4.1 and 4.2
Coefficient-specific concentration index, Satterthwaite df (CR2), leave-one-cluster-out	3 Inferential support	Concentration of estimator weights for a specific coefficient; dependence on single clusters	Measurement validity; whether a near-zero estimate is a true null	Section 4.2; Supplementary Analysis S4
Preprocessing/missing-data specifications	3	Sensitivity of estimates to handling rules	Which rule is correct	Section 4.4
Design-conditional rejection rate	3	Detectability of a specified curvature under a specified error structure, given the observed design	Probability that an effect exists; measurement validity	Section 4.5
Generating-model (target-signal) simulation	2 (illustration)	One candidate mechanism linking signal loss to recovery	Which process operated in a given sample	Section 4.6; Tables S6–S8

Note. Questions: 1 = observed differentiation; 2 = measurement validity; 3 = parameter-specific inferential support. None of the quantities alone establishes measurement validity.

Table 2. Descriptive differentiation indicators of subordinate ratings Z (all five dimensions).

Dimension	A Ceiling Rate	A Variance	A ICC(Z)	B Ceiling Rate	B Variance	B ICC(Z)
1. Task Thinking	11.7%	0.510	0.369	48.4%	0.401	0.206
2. Task Execution	18.0%	0.457	0.266	49.2%	0.383	0.229
3. Communication	31.1%	0.762	0.463	60.9%	0.334	0.206
4. Organizational Contribution	25.4%	0.732	0.269	50.0%	0.427	0.203
5. Mindset	47.5%	0.562	0.142	78.1%	0.150	0.198
Composite (five-dimension mean)	16.4%	0.485	0.322	48.4%	0.259	0.258

Note. The ceiling rate is the proportion of $Z \geq 4.5$; ICC(Z) is the intraclass correlation with supervisors as clusters (cluster-bootstrap 95% intervals in Table S10; means and skewness in Table S1). The indicators do not agree in direction everywhere—for Dimension 5 (Mindset), ICC(Z) is higher in Organization B while Organization A's ceiling rate is also high—so observed differentiation should not be judged by any single indicator; the description rests on their joint pattern.

2. Theoretical Framework

2.1. Three Questions: Observed Differentiation, Measurement Validity, and Inferential Support

Rating discriminability is rooted in the classical problems of rating quality—leniency, halo, and range restriction (Saal et al., 1980; Murphy & Cleveland, 1995)—and range restriction can severely distort observed correlations (Sackett et al., 2007). Multi-source ratings sharpen these problems: rater idiosyncrasy is the largest variance component (Scullen et al., 2000), and variance restriction distorts even the interpretation that sources hold distinct information (LeBreton et al., 2003). These concerns have been registered in SOA reviews (Fleenor et al., 2010) and appear as aggregation checks in some SOA–outcome studies (Aarons et al., 2017; Nielsen et al., 2022), but they have not been articulated as an explicit variance-source premise for interpreting dyad-level SOA response surfaces. In this study, rating discriminability refers to the systematic differentiation among supervisors observed in subordinate ratings. Measurement error and rater-related processes may alter associations involving quadratic and product terms (Busemeyer & Jones, 1983), and the variance structure of ratings shapes inference in related settings (Bono & Colbert, 2005); the present study does not infer those processes from restricted rating distributions. It examines observed differentiation and model-dependent estimation performance separately.

Three questions must therefore be kept apart. The first is the observed differentiation: whether the subordinate ratings, as distributed in the sample, concentrate near the ceiling, retain total variance, and separate supervisors (the between-supervisor variance share). The second is measurement validity: whether the ratings represent their intended referents at the relevant level. Because the focal outcome is dyad-specific LMX, within-supervisor variation in subordinate ratings may reflect meaningful dyadic experiences or unequal observation opportunities rather than error, and ICC(Z) neither identifies validity nor separates rater tendencies from genuine dyadic differences. The third is parameter-specific inferential support: how much information the observed configuration of self-ratings, subordinate ratings, and clusters provides for estimating and detecting a particular surface parameter, such as the curvature a_4 along the incongruence line. The observable indicators of the first question describe the sample; they do not establish the second, and—as Section 4.5 shows for one design—they do not by themselves determine the third. Table 1 lists the quantities used in this study, what each describes, and what it cannot judge; “weak information” is not used as a summary label for the three.

2.2. Self–Other Agreement, LMX, Response Surfaces, and Centering

SOA refers to agreement between a leader's self-rating and others' ratings, discussed through the over-estimator/in-agreement/under-estimator typology (Atwater et al., 1998); self-inflation has an experimentally documented mechanism in core self-evaluations and self-serving attribution (Cristofaro & Giardino, 2020). SOA moderates the validity of

leadership and performance predictions (Atwater & Yammarino, 1992), but its relationships with specific outcomes are context-dependent and inconsistent (Fleenor et al., 2010; Ostroff et al., 2004). LMX, the quality of the supervisor–subordinate dyadic relationship, is meta-analytically established as an antecedent of broad outcomes (Gerstner & Day, 1997; Dulebohn et al., 2012; Martin et al., 2016), and supervisor–subordinate agreement on LMX itself is only moderate (Sin et al., 2009). The combination therefore puts “whose rating counts as the substance” at issue on both sides; this study links the multisource predictor to relationship quality as perceived by the subordinate, and the discriminability of measurement has, to our knowledge, not been examined in this combination as a boundary condition on inference.

Congruence hypotheses should not be tested with difference scores alone, which conflate the independent contributions of self- and other-ratings (Edwards & Parry, 1993; Edwards, 2002). A response surface is estimated from the polynomial regression entering X , Z , their squares, and their interaction, and the slopes (a_1 , a_3) and curvatures (a_2 , a_4) along the congruence and incongruence lines are evaluated (Shanock et al., 2010); this approach has entered administrative research on leader–employee and person–environment fit (e.g., Sweis et al., 2022; Xie & Yang, 2025). Two points of procedure matter here. First, the first-order slopes a_1 and a_3 depend on the centering location, whereas a_2 and a_4 do not; an uncentered a_3 refers to $X = Z = 0$, a point that is unrealizable on a 1–5 scale, and the centering location must be stated explicitly (Yao et al., 2026). Second, a congruence effect cannot be established from a single significant parameter: broad-sense congruence requires $a_4 < 0$, $a_3 = 0$, and coincidence of the first principal axis with the congruence line (Humbert et al., 2019), and a vertex shifted along the incongruence line yields a directional asymmetry rather than a congruence effect. In these data, ceiling concentration co-occurred with sparse observations on the self-overestimation side, which is the distinction examined in the diagnostic workflow rather than a causal process we have established.

2.3. Scope of This Study

The study separates its claims by the strength of the evidence. The diagnostic proposition (Tier 1) is that the observed rating differentiation—the ceiling rate, variance, and between-supervisor variance share of subordinate ratings—can be described before analysis using the data organizations already hold, and that weak differentiation is a warning that the observed SOA may reflect properties of the rating process and that the observed support for incongruence may be narrow; it is not by itself a judgment about measurement validity or about the design’s power to detect a given surface parameter. This study illustrates the resulting sequence rather than testing the proposition, through a formal test of the between-sample surface difference, coefficient-level and missing-data diagnostics, a design-conditional simulation of curvature recovery, and a generating-model illustration of a candidate target-signal-loss mechanism; none of these attributes the observed difference to discriminability, and none validates the indicators’ diagnostic performance. Organizational explanations (Tier 2)—first-time administration, an immature feedback culture (London & Smither, 2002; Atwater et al., 2007), and appraisal purpose, which also breeds leniency (Murphy & Cleveland, 1995)—are candidate backgrounds for weak differentiation that organization $n = 2$ cannot adjudicate; the numerical indicators, the auxiliary text indicator, and the two simulations address different estimands and do not constitute converging evidence for any mechanism or cause. Practically, the descriptive step is executable on the organization’s side; the remaining steps of the workflow (Sections 5.1 and 5.3) require the analyst’s diagnostics.

3. Materials and Methods

3.1. Design, Sample, and Measures

We used 360-degree appraisal and LMX survey data from two Japanese private-sector organizations, not selected on the basis of rating discriminability; the contrast in rating differentiation was identified after the originally hypothesized common SOA–LMX surface was not supported. Organization A had operated 360-degree appraisal for more than 10 years and integrated it into its broader HR practices—institutional context was treated as background, not as a measured feedback-culture variable; Organization B implemented 360-degree appraisal for the first time. After matching and missing-data processing, the final analytic sample comprised 125 dyads (A: 61 dyads/17 supervisors, cluster sizes 2–6, median 4; B: 64 dyads/22 supervisors, cluster sizes 1–6, median 3; Table S11); for the focal Dimension 1 (Task Thinking), one Organization A construct-case missing all six items is excluded, so the analyses of Sections 4.2, 4.4 and 4.5 use 60 and 64 dyads (124), except the complete-case specification (111 dyads, 38 supervisors) and the secondary multiple-imputation specification (125). The study is cross-sectional and observational; the design permits formal comparison of the two observed response surfaces but not estimation of a general organization-level effect of implementation history, feedback culture, or rating discriminability. Under the confidentiality agreement, industry, size, and survey years are withheld, which constrains causal interpretation and generalizability (Section 5.4).

Supervisor self-rating X and subordinate rating Z were the mean of six items for each of five behavioral dimensions—Dimension 1 (Task Thinking), 2 (Task Execution), 3 (Communication), 4 (Organizational Contribution), and 5 (Mindset); representative item content is given in the note to Table S1. All items used a five-point (1–5) format with higher values more favorable; the midpoint (3) is the numeric middle of the format and carries no substantive anchor. Both organizations administered the same instrument family in three hierarchy-specific versions, with item wording adapted to level; within each dyad, both forms referred to the same supervisor and used the same hierarchy-specific item version; the sources differed in rating perspective (Table S11). Item-level equivalence across versions and organizations was not formally established (Section 5.4). LMX was measured with an adapted LMX-7-based seven-item relationship-quality scale (Graen & Uhl-Bien, 1995), translated into Japanese by the research team and administered identically in both organizations on a five-point agreement format; its measurement invariance was not independently validated. Internal-consistency reliabilities are shown in Table S12; some self-rating reliabilities are near 0.70, which may destabilize the self-rating-related quadratic and interaction terms. Descriptive statistics by organization and dimension are shown in Table S1.

3.2. Missing Data: Primary Rule and Five Reported Specifications

Missing data were confined to the subordinate 360-degree items: 143 of 3750 subordinate item responses (3.8%) were missing across 34 dyads. In the primary analysis, missing items were imputed by a prespecified within-dyad available-item mean rule following Newman (2014), supplemented by a prespecified adjacent-construct pooling rule for constructs missing three to five items; four construct-cases missing all six items were excluded listwise for the affected dimension (algorithm and counts in Supplementary Methods SM1). This is deterministic single imputation, so imputation uncertainty is not propagated into the primary-analysis standard errors. For the focal dimension, the handling was therefore further examined under four approaches, reported as five specifications: (i) the current rule; (ii) the same rule without adjacent-construct pooling, requiring at least three observed items (this changes one dyad); (iii) complete-case analysis excluding every dyad with any imputed Dimension 1 item (111 dyads, 38 supervisors); and (iv) multiple imputation

by fully conditional specification with predictive mean matching ($m = 20$; van Buuren, 2018) in a primary sensitivity specification that imputes only partially missing constructs (124 dyads) and a secondary specification that also restores the construct-case missing all six items (125 dyads). Approaches (i)–(iii) use the same null-imposed wild cluster bootstrap as the primary analysis. For (iv), scalar estimates were pooled by Rubin's (1987) rules with Barnard and Rubin (1999) degrees of freedom (complete-data $df = G - 1$), and the five-term interaction test with the multivariate D1 procedure and Reiter's (2007) degrees of freedom, and per-imputation wild-bootstrap p -values are reported as descriptive ranges, not pooled. The imputation model is not substantive-model-compatible and approximates the nested structure by a supervisor mean; these are sensitivity analyses, not replacements for the primary analysis. The reported imputation run was reproduced exactly and its pooled estimates were stable across chain lengths of 10, 20, and 30 cycles and an independent seed set (Supplementary Analysis S3). A $\pm 3SD$ outlier screen flagged seven values on six dyads, all retained as actual response values (Aguinis et al., 2013); an outlier-exclusion sensitivity analysis is shown in SM1.

3.3. Response Surface Analysis, Centering, and Inference Under Few Clusters

Following Edwards (2002), we estimated $LMX = b_0 + b_1X + b_2Z + b_3X^2 + b_4XZ + b_5Z^2 + e$ for each dimension and organization and derived $a_1 = b_1 + b_2$, $a_2 = b_3 + b_4 + b_5$, $a_3 = b_1 - b_2$, and $a_4 = b_3 - b_4 + b_5$. The main analysis used scale-midpoint centering ($c = 3.00$, identical for both organizations and all dimensions), adopted following Yao et al. (2026); a_1 and a_3 are therefore slopes at the midpoint, and organization-specific grand-mean centering ($A: c = 3.40$) is reported as a within-organization sensitivity analysis. A pooled model with an organization dummy as intercept was estimated first; because the organization $\times$ polynomial-term interaction model rejected surface homogeneity for Task Thinking (Section 4.4), its estimates were withdrawn and surfaces are reported by organization. Between-organization differences were tested in one model containing the five polynomial terms, an organization dummy ($A = 1$), and its products with all five terms, at the common centering; Δa_4 and Δa_3 are linear combinations of the interaction coefficients, and the five interaction terms are tested jointly. With up to 5 dimensions $\times$ 4 parameters examined, the study is exploratory: unadjusted estimates are reported with uncertainty intervals, sensitivity is assessed across inferential approaches, and no confirmatory pass–fail criterion is used (Section 5.4).

The data are nested, with multiple subordinates per supervisor and X being identical within supervisor (ICC of LMX 0.138 in A and 0.398 in B; 17 and 22 clusters). With few clusters, conventional cluster-robust standard errors are downward-biased (McNeish & Stapleton, 2016; Cameron & Miller, 2015)—a live constraint in administrative research, where a recent multi-team LMX study aggregated to the team level because ten clusters were too few (Bratek et al., 2026); aggregation would have discarded the dyadic variance our hypotheses concern, so we retain the dyad level and correct the inference. Null-imposed wild cluster bootstrap p -values ($B = 9999$; Rademacher weights primary, six-point Webb weights as a check; seed 777) constitute the primary small-cluster inference; cluster-bootstrap percentile intervals ($B = 4000$) quantify estimation uncertainty; analytic CR1 cluster-robust estimates, GEE, leave-one-cluster-out (LOCO), a random-intercept cross-level polynomial model (Nestler et al., 2019; Supplementary Analysis S2; random slopes for the X terms are not identified because X is constant within supervisor), and the alternative centering serve as sensitivity checks. Which reference distribution each reported p -value uses—normal-reference CR1 in the original analyses, $t(G - 1)$ for CR1 in the new diagnostics, Satterthwaite t for CR2, wild bootstrap, or multiple-imputation pooling—is stated in the corresponding table (Table S14). All analyses were implemented in Python

3.11.15 (statsmodels 0.14.6, NumPy 2.4.4, pandas 3.0.2, SciPy 1.17.1); the code with fixed seeds is available on request.

3.4. Coefficient-Level Design-Support Diagnostics

Because the wild cluster bootstrap improves the calibration of tests but does not increase the information the design contains, we added coefficient-level diagnostics for the focal dimension. Two distinct quantities are reported. The first is a coefficient-specific concentration index: for a coefficient or linear combination $c'b$, the OLS estimator weight vector of cluster g is $w_g = X_g(X'X)^{-1}c$, each cluster's share of the squared weight norm is $s_g = \|w_g\|^2 / \sum \|w_h\|^2$, and the index is the inverse Herfindahl index $1/\sum s_g^2$. It summarizes the concentration of linear-estimator weights under an identity working covariance; it is not an empirical variance share under arbitrary within-cluster dependence and not an effective sample size, and it is computed for a_4 and Δa_4 as linear combinations rather than by adding component shares. The second is the Bell and McCaffrey (2002) CR2 standard error with Satterthwaite degrees of freedom (Pustejovsky & Tipton, 2018) for a_3 , a_4 , and Δa_4 , which calibrates the parameter-specific small-sample test under the stated working model. LOCO ranges complete the set. Together, these answer how much effective information the observed design provides for a specific parameter—something neither the number of dyads nor the number of distinct self-rating values conveys (Section 4.2; formulas in Supplementary Analysis S4).

3.5. Descriptive Indicators of Rating Differentiation and Auxiliary Analyses

Observed differentiation of subordinate ratings was described by the variance, skewness, and ceiling rate of Z (the proportion with $Z \geq 4.5$, within half a scale point of the maximum—specified during analysis as a conceptually motivated threshold, not a validated cutoff), together with $ICC(Z)$, the share of total Z variance between supervisors from an unconditional random-intercept model (REML), with cluster-bootstrap intervals ($B = 2000$; Table S10). $ICC(Z)$ separates observed between-supervisor clustering from within-supervisor variation, which mixes dyad-specific relationships, rater-specific tendencies, and residual error; it does not establish that the between-supervisor component reflects true leadership quality. Systematic between-supervisor differentiation is the conceptual condition proposed here; the three indicators are provisional descriptive measures of observed differentiation, not an operationalization of target-related signal, and their diagnostic validity requires external evaluation. No composite score or pass–fail rule was applied; the indicators were examined jointly as related warning signals ($ICC(Z)$ is itself affected by total variance), and skewness, which was directionally inconsistent, is used descriptively only. Auxiliary analyses—broad-sense congruence criteria (Humberg et al., 2019; Table S4), the difference-score artifact analysis (Section 4.6), indirect checks against same-source explanations including an exploratory facet decomposition (Supplementary Discussion S1), and comment length as an organization-level text indicator (comments were optional in A and mandatory in B; Table S5)—are specified in SM1 and are exploratory.

3.6. Complementary Simulation Analyses

Two simulations serve different purposes. (a) Design-Conditional Simulation of Curvature Recovery: This holds fixed the observed self-ratings, subordinate ratings, and supervisor clusters of each organization for Task Thinking (A: 60 dyads, 17 supervisors; B: 64 dyads, 22 supervisors; midpoint centering) and asks how often each observed design recovers and detects a specified curvature along the incongruence line under a specified error structure. Taking Organization A's estimated coefficients as the base surface, the generating a_4 was set to -3.20 , -2.475 (A's estimate), -1.237 (half), and 0 by adding $(\delta/4)(X - Z)^2$, which changes b_3 , b_4 , and b_5 jointly and leaves a_1 , a_2 , and a_3 unchanged;

a generating a_4 of 0 removes the incongruence-line curvature but does not make the surface flat. Conditional means and generated outcomes are distinguished: for the A-sized condition, all conditional means lie within the 1–5 scale and 1.7% (A) and 5.1% (B) of generated outcomes fall outside it, whereas the A-anchored zero-curvature and half-curvature conditions place conditional means above the scale at part of the design (Table S22), so a within-scale zero-curvature surface anchored on Organization B's observed coefficients was added as a distinct calibration scenario. Outcomes were not truncated; a truncated variant is reported in Supplementary Analysis S5 only as a caution, because truncation itself induces curvature. Errors were $u_g + e_{ig}$, with variance components estimated from each organization's observed surface residuals (A: $\sigma^2 = 0.258$, $\rho = 0.00$; B: $\sigma^2 = 0.388$, $\rho = 0.35$). For each condition, $R = 1000$ replicates were drawn and the rejection rate of $H_0: a_4 = 0$ by the null-imposed wild cluster bootstrap (inner $B = 999$ for the conditions underlying the main conclusions, $B = 199$ otherwise) and by the CR1 $t(G - 1)$ test was recorded, with bias, RMSE, coverage of CR1 t intervals, and Monte Carlo standard errors (Morris et al., 2019); a between-organization version evaluated Δa_4 under a common surface and under A's curvature versus zero. The simulation is a post hoc evaluation of estimation performance under specified conditions; it does not estimate the probability that Organization B has a true effect, test whether discriminability caused the between-organization difference, validate the indicators, or add to the single self-overestimation-side dyad in Organization B (generating coefficients, error estimates, scale-range checks, seeds, and full performance tables in Supplementary Analysis S5).

(b) Generating-Model Illustration of Target-Signal Loss: A separate post hoc Monte Carlo study draws a latent supervisor quality T and a self-rating X , generates LMX as $4.0 - 0.8(X - T)^2 + \epsilon$ (generating $a_4 = -3.20$) or with the observed main effects restored, and lets the reported subordinate rating mix target-related signal with rater-style variance through a target-signal weight w_T , with an additive shift calibrating the ceiling rate. Each replicate comprises 64 independent dyads without clusters, estimated by OLS, so the design isolates a measurement-source mechanism and does not reproduce the nested design or yield calibrated probabilities for Organization B (equations, conditions, and seeds in the notes to Tables S6–S8 and Table S14).

3.7. Ethics

The 360-degree appraisal and the LMX survey were administered together, and both sets of responses were obtained for this research under a single procedure with the authorization of the participating organizations. Individual informed consent was obtained from participants at the time of data collection by written information: prospective participants received a notice stating the academic-research purpose of the study, the topic under investigation, and how the responses would be used and stored—for research purposes only, not for the identification or evaluation of individuals within the organization, managed under identification numbers, and not disclosed in individually identifiable form; responding to the survey after receiving that notice indicated consent, and no signed consent form was used. Participation was voluntary, and non-response carried no consequence for the individual. The notice did not address withdrawal of consent after submission. The dataset was de-identified before transfer to the authors, who had no access to direct identifiers; the organizations are not named, and potentially identifying attributes are withheld. The study was screened under the institution's ethics-review-necessity procedure and formally determined not to require review by its human research ethics committee—a determination about review necessity, not a finding that the research carries no risk of any kind (see the Institutional Review Board Statement). This study was conducted in accordance with the Declaration of Helsinki. Generative AI was not used in the study design, data

collection, analysis, or interpretation; its use was limited to translation of the authors' draft, as disclosed in the Acknowledgments.

4. Results

4.1. Descriptive Indicators of Rating Differentiation

Table 2 presents the ceiling rate, variance, and ICC(Z) of Z for all five dimensions; no validated diagnosis or pass–fail classification is implied, and the indicators' diagnostic performance is not evaluated in this study. In Organization B, subordinate ratings are compressed toward the top of the scale: the ceiling rate exceeds Organization A's in every dimension, the variance is smaller in every dimension, and the ICC(Z) point estimate is lower in four of five (ICC intervals are wide, e.g., Dimension 1: [0.00, 0.56] in A and [0.00, 0.57] in B; Table S10). Skewness was not consistent (Table S1). For the focal Dimension 1, the ceiling rate is 11.7% (A) versus 48.4% (B) and ICC(Z) 0.369 versus 0.206. The three indicators are related rather than independent, and Organization A shows dimension-specific warnings (Dimension 5). The joint pattern remained intact in all ten organization $\times$ dimension cells when all dyads with imputed items were excluded (Table S13). Discrepancy-direction coverage was markedly asymmetric: the self-rating exceeded the subordinate rating in 8 of 60 dyads in A (6 supervisors) but in 1 of 64 in B. As a post hoc description across the ten cells, in every cell meeting at least two of three sample-derived conditions (SM1), no surface parameter's 95% cluster-bootstrap interval excluded zero in the analyses that follow; this is a co-occurrence of two descriptions, not evidence of diagnostic performance—for Dimension 1 the same warning profile did not imply low power for the curvature of Organization A's size under the specified generating model (Section 4.5).

4.2. Centering and Coefficient-Level Information: Procedural Checks

A simple count of “significant parameters” is misleading: across specifications it changes from 6 vs. 0 (uncentered, uncorrected OLS) to 11 vs. 3 (midpoint-centered, uncorrected) to 4 vs. 0 (centered, supervisor cluster bootstrap), and uncentered counts include slopes at points outside the scale (Table S3). Centering and dependency correction are distinct operations: centering changes which quantities the coefficients represent, cluster correction changes the assessment of their uncertainty. For Task Thinking in Organization A, the uncentered $a_3 = +4.098$ —the slope at $X = Z = 0$ —could be misread as symmetry or as an opposite directional pattern, whereas both meaningful within-scale reference locations yield a negative directional slope: $a_3 = -2.055$ at the scale midpoint and -2.881 under pooled grand-mean centering ($c = 3.40$); the direction was consistent at the two reference levels, and method-specific uncertainty is reported in Table S18. The curvatures $a_2 = -0.595$ and $a_4 = -2.475$ are invariant to centering. Across the robustness battery (Table S18), a_3 and a_4 are supported consistently—cluster-bootstrap intervals exclude zero, wild-bootstrap p -values are, at most, 0.0033 under both weight distributions, and GEE and LOCO reproduce them—whereas a_1 is not supported and the approaches disagree on a_2 , which is therefore not treated as robust.

Coefficient-level design support shows that information concentration is coefficient-specific. In Organization A the concentration indices were 8.3 for a_4 and 8.3, 7.2, and 6.0 for the individual X^2 , XZ , and Z^2 coefficients, and 6.9 for a_3 ; the concentration is thus not confined to the self-rating direction, and these indices do not define a single effective sample size for the surface. The CR2 tests used separate Satterthwaite approximations: 6.8 degrees of freedom for a_4 and 5.2 for a_3 (nominal 16). Under these corrections $a_4 = -2.475$ remains supported (CR2 se = 0.481; $p = 0.0015$; 95% interval $[-3.62, -1.33]$), as does $a_3 = -2.055$ (CR2 se = 0.219; $p = 0.0002$; $[-2.61, -1.50]$), and LOCO estimates stay within $[-2.74, -2.13]$ for a_4 and $[-2.27, -1.92]$ for a_3 with no change in sign. For the between-organization

contrast, $\Delta a_4 = -2.630$ has CR2 se = 0.840 with 12.2 degrees of freedom ($p = 0.0085$; $[-4.46, -0.80]$) and a LOCO range of $[-3.05, -2.14]$. In Organization B, $a_4 = +0.155$ (CR2 se = 0.688, $df = 7.6$, $p = 0.83$; $[-1.45, 1.76]$) changes sign in 2 of 22 LOCO re-estimations and a_3 in 6 of 22, so near zero their direction is not determined. The concentration indices summarize how estimator weights are distributed across clusters; the Satterthwaite degrees of freedom calibrate parameter-specific small-sample tests under the stated working model. Neither is a count of supervisors or dyads, and Organization A's eight distinct self-rating values are not eight observations (Supplementary Analysis S4).

4.3. The Observed Surfaces: Directional Asymmetry in Organization A

Against Humberg et al.'s (2019) broad-sense congruence criteria (Table S4), condition (i) is satisfied ($a_4 = -2.475$, supported across approaches) and condition (ii) is rejected ($a_3 = -2.055$, negative across all estimation approaches); the first principal axis did not coincide with the congruence line, but its slope and the stationary-point classification were imprecise, so the rejection of broad-sense congruence rests on the nonzero a_3 and no clear ridge is claimed. The vertex of the one-dimensional cross-section along the midpoint-referenced incongruence line ($X + Z = 6$) lies at $u^* = -a_3/(2a_4) = -0.415$ ($u = (X - Z)/2$), where the subordinate rating exceeds the self-rating by about 0.83 points (Figures 1a and 2). This is the vertex of the cross-section, not the stationary point of the full surface, whose type is uncertain (a saddle in about a quarter of bootstrap replications). Re-expressing the fitted surface under common-constant centering changes neither its predictions at fixed original-scale coordinates nor u ; defining the cross-section by $X_c + Z_c = 0$, however, selects the original-scale line $X + Z = 2c$, so the vertex locations reported for $c = 3.00$ and $c = 3.40$ ($u^* = -0.42$ and -0.58 ; note to Table S4) refer to different parallel cross-sections of the same fitted surface, not to movement of a fixed vertex, and are not combined into a single uncertainty range or read as a general optimum. Model-implied contrasts show lower predicted LMX on the self-overestimation side at equal distances from perfect agreement, but only 8 of 60 dyads (13.3%, from six supervisors) lie on that side; prediction beyond the observed joint X–Z support is extrapolation, and the marginal range of u alone does not establish such support. What the data support is thus a local, model-implied directional asymmetry—a cross-sectional association, not a causal effect—rather than a congruence effect.

A negative a_3 alone does not establish a discrepancy-specific relationship, because it is a contrast of the two local slopes ($a_3 = b_1 - b_2$; the pattern reported by Barbuto et al., 2012, in which LMX tracks the subordinate rating alone, would also produce it). We therefore examined a separate first-order model and a joint test of self-rating terms as auxiliary interpretive checks, not prerequisites for fitting or interpreting every response surface. In Organization A's separate first-order model the self-rating carried an independent negative coefficient ($b_X = -0.625$, cluster-robust $p = 0.0001$) alongside a positive subordinate-rating coefficient ($b_Z = +0.988$), and the cluster-robust joint test of the three self-rating terms in the polynomial model was $F(3, 16) = 12.65$, $p < 0.001$; in Organization B the first-order self-rating coefficient was not statistically supported ($b_X = +0.027$, $p = 0.870$; Table S16). These checks are study-specific, not mathematical requirements for directional asymmetry; they do not eliminate suppression, measurement error, or other common-method explanations (Supplementary Discussion S1). Applying the same checks to the other four dimensions, only Dimension 1 shows a concave cross-section with a maximum ($a_4 < 0$), a shifted vertex ($a_3 < 0$), and an independent self-rating contribution; Dimension 2 supports only a directional association (Table S17). In Organization B, the estimated incongruence-line curvature was close to zero and imprecisely estimated (Section 4.2); in a separate first-order model, subordinate ratings were positively associated with LMX (Table S16), which is

distinct from the coefficient estimates of the polynomial model shown in Figure 1b. “Nearly planar” describes point estimates, not an established flat surface.

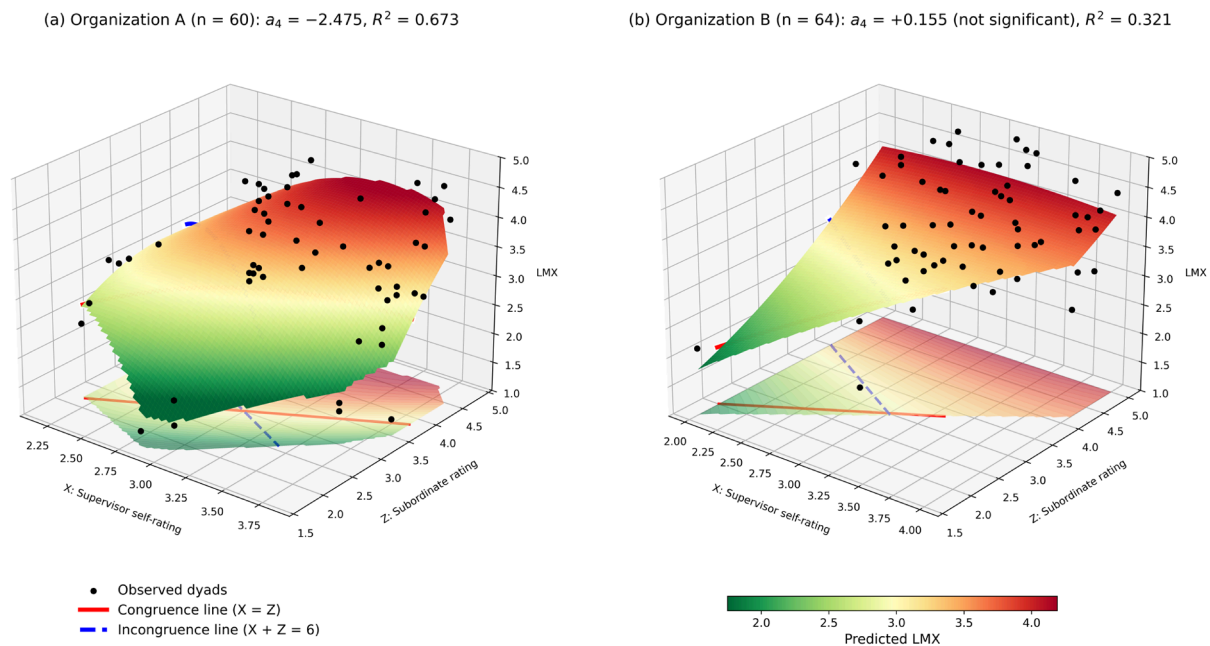


Figure 1. Response surfaces for Dimension 1 (Task Thinking) within the observed data region (floor: contour projection; red = congruence line, $X = Z$; blue = incongruence line, $X + Z = 6$ in the original metric; LMX = leader–member exchange). (a) Organization A (stronger differentiation indicators, $n = 60$): LMX declines at an accelerating rate along the incongruence direction, with the cross-section peaking where the subordinate rating exceeds the self-rating. (b) Organization B (multiple differentiation warning indicators, $n = 64$): under ratings compressed toward the ceiling, the estimated curvature along the incongruence line is close to zero and imprecisely estimated ($a_4 = +0.155$; leave-one-cluster-out range spanning zero), and the point-estimate surface is nearly planar.

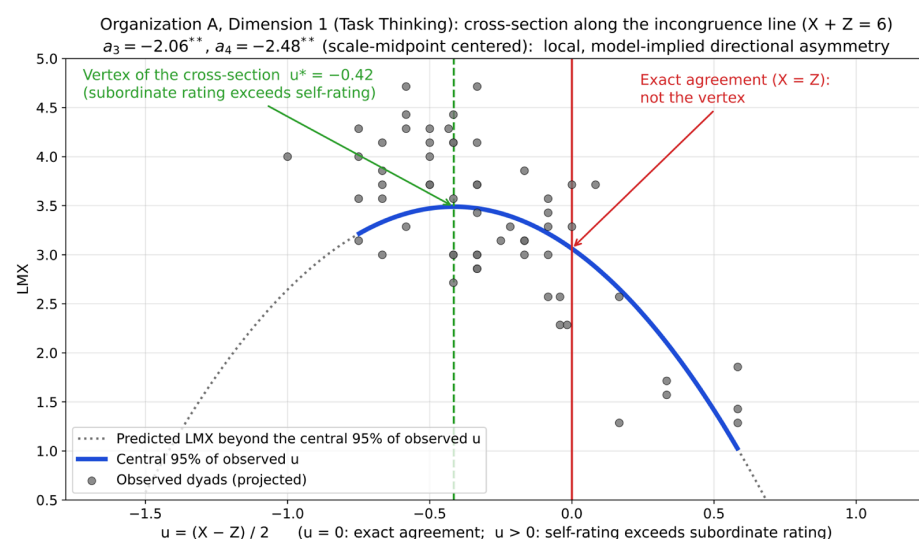


Figure 2. Organization A: one-dimensional cross-section of the fitted surface along the midpoint-referenced incongruence line ($X + Z = 6$ in the original metric). The vertex of this cross-section lies at $u^* = -0.42$ —the side where the subordinate rating exceeds the self-rating—and perfect agreement ($u = 0$, i.e., $X = Z = 3$) is not the vertex. Only 8 of the 60 dyads (from six supervisors) lie on the self-overestimation side; the dotted segment lies beyond the central 95% of observed u , a one-dimensional reference that does not by itself delimit the joint X – Z support (note to Table S4). Asterisks denote wild cluster bootstrap significance ($** p < 0.01$).

4.4. Between-Organization Contrast and Missing-Data Sensitivity

The interaction model provides a fixed-group comparison of the two organizational samples: it tests $\Delta a4$ directly and surface homogeneity as a joint test of the five interactions, not organization-level moderation in a population of organizations. For Dimension 1, $\Delta a4 = -2.630$ (wild cluster bootstrap $p = 0.006$; analytic cluster-robust $p = 0.0009$), the joint test is significant (wild $p = 0.023$; $F(5, 38) = 3.93$), and $\Delta a3 = -1.900$ is negative but not significant at the midpoint ($p = 0.123$): the surfaces differ, rejecting the homogeneity assumption of the pooled model. In the other four dimensions $\Delta a4$ is negative throughout but not significant; only for Dimension 5 are the joint test and the midpoint-referenced $\Delta a3$ (+1.723, analytic cluster-robust $p = 0.041$) significant, providing exploratory support for $\Delta a3$ but not $\Delta a4$ in that dimension (all dimensions in Table S26). What the test establishes is that the surfaces differ, not that discriminability produces the difference: with $n = 2$ organizations, discriminability is confounded with every other organizational attribute, and any attribution is a hypothesis for future testing.

The Task Thinking curvature contrast remained negative and statistically supported in all five reported missing-data specifications (Table 3), as did Organization A's $a4$ (−2.04 to −2.48). The omnibus five-term test was supported under the current rule, removal of adjacent pooling, and the primary multiple-imputation sensitivity specification (D1 $p = 0.0086$), but not under complete-case analysis ($p = 0.2207$) or the secondary specification that restores the construct-case missing all six items ($p = 0.1048$). The focal test concerns one feature of surface shape—the difference in incongruence-line curvature—whereas the omnibus test jointly assesses the five slope and curvature coefficient differences; a nonsignificant omnibus result neither establishes equivalence nor negates the focal contrast, and the focal contrast does not describe the complete pattern of coefficient differences. The comparisons do not isolate the effects of case inclusion (124 vs. 111 dyads, including one Organization A cluster), imputation, and testing procedures, and the primary multiple-imputation result is not treated as resolving the concern. Because the focal contrast was selected exploratorily, these are sensitivity results rather than independent confirmation.

Table 3. Sensitivity of the Task Thinking surface parameters to the handling of missing subordinate items (five reported specifications).

Specification	Dyads/Supervisors	Organization A $a4$	$\Delta a4$	Omnibus Five-Term Test
Current rule (available-item mean + adjacent pooling)	124/39	−2.475 (wild $p = 0.0033$)	−2.630 (wild $p = 0.0057$)	wild $p = 0.0233$
No adjacent pooling (≥ 3 observed items)	124/39	−2.430 (wild $p = 0.0037$)	−2.585 (wild $p = 0.0066$)	wild $p = 0.0226$
Complete case (all dyads with any imputed item excluded)	111/38	−2.043 (wild $p = 0.0183$)	−2.041 (wild $p = 0.0427$)	wild $p = 0.2207$
Multiple imputation, primary sensitivity specification (partially missing constructs only; $m = 20$)	124/39	−2.406 [−3.36, −1.46], $p < 0.0001$	−2.508 [−4.09, −0.93], $p = 0.0027$	D1 $F(5, 36.1) = 3.68$, $p = 0.0086$; per-imputation wild p 0.013–0.047
Multiple imputation, secondary specification (wholly missing construct-case restored)	125/39	−2.133 [−3.30, −0.97], $p = 0.0016$	−2.233 [−3.92, −0.54], $p = 0.0112$	D1 $F(5, 35.8) = 1.98$, $p = 0.1048$; per-imputation wild p 0.024–0.275

Note. Rows 1–3: null-imposed wild cluster bootstrap with Rademacher weights, $B = 9999$, seed 777 (the six-point Webb value for Organization A's $a4$ under the current rule is 0.0026; Table S18). Rows 4–5: Rubin's rules with Barnard–Rubin degrees of freedom for scalar estimates (95% intervals in brackets); D1 pooling with Reiter's (2007) degrees of freedom for the five-term test; per-imputation wild p -values ($B = 999$) are descriptive ranges, not pooled p -values. Organization B's $a4$ was +0.155, +0.155, −0.002, +0.103, and +0.100 in the five specifications (Table S19). Task Thinking only.

4.5. Design-Conditional Simulation of Curvature Recovery

Table 4 summarizes the design-conditional simulation for Task Thinking. With curvature set to Organization A's estimate ($a_4 = -2.475$), rejection rates were 84.9% for the A design and 96.7% for the B design (1000 replications; inner bootstrap $B = 999$). With half-sized curvature, the corresponding rates were 33.6% and 45.0% ($B = 199$). Under the two zero-curvature surfaces, within-organization rejection rates were 5.3–6.1%; the rate for the between-organization contrast under a common surface was 4.8%. CR1 t intervals covered the generating value in 91.1–92.4% of the rerun conditions. Table 4 distinguishes detection from null rejection, and Supplementary Analysis S5 reports Monte Carlo uncertainty, scale-range checks, and all conditions. Under these specified polynomial and error models, the B design did not have low power for an A-sized observed curvature despite its distributional warning profile. This evaluates a_4 at assumed effect sizes; it does not identify why the observed B estimate was nonsignificant, establish measurement validity, or add support on the sparsely observed self-overestimation side. Smaller effects and other generating processes remain unresolved, and the difference in rejection rates between the two designs reflects their complete design-and-error configurations, whose individual features were not isolated. Some generating conditions extend beyond the outcome scale, so these rates should not be treated as operating characteristics for all bounded LMX measurements.

Table 4. Design-conditional simulation: rejection rates of H_0 for a_4 and Δa_4 (Task Thinking).

Generating Value	Base/Null Surface	Organization A Design: Rejection (MCSE)	Organization B Design: Rejection (MCSE)	Inner B	Reading
$a_4 = -3.200$	A-anchored	95.5% (0.7)	99.2% (0.3)	199	Detection rate
$a_4 = -2.475$ (=Organization A estimate)	A-anchored	84.9% (1.1)	96.7% (0.6)	999	Detection rate
$a_4 = -1.237$ (half)	A-anchored	33.6% (1.5)	45.0% (1.6)	199	Detection rate
$a_4 = 0$	A-anchored	5.9% (0.7)	5.3% (0.7)	999	Type I error rate
$a_4 = 0$	B-anchored (conditional means within scale)	5.5% (0.7)	6.1% (0.8)	999	Type I error rate
$\Delta a_4 = 0$ (common surface, combined design)	—	Combined designs: 4.8% (0.7)		999	Type I error rate
$\Delta a_4 = -2.475$ (A curvature vs. zero, combined design)	—	Combined designs: 70.5% (1.4)		199	Detection rate

Note. $R = 1000$ replications per condition; null-imposed wild cluster bootstrap with Rademacher weights; MCSE = Monte Carlo standard error in percentage points, $\sqrt{[p(1-p)/R]}$. Errors $u_g + e_{ig}$ with $\sigma^2 = 0.258$, $\rho = 0.00$ (A) and $\sigma^2 = 0.388$, $\rho = 0.35$ (B). Under zero curvature the CR1 $t(G-1)$ test rejected in 7.9–8.9% (within) and 7.6% (between) of replications. The two Δa_4 rows use the combined two-organization design (124 dyads, 39 supervisors); their rates are placed in the Organization A column for layout only. Generating $a_4 = 0$ removes the incongruence-line curvature only. The A-anchored zero-curvature and half-curvature conditions place conditional means outside the 1–5 scale at part of the design (half-curvature: 1.7% of design points in A and 18.8% in B; zero: 13.3% and 42.2%), whereas the A-sized condition does not; see Table S22. Rates are conditional on the specified surfaces, error structures, and observed configuration; they are not probabilities that a true effect exists in Organization B. Bias, RMSE, coverage, and a truncated-outcome caution are in Supplementary Analysis S5.

4.6. Auxiliary Analyses

Four auxiliary analyses are summarized here and reported in full in the Supplementary Materials. First, the composite-SOA difference score correlates with LMX at -0.371 in Organization B, yet the partial correlation controlling for Z is $+0.050$; in Organization A it attenuates from -0.764 to -0.223 (Table S9): difference scores inherit the strong contribution

of $-Z$, so their association with LMX does not by itself establish a discrepancy effect, which is the rationale on which Edwards and Parry (1993) rejected the difference scores. Second, improvement comments were longer in Organization A whereas comments on strengths did not differ (Table S5); the differing response regimes make this an auxiliary observation that provides no independent validation, and none of the differentiation indicators alone establishes measurement validity or determines parameter-specific inferential support. Third, common method variance (Podsakoff et al., 2003) cannot be ruled out: the directional-asymmetry claim rests on the cross-source coefficient b_X , the nonlinear term a_4 is not a direct consequence of a uniform additive same-source factor, and the effects are dimension-specific in limited respects, but without supervisor-rated LMX or objective indicators CMV remains unresolved (Supplementary Discussion S1). Fourth, the generating-model simulation (Section 3.6, part (b)) illustrates one candidate measurement mechanism: ceiling truncation alone left a_4 strongly present, whereas estimates deteriorated toward zero with indeterminate sign as rater style replaced target-related signal, and the estimator was biased negative even at $w_T = 1.00$ (Tables S6–S8; Figure S1). This illustration includes no true-null condition, is not calibrated to Organization B, and cannot explain its nonsignificant estimate; the result of Section 4.5 does not imply, by elimination, that target-signal loss operated there.

5. Discussion

5.1. Methodological Contribution: An Integrated Diagnostic Workflow

The primary contribution is an applied diagnostic workflow that links observed evidence to the scope of RSA-based SOA–LMX claims. It separates rating differentiation, measurement validity, and parameter-specific inferential support (Table 1). The two samples illustrate why these questions require different evidence: nominal sample size does not describe coefficient-level information concentration (Section 4.2); sensitivity can differ between a focal contrast and an omnibus test (Section 4.4); and a distributional warning profile need not imply low power for a specified curvature (Section 4.5). The workflow integrates established methods rather than introducing a new estimator, and its diagnostic performance and transferability remain to be evaluated externally. Within it, the procedural elements demonstrated earlier play specific roles: explicit centering determines where first-order parameters are evaluated and therefore what a directional claim means (Yao et al., 2026), and we propose disclosure of the centering location as a reporting standard; cluster-robust inference with small-sample corrections, LOCO, and coefficient-specific concentration indices serve the assessment of parameter-specific support, under which several coefficients significant under uncorrected OLS no longer received consistent support; the organization $\times$ polynomial-term interaction model provides the formal test of a fixed-group difference between two observed surfaces; and, unlike a main-effect gate that withholds the polynomial analysis itself (Barbuto et al., 2012), the workflow does not use the significance of lower-order terms as a gate for fitting the polynomial model—it presupposes commensurable measurements and an estimable design and restricts interpretation according to the diagnostics.

The simulation illustrates why distributional warnings and parameter-specific inferential support require separate evidence. In this design, the warning profile did not imply low power for the larger curvature under the specified generating model. Conversely, conditional detectability does not establish that ratings capture their intended referents or that a sparsely observed discrepancy direction is well represented. The workflow therefore treats descriptive indicators as prompts for further review, not as diagnoses of invalid measurement or explanations of nonsignificant estimates. The study recasts SOA research's implicit measurement assumption as a boundary condition on inference to be examined

in each of the three questions, and the answers to RQ1 and RQ2 follow from this: the indices describe observed differentiation and distinguished the two samples, but what they do not establish—validity and parameter-specific support—must be examined separately; and inference from rating data requires procedural premises independent of discriminability (centering, few-cluster inference, specification sensitivity) whose neglect changes the conclusions drawn from the same data.

5.2. Empirical Implications and Alternative Explanations

The inconsistency of SOA–outcome relationships (Fleenor et al., 2010) recurs across the three direct tests, and whether variation in rating differentiation contributes to it is an untested hypothesis that this study makes examinable rather than a finding: Barbuto et al.’s first-order pattern was reproduced in Organization B but not in Organization A under one identical procedure, and that difference is not itself evidence of low discriminability. Ertürk et al. (2018) report no differentiation indicators, so their sample cannot be located on that axis—which illustrates the proposed reporting standard. For the 360-degree feedback literature, the study adds measurement discriminability as a consideration to the finding that feedback effectiveness depends on design and context (Kluger & DeNisi, 1996; DeNisi & Kluger, 2000; Smither et al., 2005); it contributes at the measurement and interpretation stages and does not test whether feedback changes behavior. The directional asymmetry in Organization A qualifies the implication, carried by the over/under/agreement typology, that exact agreement is uniformly optimal: a quadratic function of the signed difference can represent the fixed-mean cross-section, but a difference-only model cannot, in general, represent the unrestricted two-predictor surface, and the contribution of polynomial regression here is to characterize the fitted surface at an explicitly specified reference level, not to make an otherwise impossible shape representable. The pattern is cross-sectional, concerns one of five dimensions, and rests on eight dyads from six supervisors.

For the non-detection in Organization B, the explanations are non-exclusive and the design cannot adjudicate among them. Organization B may have a smaller, a genuinely different, or no curvature along the incongruence line; the conjecture that its design necessarily had low power for curvature of Organization A’s size is not supported under the specified models, and the simulations did not identify the cause of the nonsignificant estimate—a smaller true curvature, a substantively different relationship, differences in the measurement process, a misspecified error or functional form, and sampling variation remain possible. Rating compression and loss of target-related signal remain a candidate account, consistent with the warning profile and the generating-model illustration but not an identified cause and not implied by elimination from the design-conditional result; genuinely homogeneous supervisors, genuinely dyad-specific behavior, unequal observation opportunities, and instrument differences without established invariance bear on the organizational hypotheses. Common method variance and suppression are narrowed but not excluded (Supplementary Discussion S1), and unmeasured power distance, feedback culture, anonymity arrangements, appraisal purpose, and rater training could differ between organizations, although follower perceptions of leadership behavior were largely invariant across cultural clusters in a 77-country comparison (Caza et al., 2021). Independently of explanation, self-rating exceeds subordinate rating in 8 of 60 dyads in A but 1 of 64 in B, so in B the incongruence-line parameters are estimated from observations predominantly on one side—a limitation of discrepancy coverage that is distinct from the information available for a4 as such.

5.3. Practical Implications

Organizations can use the three indicators as triggers for measurement review before commissioning SOA analysis; the indicators do not certify appraisal data for personnel decisions, do not by themselves show that an analysis would be underpowered, and do not justify excluding a dataset. Table 5 sets out what an analyst would do at each step of the workflow, what is recorded, and what each step cannot settle. For example, a near-ceiling rate of 50% and an ICC(Z) of 0.20 on a dimension would initiate an item- and protocol-level review rather than automatic data exclusion: the analyst would examine whether ceiling concentration is widespread or item-specific, whether raters had opportunities to observe the behaviors, and whether the confidentiality and appraisal-purpose instructions could have influenced responses; would separately inspect the joint X–Z support and the coefficient-specific diagnostics before deciding how narrowly the fitted surface can be interpreted; and would record which conclusions survive alternative specifications. No step aims to increase variance artificially or to make a result significant. Multi-source designs in administrative settings often involve few ratees and many ratings per ratee (e.g., Maurin Saldaña et al., 2026), precisely the configuration in which rater-side variance can masquerade as information about ratees. Rater training in behavioral observation is a candidate intervention for future experimental evaluation, not a validated remedy; because “first-time administration impairs discriminability” remains a candidate hypothesis, no prescription based on years of operation follows; and the exploratory asymmetry motivates replication with better coverage of both discrepancy directions—it does not support a prescription to increase or decrease managers’ self-ratings. The sample-derived anchors used in Section 4.1 (two or more of: ceiling rate above 40%, variance below 0.45, ICC(Z) below 0.25) are post hoc illustrative values from ten cells, not validated thresholds (SM1).

Table 5. The diagnostic workflow as actions, records, and limits.

Step	What Is Done	What Is Recorded/Decided	What It Does Not Settle
1 Inferential target and scales	Specify what is compared as the same referent; check self and subordinate scales, item versions, and timing	The range of comparable interpretation	Model adequacy (commensurability alone does not establish it)
2 Distribution and observed support	Ceiling rate by item and dimension, supervisor-level distributions, ICC(Z) with its interval, joint X–Z distribution and discrepancy-direction coverage	Where compression and sparse regions lie	Validity; power. No automatic exclusion, no manipulation of ratings to increase variance
3 Measurement-process review (if flagged)	Inspect instructions, anonymity, appraisal purpose, observation opportunity, item distinctiveness	A record distinguishing genuinely similar behavior, limited observation, and response pressure; items or “unable to observe” options to pilot in the next administration	Whether the surface is estimable
4 Estimation and sensitivity	Common centering; cluster-robust inference with small-sample corrections; coefficient-level concentration diagnostics; leave-one-cluster-out; missing-data specifications	Estimates and intervals; which conclusions survive alternative specifications	Causes of a null
5 Conditional performance (optional)	Simulate recovery under specified effect sizes and error structures with the observed design fixed, stating the evaluation point	Detection and coverage rates under those conditions	The cause of a nonsignificant result; the probability that an effect exists
6 Interpretation and further collection	Limit claims to the supported region and specifications; identify missing information	Independent measures, observation opportunities, data in sparse regions as next steps	— (does not aim at improved distributions or significance)

Note. A proposal derived from two samples; no intervention named here has been evaluated in this study.

5.4. Limitations and Future Research

The study is cross-sectional and a two-organization comparison; causal inference is not possible, and the attribution of the between-organization difference to discriminability cannot escape the organization $n = 2$ confounding (industry, size, and rating-tool design included). The boundary-condition proposition and the workflow were generated post hoc, and the coefficient-level diagnostics, multiple-imputation analyses, and design-conditional simulation were added after peer review. The number of clusters is at most 39, a level at which sandwich estimators can be biased (McNeish & Stapleton, 2016); only one of five dimensions is consistently supported; X is constant within supervisor, so only eight distinct self-rating values underlie Organization A's 60 dyads (11 for B's 64), and the coefficient-level diagnostics quantify the resulting concentration for each parameter separately without defining an effective sample size for the surface. Overestimation-side dyads number 8 of 60 from six supervisors, and 17/17 LOCO significance is weaker evidence for that side than it appears. With up to 20 surface parameters and a focal contrast selected exploratorily, multiplicity is a limit—illustratively, the Task Thinking a4 (Webb-weight wild $p = 0.0026$) would be borderline under a Bonferroni-style correction for the 20-coefficient family (0.052)—and the sensitivity analyses are not independent replications. The stationary point is a saddle in about a quarter of bootstrap replications, so “ridge” remains an approximation, and the negative self-rating main effect could reflect a supervisor-side third variable. Measurement invariance of the adapted LMX scale and item-level equivalence of the instrument versions were not validated; ICC(Z) is estimated with wide uncertainty from few clusters; CMV is unresolved; the primary missing-data handling is a single imputation, and the multiple-imputation sensitivity analyses use an imputation model that is not substantive-model-compatible, approximates the nested structure by a supervisor mean, and was applied to one dimension only. The omnibus five-term difference was not supported under the complete-case and secondary multiple-imputation specifications (Table 3).

The two simulations bound generalizability in different ways. The design-conditional simulation used one error structure per organization, held the observed ratings fixed, examined four curvature levels for one dimension, and evaluated the detectability of a4 given the observed design, not the full preprocessing-to-inference pipeline, measurement validity, or coverage of the self-overestimation direction; some generating conditions place conditional means outside the scale. The generating-model simulation used independent dyads without the empirical cluster structure and includes no true-null condition. Text evidence is limited to comment length, cultural variables such as power distance were not measured, the sample-derived anchors were not externally validated, and participant sex and gender were not available to the authors. Future research follows from these limits. The most natural extension is a two-study design using Organization B's second administration as a validation sample—a path toward preregistration-style testing of this study's post hoc proposition (Hollenbeck & Wright, 2017; Woo et al., 2017). External validation of the indicators should compute them before outcome analysis in many organizations and dimensions, with preregistered rules and out-of-sample evaluation of interval width, sign stability, parameter recovery, and false-warning rates; this study offers no diagnostic validation and one conditional illustration in which a warning profile did not imply low power for a specified curvature. The design-conditional simulation should be extended to varying error structures, measurement reliability, dyad-specific variance, ceiling and rater-style processes, and other dimensions, and combined with the generating-model design in one clustered framework. Replication of the Task Thinking pattern calls for sampling that balances discrepancy directions, a larger supervisor sample, supervisor-rated LMX and external outcomes, and preregistration; rater-side interventions—training

in behavioral observation, anonymity arrangements, rating purpose, and behaviorally anchored items—should be evaluated in randomized or stepped-wedge designs.

6. Conclusions

This study proposes an applied workflow for evaluating RSA-based SOA–LMX inference by separating observed rating differentiation, measurement validity, and parameter-specific inferential support. Its application to two organizational samples showed that the focal Task Thinking curvature contrast remained supported across the examined missing-data and small-sample specifications, whereas the omnibus interaction test was specification-sensitive. Under the specified simulation models, a distributional warning profile did not imply low power for an A-sized observed curvature. This result neither identifies the cause of Organization B’s nonsignificant estimate nor validates its measurements. The local asymmetry observed in Organization A remains exploratory because support on the self-overestimation side was sparse. The methodological contribution is the explicit connection between diagnostic evidence and the scope of claims it supports. External evaluation is needed before the proposed workflow or its indicators can be treated as general diagnostic rules.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/admsci16110522/s1>, Figure S1: Sweep of the target-signal weight w_T (1,000 replications each, ceiling rate fixed at 69–79%); Table S1: Descriptive statistics of the main variables (by organization and dimension); Table S2: Full polynomial regression coefficients (scale-midpoint centered, $c = 3.00$) by organization and dimension; Table S3: Descriptive summary of RSA results by organization; Table S4: Satisfaction of the broad-sense congruence criteria of [Humbert et al. \(2019\)](#) (Organization A, Dimension 1 (Task Thinking)); Table S5: Auxiliary text-based indicators from subordinate comments; Table S6: Simulation: Estimated a_4 under a true $a_4 < 0$ (500 replications each); Table S7: Sweep of the target-signal weight w_T (ceiling rate fixed at 69–79%, 1000 replications each); Table S8: Sweep with a generating model including main effects (1000 replications each); Table S9: Correlations among key variables (composite SOA); Table S10: Cluster-bootstrap 95% confidence intervals for ICC(Z) ($B = 2000$); Table S11: Sample composition and instrument versions; Table S12: Internal-consistency reliability (Cronbach’s α) by source, organization, and dimension; Table S13: Imputation-exclusion sensitivity of the screening indicators (full sample $\rightarrow$ dyads with imputed items excluded); Table S14: Reproducibility specification of the inferential procedures; Table S15: Comparison of the three direct SOA–LMX studies with the present study; Table S16: Independent contribution of the self-rating (Dimension 1 (Task Thinking), by organization); Table S17: Correspondence between the robustness of a_3 and the independence of the self-rating (Organization A, all five dimensions); Table S18: Surface parameters of Dimension 1 (Task Thinking): centering location and robustness across multiple inferential approaches (Organization A, $n = 60$, 17 supervisors); Table S19: Missing-data sensitivity of the Dimension 1 (Task Thinking) surface parameters: five reported specifications and six stability-check runs; Table S20: Coefficient-level design-support diagnostics, Dimension 1 (Task Thinking); Table S21: Generating polynomial coefficients by condition (scale-midpoint centering); Table S22: Scale-range check of the generating surfaces: conditional means and generated outcomes; Table S23: Full performance results of the design-conditional simulation ($R = 1000$ replications per row); Table S24: Auxiliary truncated-outcome variant: rejection rates only (caution, not a type I error or power evaluation); Table S25: Scenario–seed map of the design-conditional simulation (all runs; seeds from the execution scripts); Table S26: Between-organization differences in the response surface via organization $\times$ polynomial-term interactions, all five dimensions (cluster-robust); Table S27: Evidential status of the main claims (Task Thinking unless stated); Supplementary Methods SM1: Implementation Details of Preprocessing, Robust Inference, and Auxiliary Analyses; Supplementary Discussion S1: Indirect Checks Against Simple Same-Source Explanations (full exposition of Section 4.6); Supplementary Analysis S2: Random-Intercept Cross-Level Polynomial Model and Cluster-Bootstrap Confidence Intervals

for ICC(Z); Supplementary Analysis S3: Imputation-Exclusion and Multiple-Imputation Sensitivity of the Focal Surfaces and Organization Contrast; Supplementary Analysis S4: Coefficient-Level Design-Support Diagnostics for Dimension 1 (added after peer review); Supplementary Analysis S5: Design-Conditional Simulation of Curvature Recovery (added after peer review). Analysis scripts with fixed seeds are available as stated in the Data Availability Statement.

Author Contributions: Conceptualization, T.H. and D.S.; Methodology, T.H.; Formal analysis, T.H.; Investigation, T.H.; Data curation, T.H.; Writing—original draft, T.H.; Writing—review and editing, D.S.; Supervision, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study as this study was reviewed under the research-ethics screening procedure of the authors' institution and was formally determined not to require review by its human research ethics committee (the declaration of non-applicability was submitted by the principal investigator and its acceptance has been confirmed). The determination followed the institution's ten-item ethics-review-necessity checklist, under which the study is not "life science and medical research" within the meaning of the Japanese national Ethical Guidelines for Life Science and Medical Research Involving Human Subjects, issued jointly by the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labour and Welfare, and the Ministry of Economy, Trade and Industry in 2021 and amended in 2022 and 2023, and involves neither invasive procedures, intervention, or biological specimens, nor sensitive personal information. The study is a non-interventional, observational secondary analysis of de-identified 360-degree appraisal and survey data collected by the participating organizations for their own personnel-development purposes; the data were provided to the authors with the consent of the participating organizations and participants, and no individually identifiable information was accessible to the authors.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author, subject to approval by the participating organizations, due to organizational confidentiality obligations. All analysis scripts (with fixed random seeds) are likewise available on request from the corresponding author.

Acknowledgments: The authors gratefully acknowledge financial support for the article processing charge (APC) from 'DLab+ Challenge 2026', DLab+, Institute of Future Science, Institute of Science Tokyo. During the preparation of this manuscript, the authors used Claude (Anthropic; 2026 version) for the purpose of translating the manuscript from the authors' original-language draft into English. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Aarons, G. A., Ehrhart, M. G., Farahnak, L. R., Sklar, M., & Horowitz, J. (2017). Discrepancies in leader and follower ratings of transformational leadership: Relationship with organizational culture in mental health. *Administration and Policy in Mental Health and Mental Health Services Research*, 44(4), 480–491. [\[CrossRef\]](#) [\[PubMed\]](#)
- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2), 270–301. [\[CrossRef\]](#)
- Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., & Metspalu, A. (2010). Variance determines self-observer agreement on the Big Five personality traits. *Journal of Research in Personality*, 44(4), 421–426. [\[CrossRef\]](#)
- AlReshaideh, F., Ertürk, A., Alkhayat, R., Abdallah, F., Abidi, O., & de la Roche, M. (2025). Are we on the same wavelength? Supervisor-subordinate cognitive style congruence and its association with supervisors' self-awareness through leader member exchange. *Frontiers in Psychology*, 16, 1583837. [\[CrossRef\]](#) [\[PubMed\]](#)

- Atwater, L. E., Brett, J. F., & Charles, A. C. (2007). Multisource feedback: Lessons learned and implications for practice. *Human Resource Management*, 46(2), 285–307. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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(1), 141–164. [\[CrossRef\]](#)
- Barbuto, J. E., Jr., Singh, M., Wilmot, M. P., & Story, J. S. P. (2012). Self-other rating agreement and leader-member exchange (LMX): A quasi-replication. *Perceptual and Motor Skills*, 114(2), 479–484. [\[CrossRef\]](#) [\[PubMed\]](#)
- Barbuto, J. E., Jr., Wilmot, M. P., & Story, J. S. P. (2011). Self-other agreement and leader-member exchange (LMX). *Perceptual and Motor Skills*, 113(3), 875–880. [\[CrossRef\]](#) [\[PubMed\]](#)
- Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948–955. [\[CrossRef\]](#)
- Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2), 169–181.
- Bono, J. E., & Colbert, A. E. (2005). Understanding responses to multisource feedback: The role of core self-evaluations. *Personnel Psychology*, 58(1), 171–203. [\[CrossRef\]](#)
- Bratek, M., Hruška, D., & Sertić, A. (2026). The relationship between leader-member exchange and team cohesion: A multi-team study in the Croatian banking sector. *Administrative Sciences*, 16(1), 23. [\[CrossRef\]](#)
- Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93(3), 549–562. [\[CrossRef\]](#)
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. [\[CrossRef\]](#)
- Caza, A., Caza, B. B., & Posner, B. Z. (2021). Transformational leadership across cultures: Follower perception and satisfaction. *Administrative Sciences*, 11(1), 32. [\[CrossRef\]](#)
- Cristofaro, M., & Giardino, P. L. (2020). Core self-evaluations, self-leadership, and the self-serving bias in managerial decision making: A laboratory experiment. *Administrative Sciences*, 10(3), 64. [\[CrossRef\]](#)
- DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *Academy of Management Perspectives*, 14(1), 129–139. [\[CrossRef\]](#)
- Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., & Ferris, G. R. (2012). A meta-analysis of antecedents and consequences of leader-member exchange: Integrating the past with an eye toward the future. *Journal of Management*, 38(6), 1715–1759. [\[CrossRef\]](#)
- Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow, & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 350–400). Jossey-Bass.
- Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, 36(6), 1577–1613. [\[CrossRef\]](#)
- Ertürk, A., Van den Broeck, H., & Verbrigghe, J. (2018). Self-other agreement on transformational leadership and subordinates' assessment of supervisor's performance: Mediating role of leader-member exchange. *Leadership & Organization Development Journal*, 39(2), 291–308. [\[CrossRef\]](#)
- Fisher, G., & Aguinis, H. (2017). Using theory elaboration to make theoretical advancements. *Organizational Research Methods*, 20(3), 438–464. [\[CrossRef\]](#)
- 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(6), 1005–1034. [\[CrossRef\]](#)
- Gerstner, C. R., & Day, D. V. (1997). Meta-analytic review of leader-member exchange theory: Correlates and construct issues. *Journal of Applied Psychology*, 82(6), 827–844. [\[CrossRef\]](#)
- Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly*, 6(2), 219–247. [\[CrossRef\]](#)
- Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, 43(1), 5–18. [\[CrossRef\]](#)
- Humberg, S., Nestler, S., & Back, M. D. (2019). Response surface analysis in personality and social psychology: Checklist and clarifications for the case of congruence hypotheses. *Social Psychological and Personality Science*, 10(3), 409–419. [\[CrossRef\]](#)
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford Press.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. [\[CrossRef\]](#)

- LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6(1), 80–128. [\[CrossRef\]](#)
- London, M., & Smither, J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review*, 12(1), 81–100. [\[CrossRef\]](#)
- Martin, R., Guillaume, Y., Thomas, G., Lee, A., & Epitropaki, O. (2016). Leader-member exchange (LMX) and performance: A meta-analytic review. *Personnel Psychology*, 69(1), 67–121. [\[CrossRef\]](#)
- Maurin Saldaña, E., Pérez-Delgado, M.-L., & Canales, J. (2026). An exploratory analysis of managerial competencies through 360° evaluation and linear regression: A case study and preliminary validation of the MLPD model. *Administrative Sciences*, 16(5), 216. [\[CrossRef\]](#)
- McNeish, D. M., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, 51(4), 495–518. [\[CrossRef\]](#) [\[PubMed\]](#)
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. [\[CrossRef\]](#) [\[PubMed\]](#)
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.
- Nestler, S., Humberg, S., & Schönbrodt, F. D. (2019). Response surface analysis with multilevel data: Illustration for the case of congruence hypotheses. *Psychological Methods*, 24(3), 291–308. [\[CrossRef\]](#) [\[PubMed\]](#)
- Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372–411. [\[CrossRef\]](#)
- Nielsen, K., Tafvelin, S., von Thiele Schwarz, U., & Hasson, H. (2022). In the eye of the beholder: How self-other agreements influence leadership training outcomes as perceived by leaders and their followers. *Journal of Business and Psychology*, 37(1), 73–90. [\[CrossRef\]](#)
- Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology*, 57(2), 333–375. [\[CrossRef\]](#)
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4), 672–683. [\[CrossRef\]](#)
- Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2), 502–508. [\[CrossRef\]](#)
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. [\[CrossRef\]](#)
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428. [\[CrossRef\]](#)
- Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*, 92(2), 538–544. [\[CrossRef\]](#) [\[PubMed\]](#)
- Saruhashi, M., Li, R., & Kida, N. (2025). Dual pathways to subjective well-being in Japanese employees: Transformational leadership and trait intrinsic motivation. *Administrative Sciences*, 15(11), 431. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggstad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology*, 25(4), 543–554. [\[CrossRef\]](#)
- Sin, H.-P., Nahrgang, J. D., & Morgeson, F. P. (2009). Understanding why they don't see eye to eye: An examination of leader-member exchange (LMX) agreement. *Journal of Applied Psychology*, 94(4), 1048–1057. [\[CrossRef\]](#) [\[PubMed\]](#)
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58(1), 33–66. [\[CrossRef\]](#)
- Sweis, R. J., Aldaod, S., Alsayed, N. M., & Sukkari, L. S. (2022). Self vs. other raters' assessment of emotional intelligence in private and public hospitals: A comparative study. *Administrative Sciences*, 12(4), 194. [\[CrossRef\]](#)
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. [\[CrossRef\]](#)
- Woo, S. E., O'Boyle, E. H., & Spector, P. E. (2017). Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review*, 27(2), 255–264. [\[CrossRef\]](#)
- Xie, D., & Yang, Y. (2025). When does authenticity benefit employee well-being: A relational framework of authenticity at work. *Administrative Sciences*, 15(11), 449. [\[CrossRef\]](#)

- Yao, Y. A., Guo, G., & Yao, H. (2026). Research questions matter: A five-criterion framework for centering decisions in congruence research. *The Leadership Quarterly*, 37(3), 101953. [[CrossRef](#)]
- Zhang, Z., Wang, M., & Shi, J. (2012). Leader-follower congruence in proactive personality and work outcomes: The mediating role of leader-member exchange. *Academy of Management Journal*, 55(1), 111–130. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.