Reflect 360 Methodology Brief

A 360-degree (multi-rater) feedback instrument for individual development: how it was built, where its competencies and behaviors come from, and our approach to establishing validity.

Version 1.2 · 17 November 2026

1. Why this exists

Clients ask three questions about any 360 feedback tool: *Where do the competencies and behavior items come from? How do you protect rater anonymity and data quality? What can a 360 score be used for, and what can it not be used for?* This brief answers those questions for Reflect 360. It follows the structure of the AI Readiness Compass, Assessment Center, Fluent and Techno briefs so that the product family reads consistently.

2. Construct definitions

Reflect 360 measures **observable workplace behavior** as seen from multiple rater perspectives: self, manager, peers and direct reports, with skip-level managers and other raters (for example internal clients or cross-functional colleagues) available as optional groups. The content model is a **framework → competency → behavior** hierarchy:

- **Frameworks** - a competency set, either a reusable library template or an engagement-specific framework.
- **Competencies** - the behavioral dimensions being rated.
- **Behaviors** - the concrete, observable items raters respond to, each tied to one competency and optionally tagged with a leadership level.

Reflect 360 is explicitly a **developmental feedback** instrument. It tells a participant how their behavior is perceived across rater groups, to drive an individual development plan - not to rank or select people.

3. Competency and behavior development

Sourcing and format

Each competency has a name and a definition. Each behavior is an observable statement tied to one competency. A framework can be cloned from the reviewed library, written by a consultant, or drafted with AI assistance from the client's own values and competencies.

Every behavior records its origin: written by a consultant or proposed by AI. AI-proposed behaviors carry a short rationale and are reviewed before use. A consultant must approve a manually built or AI-assisted framework before the engagement can be launched and raters invited; a framework cloned from the reviewed library is treated as approved when it is created.

Behaviors can be tagged with a leadership level (executive, senior manager, manager, individual contributor, or all levels). In the current release, every behavior in the engagement's framework is presented to all raters.

Research anchors

Library competencies carry a per-competency **validation-evidence trail** where one has been reviewed. The trail anchors the competency to the multisource-feedback and competency-modeling literature (§6) from a curated, closed bibliography, records a confidence level, and counts as reviewed only once a person has verified it. As of 16 November 2026, none of the 120 library and engagement competencies yet has a verified validation-evidence trail. Coverage is tracked and reviewed internally by VIFM.

Behavioral anchoring

Items are written as observable behavior statements rather than trait adjectives and are rated for frequency, in the tradition of behavior-based rating formats such as behaviorally anchored rating scales (BARS) (§6). This keeps ratings grounded in what a rater actually saw, which is the foundation of both face validity and rater reliability.

---

4. Validity

Content validity

Every behavior maps to exactly one competency, and competencies are defined before behaviors are written. A consultant approves each manually built or AI-assisted framework before launch; frameworks cloned from the reviewed VIFM library are used as reviewed. This is the strongest evidence the product holds today.

Face validity

Behavioral anchors give raters concrete statements to respond to; we mark this *partial* until a structured face-validity review (documenting that raters interpret each behavior as intended) is recorded.

Construct validity

A formal construct-validity study (factor structure of the competency model; convergent/discriminant patterns across rater sources - multitrait-multimethod) has **not** been run. The framework rests on expert content review, not on empirical confirmation - a disclosed gap.

Criterion validity

Reflect 360 is a **developmental** instrument; we make no selection or job-performance prediction claim. Criterion-related predictive validity has not been established and is not claimed for this use. The

published 360 literature is candid that multisource ratings are best used for development, not high-stakes decisions (§6).

---

5. Rater confidentiality and reliability

Rater confidentiality (anonymity threshold)

The anonymity threshold is a confidentiality rule that controls when results are released. It protects rater candor. It is not a measure of inter-rater reliability or of statistical agreement between raters. The platform applies it as follows:

- **Threshold.** Each engagement has a minimum number of raters per group. The minimum is three and the platform does not allow a lower setting; a consultant may set a higher minimum, up to ten.
- **Protected groups.** Peers, direct reports, skip-level managers and other raters. A protected group's average is shown only when at least the threshold number of raters in that group have given ratings. This is checked separately for the overall result, for each competency and for each behavior, and "Not observable" responses do not count.
- **Self and manager.** These two groups are not anonymized. The self-rating is always shown, and the line manager's ratings and comments are shown under Manager even when there is only one manager, because the manager is identifiable by role. Manager ratings are therefore not anonymous.
- **Small groups are withheld, not merged.** A protected group below the threshold is shown as hidden. It is not merged into another named group. Its ratings still count toward the combined result for all others (all raters except self). That combined average, and the self-versus-others gap based on it, are withheld when the protected raters who contributed number at least one but fewer than the threshold. When the manager is the only other rater, the combined average is the manager's rating.
- **Written comments.** Comments from a protected group are shown only when at least the threshold number of raters in that group wrote comments. This is counted separately for comments on individual behaviors and for answers to the open-ended questions.

Inter-rater agreement

This is the reliability dimension most natural to a 360, because multiple raters respond per participant. No inter-rater reliability or agreement coefficient is computed or reported today - a disclosed gap. The item-level detail of the participant report flags a rater group whose ratings of a behavior differ by three or more points on the 5-point scale, so that the average is read with care. Within-group agreement (and the meaningful self-other gap) is computable from the response data and is interpreted per the multisource-feedback literature (raters in different roles legitimately see different behavior, so perfect agreement is neither expected nor desirable).

Internal consistency

Per-competency internal consistency is **not currently computed**; it can be derived from behavior-level responses once response volume per framework is sufficient. A disclosed gap.

Test-retest reliability

Stability across repeated 360 cycles is **not currently tracked** - a disclosed gap (re-survey cycles make it computable in future). Where a participant is reassessed, the report shows the change in others' ratings against the previous cycle; this is a development indicator, not a test-retest coefficient.

---

6. Reference frameworks and published instruments

Competencies and behaviors content-align with the works below; this is content alignment, not republication. The per-competency validation-evidence trail anchors to these.

Multisource (360) feedback method and psychometrics

- Bracken, D. W., Timmreck, C. W., & Church, A. H. (Eds.) (2001). *The handbook of multisource feedback.* Jossey-Bass.
- Conway, J. M., & Huffcutt, A. I. (1997). *Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings.* Human Performance, 10(4), 331–360.
- 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.
- 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.
- DeNisi, A. S., & Kluger, A. N. (2000). *Feedback effectiveness: Can 360-degree appraisals be improved?* Academy of Management Executive, 14(1), 129–139.

Behavioral rating scales and rater accuracy

- Smith, P. C., & Kendall, L. M. (1963). *Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales.* Journal of Applied Psychology, 47(2), 149–155.
- Woehr, D. J., & Huffcutt, A. I. (1994). *Rater training for performance appraisal: A quantitative review.* Journal of Occupational and Organizational Psychology, 67(3), 189–205.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Sage.

Competency modeling

- Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). *Doing competencies well: Best practices in competency modeling.* Personnel Psychology, 64(1), 225–262.
- Bartram, D. (2005). *The Great Eight competencies: A criterion-centric approach to validation.* Journal of Applied Psychology, 90(6), 1185–1203.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* AERA.

7. Limitations and disclosures

- **Developmental use only.** Reflect 360 is not validated for selection, promotion or pay decisions, and should not be used for them.
- **Self-report and rater perception.** Scores reflect *perceptions* of behavior; rater leniency, politics and halo are real. Anonymity gating and behavioral anchors mitigate but do not eliminate them.
- **No inter-rater agreement coefficient yet.** The anonymity threshold is a confidentiality rule; it is not evidence that raters agree.
- **No empirical factor structure or internal-consistency report yet.** Both are computable as response volume grows; neither is published today.
- **No test-retest evidence yet.**
- **No fairness/DIF analysis** - marked *not applicable* given the non-selection use, but group-difference reporting would need care if use ever changed.

8. Update cadence

- Frameworks and behaviors: curated per engagement; library templates revised as the competency models evolve.
- Methodology brief: re-issued as the product matures. This document is v1.2.
- Validation-evidence trails: reviewed, and their coverage tracked, internally by VIFM.

9. Contact

For methodology questions or research-collaboration inquiries, contact the VIFM consulting team: `contact@viftraining.com`.

Reliability and Validity Evidence

Two kinds of evidence: **(A)** established coefficients for the *method* (multi-source / 360-degree feedback), from the published meta-analytic literature; and **(B)** VIFM's own reliability procedures and their current status. The published figures in (A) describe 360-degree feedback as a general method; they are not results obtained with Reflect 360.

A. Established coefficients for the method (published literature)

- **Self-other agreement:** meta-analytic corrected correlations of about **.35** (self-supervisor) and **.36** (self-peer) (Harris & Schaubroeck, 1988). Low-to-moderate agreement is expected and *informative* in 360 feedback - the self-vs-others gap is the signal, not measurement error.
- **Multi-source design** reduces single-rater bias and idiosyncrasy (Bracken, Timmreck & Church, 2001).
- **Impact:** published meta-analytic evidence shows that performance improvement after 360 feedback is generally small and depends on conditions such as how the feedback is received and followed up (Smither, London & Reilly, 2005, meta-analysis).

Source for the self-other agreement figure (the other works are listed in §6): Harris, M. M., & Schaubroeck, J. (1988). *A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings.* Personnel Psychology, 41(1), 43–62.

B. VIFM's own reliability - method, threshold, and current status

- **Content validity:** behaviors come from a reviewed library framework or are derived from the *client's own* values and competencies (consultant-reviewed), then rated by self, manager, peers and direct reports (and, optionally, skip-level managers and other raters) on a 5-point frequency scale: 1 = Almost never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Almost always. A "Not observable" response is excluded from all averages. Participant and cohort reports show a fixed reference band of 3.5 to 4.25 (the "Favorable Zone"). It is a reference band set by VIFM, not a published norm or a benchmark derived from a comparison group.
- **Anonymity threshold (minimum raters per group, at least 3):** group averages and written comments from peers, direct reports, skip-level managers and other raters are withheld until the group reaches the threshold; self and manager results are not anonymized (see §5). This is a confidentiality rule that protects rater candor. It is not a reliability statistic.
- **Internal consistency** per competency (across its behaviors) has not yet been computed. It is a planned analysis once rater volume is sufficient, with a target of **α ≥ .70**. No internal-consistency result is established or reported today.
- **Computed today:** the self-versus-others gap (the self average minus the combined average of all other raters, at overall, competency and behavior level) and, where a participant is reassessed, the change in others' ratings against the previous cycle.
