Assessment Center Methodology Brief

Job-relevant exercises observed and scored by trained assessors against a defined competency framework, delivered as an Assessment Center or a Development Center.

Version 1.4 · 25 November 2026

1. Why this exists

This brief explains how the assessment-center framework was built, where the competencies come from, and our approach to establishing validity.

Consultants and clients ask three questions about any assessment-center program: *Where did the competencies and exercises come from? How do you know they measure what you say they measure? How will accuracy improve over time?* This brief answers those three questions for the Assessment Center.

It is the Assessment Center counterpart to the AI Readiness Compass methodology brief and follows the same structure, so that the two products read consistently.

Sections 2 to 9 describe the center as delivered with trained assessors. Section 10 describes the **virtual (AI) delivery mode**, in which the conversations are run and rated by AI with no human review, and states where that mode departs from the standards described here.

2. Construct definitions

The framework

The Assessment Center measures behavior against **21 competencies** (framework v2, active since 27 September 2026), each elaborated by behavioral indicators (**205 in service**, positive and negative) and development tips (**116 in service** as of 17 November 2026, at least 3 per competency), organized into **8 clusters** under **four domains**:

- **THINKING** - reasoning, problem-solving, judgment, strategy.
- **RESULTS** - achievement, drive, execution, accountability.
- **PEOPLE** - interpersonal effectiveness, leadership, influence, communication.
- **SELF** - integrity, adaptability, resilience, self-management.

The hierarchy of domain, cluster, competency and indicator is held in the platform's database, so every competency has an explicit place in the model and every behavioral indicator rolls up to exactly one competency.

Each center is built on the specific competencies selected when it is designed. When the framework is revised, a new framework version is issued and superseded competencies are retired rather than deleted, so the record of a completed center keeps the competencies it was rated on; ratings are locked once a center is completed. Each competency also carries descriptive anchors for the points of the rating scale (see §3). The behavioral indicators and the rating-scale anchors have been through an expert review, loaded on 10 November 2026: 205 behavioral indicators and 116 development tips are approved and in service, 30 were rejected and withdrawn, and the anchors for all 21 competencies were reviewed. The Arabic names and definitions of the competencies remain pending human review. Wording changes from expert review are applied to the current framework version and recorded with the reviewer and date; a change to the set of competencies is issued as a new framework version.

Why a behavioral framework

An assessment center is a *method*, not a single test: trained assessors observe candidates in job-relevant exercises (in-trays, role-plays, group exercises, case studies, presentations and competency-based interviews) and rate the behaviors they see against defined competencies. The validity of the method rests on (a) competencies defined before exercises are designed, (b) each exercise mapped to the competencies it is built to elicit, and (c) multiple independent observations per competency. The Assessment Center framework encodes all three.

---

3. Competency and exercise development

Sourcing

The competency model was authored by VIFM consultants drawing on the competency-modeling literature (§6) and refined against the behaviors consultants already use with clients. Each competency carries a definition and a set of positive and negative behavioral indicators that tell an assessor what "good" and "poor" look like in observable terms.

Exercise-competency mapping

Exercises are linked to the competencies they are designed to elicit through an explicit exercise-competency matrix, so no competency is assessed by a single exercise and no exercise is scored against a competency it was not designed to surface. Each center assesses a set of competencies selected for the role (between 4 and 15), not the full framework. This matrix is the backbone of content validity (§4).

Research anchors

Every competency carries a per-competency **research-anchor trail**. Each trail anchors the competency to one or more published instruments or works from a curated, closed bibliography (§6), records a confidence level (*direct adaptation*, *construct-aligned* or *novel*), and is **human-verified** before it is surfaced on any client-facing deliverable. As of 17 November 2026, the trails for 8 of the 21 competencies have been verified or edited by a reviewer. Anchors proposed with AI assistance remain

internal until a VIFM reviewer has verified them: human verification and quality assurance. Coverage is tracked in VIFM's internal evidence and validity map.

Rating format

Behaviors are rated on a **5-point competency rubric** (1 to 5), criterion-referenced to the behavioral-indicator anchors rather than to a norm group, with a descriptive anchor for each scale point per competency and exercise-specific indicator examples for assessors. The scale points are labeled: 1 Significant Development Needed, 2 Development Needed, 3 Competent, 4 Strength, 5 Significant Strength.

Ratings are reached in four steps, and the platform keeps a record of each step:

1. **Independent exercise ratings.** In each exercise, every assessor assigned to a participant records the behaviors observed and rates each competency mapped to that exercise on the 1 to 5 scale, with a written justification. Each assessor works in a separate record, and access rules restrict assessors to their own observations and ratings. These ratings are stored per competency, per assessor and per exercise, preserving the raw observations behind every score.
2. **Preliminary ratings.** After the exercises, each assessor reviews their own evidence across the exercises and records one preliminary rating per competency for each participant, with notes.
3. **Consensus integration.** At a chaired integration meeting, the assessors' preliminary ratings are set side by side for each competency, with their average and a flag where they differ by more than one scale point. The panel then agrees one rating per competency (a whole number from 1 to 5) and records its discussion notes. The agreed rating is a panel judgment based on the evidence, not an automatic average. The overall rating cannot be recorded until the chair is named and has confirmed that the evidence of every contributing assessor was heard.
4. **Overall rating.** For a **selection** center the overall rating is **calculated**, not agreed in discussion: it is the weighted average of the agreed competency ratings, using the competency weights set in the center design. A VIFM administrator must confirm the weights before the platform will calculate any overall rating; where no weights are set, all competencies count equally. The figure is kept to two decimal places and is also shown rounded to the nearest whole point on the 1 to 5 scale. The overall rating is produced only once every competency in the center has an agreed rating; until then the chair sees which competencies are still missing one. The panel cannot enter or change it; it may record that it disagrees and why, and the calculated rating remains the center outcome.
   Development and **succession** centers reach the overall rating (a whole number from 1 to 5) by assessor discussion; a succession center whose outcome will inform an appointment is set up as a selection center. In every center, the readiness recommendation (Ready Now, Ready with Development or Not Ready) is the panel's judgment, recorded with the overall rating.

This design is aligned with the provision of the British Psychological Society standard on arithmetic integration of ratings for selection decisions (§4).

---

4. Validity

Content validity

Every exercise maps to defined competencies through the exercise-competency matrix, and every competency is elaborated by behavioral indicators. The framework was reviewed by VIFM consultants and the framework owner before exercises were accepted. This is the strongest evidence the program holds today.

Face validity

Competencies are expressed as observable behaviors (the indicators) rather than abstract traits, which is the bar for face validity. We have **not** yet run a formal face-validity study (structured respondent and assessor interviews documenting that each indicator is interpreted as intended); VIFM's internal evidence map records this as *partial* until that study is documented.

Construct validity

A formal construct-validity study (e.g., confirmatory factor analysis or a multitrait-multimethod analysis testing whether assessors rate *competencies* rather than *exercises* - the classic assessment-center "exercise effect") has **not** been run. The framework is treated as a content-based model, designed by mapping exercises to competencies and reviewed by experts, not an empirically confirmed factor structure, until that work is done. This is a disclosed gap.

Criterion validity

No criterion-related validity evidence yet exists for VIFM's own centers. Published **meta-analyses** (Gaugler et al., 1987; Arthur et al., 2003 - §6) show that well-designed assessment centers, as a method, can predict later job performance. Those findings describe the method in general; they do not validate VIFM's exercises, assessors or scoring, and VIFM makes no predictive-validity claim for its own centers until its local study reports.

The structure for a local study now exists. For every participant with a finalized overall rating, VIFM schedules a **6- and 12-month outcome follow-up** in the platform, counted from the last day of the center: whether the person went into the role, and the line manager's rating of their performance on the center's own 1 to 5 scale, collected by VIFM from the client. The correlation between center rating and later performance is computed from these, for the participants who went into the role, and reported with its sample size: no figure is shown below 10 participants, it is marked an early indication below 30, and provisional below the standard's guide of about 100. Whether a client agrees to this collection is recorded per center in the evaluation plan and in the center agreement.

Alignment with the BPS standard

VIFM has mapped the program clause by clause against the British Psychological Society's *Design and Delivery of Assessment Centres* (Division of Occupational Psychology, 2015), and the program is designed to be aligned with it. The platform now provides for each requirement that software can support: purpose recorded per center and driving the integration rule; a center agreement and plan generated from the design; a participant joining pack, consent, adjustments, concerns and appeals;

center roles with competence records; staffing rules; a delivery log; exercise quality checks; fairness monitoring; a guidance note for decision makers with every set of reports; and the validity evidence described above, summarized per center in an evidence pack. This is VIFM's own assessment of alignment. It has not been audited, certified or accredited by the Society, and the platform features support, but do not by themselves ensure, that a given center is run to the standard. Alignment is VIFM's own mapping against the published standard or law; it is not a certification, accreditation, audit or legal opinion by the body concerned. The clauses that remain open are those only data can close: a validation sample and the evaluations that run once centers have been delivered.

---

5. Reliability

Inter-rater reliability

The platform records each assessor's rating per competency for each exercise, so inter-rater agreement (ICC(2,1): two-way random effects, absolute agreement, single measure) is **computed** from the audit trail, on complete cases only: the pair of assessors who rated the most participant-competency combinations in common, with no imputation of missing ratings. It is reported per center in the evidence pack and pooled across centers in VIFM's assessor analytics console. This is the reliability evidence most natural to an observational method.

Internal consistency

Cronbach's alpha is **not applicable** in the usual sense: assessment-center scores are observational competency ratings, not responses to a multi-item scale. We therefore do not report alpha for the Assessment Center and mark the cell *n/a* rather than *missing*.

Test-retest reliability

Stability of competency ratings across repeated assessments is **not currently tracked**. This is a disclosed gap.

---

6. Reference frameworks and published instruments

Items, competencies and exercises were authored to content-align with established works; this is *content alignment*, not republication of a validated instrument, so we do not claim the source instruments' coefficients. The list below is what every per-competency research-anchor trail anchors back to. Citations are deliberately limited to seminal, well-established works so that each can be readily verified.

Assessment-center method and standards

- International Taskforce on Assessment Center Guidelines (2015). *Guidelines and ethical considerations for assessment center operations* (6th ed.). Journal of Management, 41(4), 1244–1273.
- Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development.* Lawrence Erlbaum.
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). *Meta-analysis of assessment center validity.* Journal of Applied Psychology, 72(3), 493–511.
- Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). *A meta-analysis of the criterion-related validity of assessment center dimensions.* Personnel Psychology, 56(1). https://doi.org/10.1111/j.1744-6570.2003.tb00146.x
- British Psychological Society, Division of Occupational Psychology (2015). *The design and delivery of assessment centres.* British Psychological Society.
- 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.
- Shrout, P. E., & Fleiss, J. L. (1979). *Intraclass correlations: Uses in assessing rater reliability.* Psychological Bulletin, 86(2), 420–428.
- ISO 10667-1/2 (2020). *Assessment service delivery - Procedures and methods to assess people in work and organizational settings.* International Organization for Standardization.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* AERA.

Combining assessment information (virtual mode, §10)

- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). *Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis.* Journal of Applied Psychology, 98(6), 1060–1072.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). *Clinical versus mechanical prediction: A meta-analysis.* Psychological Assessment, 12(1), 19–30.
- Koo, T. K., & Li, M. Y. (2016). *A guideline of selecting and reporting intraclass correlation coefficients for reliability research.* Journal of Chiropractic Medicine, 15(2), 155–163.

Competency modeling

- Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance.* Wiley.
- Boyatzis, R. E. (1982). *The competent manager: A model for effective performance.* Wiley.
- 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.

Domain-specific anchors

- **THINKING** - Schmidt, F. L., & Hunter, J. E. (1998). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.* Psychological Bulletin, 124(2), 262–274; Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). *Leadership skills for a changing world: Solving complex social problems.* The Leadership Quarterly, 11(1), 11–35.
- **RESULTS** - McClelland, D. C. (1973). *Testing for competence rather than for intelligence.* American Psychologist, 28(1), 1–14; Locke, E. A., & Latham, G. P. (2002). *Building a practically useful theory of goal setting and task motivation: A 35-year odyssey.* American Psychologist, 57(9), 705–717.
- **PEOPLE** - Mayer, J. D., Salovey, P., & Caruso, D. R. (2008). *Emotional intelligence: New ability or eclectic traits?* American Psychologist, 63(6), 503–517; Yukl, G. (2012). *Effective leadership behavior: What we know and what questions need more attention.* Academy of Management Perspectives, 26(4), 66–85.
- **SELF** - Barrick, M. R., & Mount, M. K. (1991). *The Big Five personality dimensions and job performance: A meta-analysis.* Personnel Psychology, 44(1), 1–26; Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). *Adaptability in the workplace: Development of a taxonomy of adaptive performance.* Journal of Applied Psychology, 85(4), 612–624.

> **A note on confidence.** Every competency was authored against its definition, not by lifting wording from these sources. The per-competency research-anchor trail flags each anchor as *direct adaptation*, *construct-aligned* or *novel*, and is human-reviewed before publication.

---

7. Limitations and disclosures

- **No empirical factor structure yet.** Construct validity (including the exercise-versus-dimension question) and differential item functioning (DIF) analyses have not been run. An adverse-impact check (the four-fifths rule on center recommendations) is computed per selection or succession center where participants disclose demographics; with small groups it carries low confidence. The framework is a content-based model, designed by mapping exercises to competencies and reviewed by experts, not empirically confirmed.
- **No criterion validity evidence for VIFM centers yet.** Published assessment-center meta-analyses describe the method in general and do not validate VIFM's own scoring. The follow-up structure for a local study exists, but no center has yet accumulated the follow-ups to report from, and about 100 participants are needed for a study with reasonable power.
- **Framework v2 content was drafted with AI assistance and then expert-reviewed.** Competency definitions, behavioral indicators and rating-scale anchors were drafted with AI assistance. An expert review of the indicators and anchors was loaded on 10 November 2026; the Arabic names and definitions of the competencies remain pending human review. Each report states the review status of the content used.
- **Rater effects are real.** Observational ratings carry leniency and halo risk. Assessor training and multi-assessor designs mitigate but do not eliminate it; inter-rater agreement is the reliability metric we report.
- **Test-retest stability is not yet tracked.**
- **The virtual (AI) mode has no agreement evidence yet.** Its ratings are made by AI with no human review, and the pilot that compares them with assessors has not yet run (§10).

---

8. Update cadence

- Competency framework: re-versioned as the framework owner revises it; the active framework is the source of truth for new assessments.
- Methodology brief: re-issued with each framework revision, and when the virtual (AI) pilot reports. This document is v1.4 (framework v2).
- Research-anchor trails: reviewed on a continuing basis by VIFM; coverage is tracked in VIFM's internal evidence and validity map.

---

9. Contact

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

---

10. Virtual (AI) delivery

What it is

A virtual center is an Assessment Center engagement delivered fully online, for clients who need to assess large numbers in a short period. It uses the same competency framework and rating scale (§2, §3). Each assessee receives a personal link that stays open for a window the client sets, and may complete the center in more than one sitting; every answer is saved as it is given. The center has three AI-run conversations and two questionnaires:

- **An AI competency interview.** The questions are taken word for word from a content library approved by VIFM's subject-matter expert. The AI interviewer asks up to two follow-up questions per question, written from the approved follow-up probes.
- **Two AI role-plays** of about 20 minutes each, with a character played by AI from an approved character brief (for example, an unhappy client or a team member raising a concern).
- **Persona and Logica** taken alongside. Their results are shown in the report as supporting information and are **never counted** in the competency ratings. Logica is reported as indicative while its items are under review.

The assessee takes part in English or Arabic, and may speak or type. Spoken answers are turned into text, and the assessee can correct the text before sending it. The client decides, per engagement, whether the conversations are also video-recorded. **The AI scores the words in the transcript only, never the video, the voice or the appearance of the assessee.** Each competency assessed is covered by at least two of the three conversations; the platform will not invite assessees until the selected content meets this.

How the ratings are reached

1. **Conversation ratings.** After the assessee finishes, the AI rates each competency in each conversation where it was targeted, on the 1 to 5 scale (§3), with its reasons and the assessee's own words as evidence. Every quote is checked against the transcript, word for word, before it is kept. A rating whose quotes cannot be found in what the assessee said is not kept: that conversation counts as giving insufficient evidence for the competency. The AI may also judge that a conversation gave too little evidence to rate a competency fairly.
2. **Final competency ratings.** The AI then reads the evidence of all the conversations together and gives one final rating per competency, as an assessor panel does at the integration meeting. The final rating is a judgment from the evidence, not an average.
3. **Overall rating.** The AI gives the overall rating (a whole number from 1 to 5) as a judgment of the whole performance, taking the competency weights into account.
4. **Recommendation.** The readiness recommendation follows fixed bands of the overall rating: 4 or 5 Ready Now, 3 Ready with Development, 1 or 2 Not Ready.

If any competency of the center has insufficient evidence, the whole report is held and no ratings are issued. A VIFM administrator then decides: reopen one conversation for the assessee to do again, or close the case with a recorded reason. A reply from the AI that is incomplete or cannot be read is never

filled in with a default rating: the step is tried again, and if it still fails, it waits for VIFM to run it again. Every AI rating records the model that produced it.

Rated by AI with no human review, and where this departs from the standards

Every virtual report, and the consent the assessee gives before starting, states that the ratings are made by AI with no human review. For a center whose purpose is selection, both also state that the results are advisory and that the client makes any decision. This is a deliberate design choice for the virtual mode, made so that large numbers can be assessed in a short period. It departs from the standards described in §4 in three ways, which VIFM discloses here rather than describing the mode as aligned:

- **No trained human assessors.** The BPS standard and the international guidelines (§6) describe a center in which trained assessors observe and rate behavior. The virtual mode has no assessors, so the staffing provisions of the BPS standard (more than one assessor per participant, 5.18; at least one assessor per three participants, 5.21) do not apply to it, and the platform does not check them for virtual engagements.
- **The overall rating is an AI judgment, not a calculation.** For in-person selection centers, the overall rating is calculated from the agreed competency ratings (§3), in line with the BPS provision on arithmetic integration (7.4). In the virtual mode the overall rating is the AI's holistic judgment, for selection centers as well. The research on combining assessment data favors calculation: a meta-analysis of selection and admissions decisions found that combining the same information by a formula predicted later outcomes markedly better than holistic judgment, an improvement in prediction of more than 50% (Kuncel, Klieger, Connelly & Ones, 2013), and an earlier meta-analysis across fields found formula methods about 10% more accurate on average (Grove, Zald, Lebow, Snitz & Nelson, 2000). Those studies concern human judgment; whether the same holds for an AI judge has not been established. The pilot below compares the AI's overall ratings with assessors' overall ratings, and the result will be reported here.
- **No panel discussion of the recommendation.** In-person, the recommendation is the panel's judgment; in the virtual mode it is set by the fixed bands above.

Integrity signals (leaving the page, pasting text, answer timing, and webcam pictures where the client turns them on) are advisory. They are shown in the report and never change a rating or fail an assessee automatically.

Data

Transcripts, ratings and reports are stored on Caliber's cloud database in Singapore. The conversations are processed by AI and speech services from Anthropic (the AI interviewer, characters and ratings), OpenAI (turning speech into text) and Microsoft (the AI voice), which may process data outside the assessee's country. The text of the conversations, the quotes taken from it and the integrity signals are deleted automatically when the engagement's retention period ends, counted from the day the assessee finished; the ratings and reports are kept as the client's record. Video recordings are deleted after the number of days the client sets, and webcam pictures after 90 days.

Validation: the pilot

Until the pilot has run, VIFM makes **no reliability or validity claim** for the AI's ratings. The pilot is set up as follows, and the pass bar was fixed on 25 November 2026, before any pilot data existed:

- **Participants:** 10 to 20 volunteers, from VIFM and outside it, take a virtual center; none comes from a client organization. Their consent states that two VIFM assessors also rate their conversations, and their results are not released to any employer.
- **Assessors:** two VIFM assessors rate each volunteer independently and blind: they read the transcripts (and watch the video where it was recorded), never see the AI's ratings or each other's, and their ratings lock when submitted. They rate each competency in each conversation, give a final rating per competency, and an overall rating, as the AI does.
- **Agreement measures:** ICC(2,1), absolute agreement, single measure (Shrout & Fleiss, 1979), between the AI and the average of the two assessors, and between the two assessors as the human benchmark; the share of AI ratings within one point of the assessors' average; and the average difference, which shows whether the AI is more lenient or harsher than the assessors. Each is reported for single conversations, final competency ratings and the overall rating, by competency and by conversation. Ratings where anyone found the evidence insufficient are left out and counted.
- **Pass bar, on the final competency ratings:** (1) the AI's ICC with the assessors is at least **.70**, and no more than **.10** below the two assessors' ICC with each other; (2) at least **90%** of the AI's ratings are within one point of the assessors' average; (3) the average difference is within **0.3** points. With fewer than 30 complete ratings, no verdict is given. ICC values from .75 to .90 are commonly read as good reliability (Koo & Li, 2016); VIFM's own target for assessor agreement is .70 (§5).

The result, whether it meets the bar or not, will be added to this section with its sample size. Because a change of AI model can change the ratings, the comparison is repeated when the model used for scoring changes.

Limitations of the virtual mode

- **No agreement or validity evidence yet.** The pilot has not run; no criterion validity study exists for the virtual mode.
- **AI judgment can be wrong in ways that are hard to see.** The quote check ensures that evidence was really said, not that it was weighed correctly. Language models can also be influenced by fluency or length; the instructions tell the AI not to reward either, but this has not been tested.
- **Arabic.** Conversations in Arabic are rated in Arabic. The pilot reports all languages together; whether agreement differs for Arabic sittings can only be checked once enough of them exist.
- **Not an assessment center in the sense of the standards.** For the reasons above, a virtual center does not meet the definition of an assessment center used by the BPS standard and the international guidelines, and VIFM does not describe it as aligned with them.

Reliability and validity evidence

Two kinds of evidence: **(A)** established coefficients for the *method* (the assessment-center method and BARS), from the published meta-analytic literature; and **(B)** VIFM's own reliability procedures and their current status. We report (A) as evidence for the method, not as a claim about a VIFM-specific coefficient.

A. Published research on the assessment-center method (not VIFM results)

- **Overall criterion validity of the method:** a meta-analysis of 50 studies reported a corrected mean validity of about **.37** across criteria such as job performance, potential and career advancement (Gaugler, Rosenthal, Thornton & Bentson, 1987). These figures describe assessment centers in general and are not coefficients for VIFM's centers.
- **Dimension-level validity:** across six broad assessment-center dimensions, estimated criterion-related validities ranged from about **.25 to .39** (Arthur, Day, McNelly & Edens, 2003). Whether assessors rate dimensions or exercises remains a well-documented open question in the literature.
- **Rating method:** Behaviorally Anchored Rating Scales were developed to give raters unambiguous, behavior-based anchors (Smith & Kendall, 1963). VIFM uses anchored scales to support consistent reading of the scale; it has not tested their effect on halo or leniency in its own centers.

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

- **Inter-rater reliability:** Intraclass correlation (**ICC**; Shrout & Fleiss, 1979) between assessors' independent ratings, which are then integrated at the chaired consensus meeting; target **ICC ≥ .70** (acceptable agreement; .80 and above is rated good). Status: ICC is computed per center wherever two assessors have rated the same participants on the same competencies (at least two such ratings in common); a pooled figure across centers is already computed for internal monitoring and will be published once enough multi-assessor centers have run.
- **Content validity:** each competency assessed at a center is mapped to **≥ 2** exercises (the exercise-competency matrix; enforced when the center is designed) - the documented coverage design. Criteria overlap is checked once 10 or more participants are rated on a pair (pairs correlating above 0.7 are flagged for merging, per the standard).
- **Criterion validity:** the correlation between center rating and the 6- and 12-month manager performance rating, per center, reported only with its sample size. Status: structure live since 30 September 2026; no follow-ups collected yet.