A note on how the assessment is built

Appendix · Building Beyond Engagement field guide
Archetype Learning Solutions

Every assessment makes three quiet decisions before it asks its first question: what it asks about, how it lets you answer, and what it does with the answers. Most instruments never explain those decisions to the person taking them, which is part of why so many managers have learned to distrust the whole genre. This one is built on three choices, and they are worth stating plainly.

The scale asks how often, not whether you agree

The twelve statements ask you how frequently you do something, and every point on the scale carries a label: almost never, occasionally, about half the time, usually, almost always. It would have been easier to write the items as agreement statements — "I give recognition regularly," strongly disagree to strongly agree — because that is how most workplace surveys are written. It would also have been worse.

Agreeing with a statement about yourself is not the same kind of mental act as counting. When you agree, you are consulting your self-image; when you estimate frequency, you are consulting your calendar. The first has no external referent, so nothing can check it. The second does. If you answer "usually" to the item about one-to-ones and your calendar shows two in the last quarter, the item has caught something, and you are the one who catches it. That is the whole design intent: an item you can be wrong about in a way you can notice.

Labelling every point rather than only the ends matters for the same reason. An unlabelled middle invites you to answer by feel. A labelled middle — "about half the time" — makes you do the arithmetic. This is not a perfect solution, and it is worth being honest about the limitation: words like "usually" do not mean exactly the same thing to everyone, and two managers who both answer "usually" may be describing genuinely different behavior. What keeps that from wrecking the instrument is the pairing of a vague word with a specific behavior. "Usually" attached to a named practice on a known cadence is a much narrower target than "usually" attached to a trait.

The four elements do not average against each other

The scoring rule that matters most is the one that refuses to trade. You get an overall figure, but you also get a floor on every element, and both have to be met. A manager who is exceptional at recognition and absent from growth and development does not clear the badge by averaging the two.

This is a deliberate departure from how scores are normally combined. Most scoring adds everything up and asks whether the total is high enough, which treats a strength in one area as genuine compensation for a gap in another. That works when you are measuring a single

underlying thing. It does not work here, because the four elements are not four measurements of one quality — they are four separate obligations a manager owes a team, and the person on the receiving end of a neglected one does not experience it as offset. An employee who is thanked warmly and never developed has an unmet need, not an averaged one.

The practical consequence is that the floor, not the total, is usually what holds someone back, and the report tells you which element did it. That is the most useful sentence the instrument produces. A total score tells you roughly where you stand. A floor tells you what to do on Monday.

The manager answers, and that is a limitation with a purpose

Your employees do not fill this in. You do. That is unusual enough to need defending, because the obvious objection is the right one: people rate themselves generously, managers included, and the research on this is not ambiguous. When managers' self-ratings are compared with the ratings their own teams give them, the managers come out higher — reliably, across every dimension measured.

So why ask the manager? Because of what the instrument is for. There are two very different reasons to measure something. One is to decide — to award, rank, promote, or certify. The other is to learn — to show someone their own practice back to themselves clearly enough that they can act on it. The first requires accuracy, because a decision made on a wrong number is a wrong decision. The second requires honesty and specificity, but tolerates a good deal of imprecision, because the person reading the result is the person who will act on it and no one else is affected by the error.

This assessment is the second kind. It is a mirror. Nobody scores you, nothing is stored, no one else sees the result, and there is no incentive to inflate — inflating it only wastes your own afternoon. Ask an employee to rate their manager and you have made something with consequences, which changes what people write. Ask the manager, privately, with no stakes attached, and the generous self-rating becomes a manageable problem rather than a fatal one.

It stops being manageable the moment a result is used to award something. That is exactly where this framework stops relying on self-report. Levels 1 and 2 you may claim from your own answers, because nothing turns on them. Level 3 requires the same twelve statements answered about you by your team, and applies the thresholds to their data, not yours — with the gap between your rating and theirs recorded as a finding in its own right. Level 4 is not assessed on your scores at all; it is assessed on whether another manager reached verified standing under your mentoring.

The rule underneath all of it: self-report is enough to start a conversation and never enough to issue a credential.

Sources

- Marfeo, E.E., Ni, P., Chan, L., Rasch, E.K., & Jette, A.M. (2014). Optimizing psychometrics in measuring behavioral health functioning: combining agreement and frequency rating scales. *Journal of Clinical Epidemiology*, 67(7), 781–784. https://pmc.ncbi.nlm.nih.gov/articles/PMC4066462/
- Nelson Laird, T.F., Korkmaz, A., & Chen, P-S.D. (2008). How often is "often" revisited: the meaning and linearity of vague quantifiers used on the National Survey of Student Engagement. Annual Meeting of the American Educational Research Association. https://www.purdue.edu/idata/documents/Laird-et-al-2008-vague-quantifiers.pdf
- Meyers, J.L. (2018). Scoring models in competency-based educational assessment. *Competency-based Education*, 3(2), e01173. https://onlinelibrary.wiley.com/doi/full/10.1002/cbe2.1173
- Herbst, T.H.H., & Conradie, P.D.P. (2011). Leadership effectiveness in higher education: managerial self-perceptions versus perceptions of others. *SA Journal of Industrial Psychology*, 37(1), Art. 867. https://scielo.org.za/pdf/sajip/v37n1/v37n1a01
- Formative and summative assessment serve different purposes and carry different evidentiary burdens. Carnegie Mellon University Eberly Center. https://www.cmu.edu/teaching/assessment/basics/formative-summative.html