CALIBRATION · GUIDE

How to run a skills *calibration session*

A skills matrix is only as trustworthy as the consistency of its scores, and the single best way to get that consistency is a calibration session: managers in a room, agreeing what each level really means against the evidence. Done well, it makes a "3" mean the same on every team. This is how to run one, the pre-work, the agenda, and the bias checks that make it fair.

Dr Alex J. Martin-Smith
CMGR · MBA · LLM · DBA

Reading time 12 min · **Method** Upleashed 0 to 5 capability framework · **Updated** May 2026

THE SHORT ANSWER

To run a skills calibration session, have managers complete their ratings with evidence beforehand, then meet for 60 to 90 minutes with a facilitator to compare scores, discuss the outliers and large gaps, run explicit bias checks, and agree adjustments so a given level means the same across every team. Document the decisions and update the matrix. In short: **prepare ratings in advance, align them against evidence in a structured session, check for bias, and record the agreed result.**

KEY TAKEAWAYS

- **Calibration makes a level mean one thing.** Its whole purpose is that a "3" means the same no matter which manager scored it.
- **Pre-work is non-negotiable.** Managers bring completed, evidence-backed ratings; the session is for alignment, not for scoring from scratch.
- **Keep it short and facilitated.** 60 to 90 minutes, a neutral facilitator, ground rules of evidence over opinion.
- **Focus on the outliers.** Spend the time on the ratings that disagree, not the ones everyone already agrees on.
- **Check for bias explicitly.** Behavioural anchors and bias prompts measurably cut disagreement between raters.

START HERE

Calibration is how a "3" *means one thing*

A skills matrix promises comparable data: scores you can read across people, teams and time. But that promise only holds if every rater applies the scale the same way, and left to themselves, they will not. A calibration session is the structured conversation that fixes this, the step that turns a collection of individual opinions into consistent, trustworthy data.

The problem it solves

Without calibration, a "3" from a generous manager and a "3" from a demanding one mean different things, so the moment you compare scores across teams you are comparing noise. This is the quiet flaw that undermines otherwise well-built matrices. Calibration solves it by bringing raters together to agree, against evidence, **what each level actually looks like**, so that a given score carries the same meaning regardless of who assigned it. It is the difference between a measuring instrument and a pile of unrelated judgements.

It is alignment, not a scoring meeting

The most important thing to understand about calibration is what it is *not*: it is not a session for completing ratings. Managers must arrive with their scoring already done and evidence ready; the meeting itself is for **discussion and decisions, not discovery**. When people turn up unprepared, the session collapses into a slow group scoring exercise and never reaches its real job,

aligning standards. The goal in the room is alignment, achieved by comparing, challenging and agreeing, not by rating from scratch.

It runs on evidence, not seniority

A good calibration session has one golden rule: **evidence over opinion**. When two managers disagree on a level, the question is not who is more senior or more insistent, but what the evidence shows. Behavioural anchors, clear descriptions of what each level looks like in practice, give everyone a shared reference, and research finds they measurably reduce disagreement between raters. Run this way, calibration is not about winning an argument; it is about matching real evidence to a shared, defined scale.

WHY IT MATTERS NOW

Uncalibrated scores *cannot be compared*

Every benefit of a skills matrix, gap analysis, coverage, succession, fair development, rests on scores meaning the same thing across the organisation. Skip calibration and that foundation cracks: the data looks comparable but is not, and decisions made on it are quietly unfair.

40%

MCKINSEY, VIA SPRAD 2025

less disagreement between raters when behavioural anchors are used, the core mechanism of calibration.

8%

GARTNER, 2024

of organisations have reliable workforce skills data; inconsistent rating is a leading reason why.

60–90

WIDELY RECOMMENDED

minutes is the sweet spot for a calibration session, long enough to align, short enough to stay sharp.

The stakes are higher than they look, because uncalibrated data does not announce itself. A matrix full of inconsistent scores looks exactly like a good one until a decision goes wrong: the wrong person identified as a gap, an unfair comparison in a promotion, a coverage figure that is not real. Calibration is the inexpensive insurance against all of it. An hour or so with the right managers, run with structure and evidence, converts a set of

individual opinions into **data the whole organisation can trust**, which is the entire point of building a matrix in the first place.

WHAT IT FIXES

Four things calibration puts right

A calibration session targets four specific problems that creep into any rating process. Each, left unchecked, quietly corrupts the data; together they are why calibration is worth the hour.

FIXES 01

Inconsistent standards

Different managers reading the scale differently is the core problem. Calibration aligns them so a level means the same across every team.

FIXES 02

Rating inflation

Generous scoring that flatters a team but hides gaps gets surfaced when ratings are compared against peers and evidence.

FIXES 03

Individual bias

Recency, leniency, the halo effect and central tendency all distort scores. Explicit bias checks catch them before they reach the data.

FIXES 04

Unfair comparison

When scores are not comparable, every cross-team decision is unfair. Calibration makes comparison legitimate, and decisions defensible.

Notice that all four share a root cause: **judgement made in isolation drifts**. A single manager, however conscientious, cannot see how their scoring compares to everyone else's, so their scale slowly diverges. Calibration is simply the act of bringing those isolated judgements together and reconciling them against a shared standard and real evidence. That is why it works, and why no amount of careful individual rating can replace it: consistency is a property of the group, not of any one rater.

THE SCALE BEHIND THE SCORES

The 0 to 5 capability framework

Calibration is far easier when the scale already has clear, behavioural definitions to anchor on. This framework, developed by Dr Alex J. Martin-

Smith, gives each level an observable meaning, so the session aligns on evidence against a description rather than arguing about adjectives.

0

No skill required or desired EXCLUDED

The skill is not needed for this role within the next year. In calibration, agreeing what is genuinely out of scope is itself useful, it stops irrelevant skills muddying the comparison.

1

In training / Trainee WEIGHTING 25%

Up to 75% trained and does not yet fully understand the quality requirements. A clear behavioural anchor: still learning, not yet working alone.

2

Developing capabilities WEIGHTING 50%

More than 75% trained; can probably work alone, but consistent quality is not yet evidenced, so complex output still needs checking. The classic "is this a 2 or a 3?" boundary calibration resolves.

3

Capable WEIGHTING 75% · THE KEY LINE

100% trained, consistent quality, works unsupervised. The most-discussed line in any calibration: the evidence test is "do they reliably do this alone, to standard?"

4

Subject Matter Expert / Trainer WEIGHTING 100%

Prolonged expertise; works autonomously and can train others. Anchor: not just capable, but the person others learn from. If unused three months, drop to Level 3.

5

Strategic ownership / Leadership WEIGHTING 100%

Defines processes and standards, shows cross-function expertise and leadership. Anchor: shapes how the skill is done across the team, not just performs it.

Behavioural anchors are the calibration tool

The reason a defined scale matters so much for calibration is that its level descriptions act as **behavioural anchors**, shared, observable reference points. Instead of debating whether someone is "strong", managers ask which described level the evidence fits. This is exactly the mechanism research credits with cutting rater disagreement by around 40%. The framework's clearest anchor, Level 3's "works unsupervised, consistent quality", resolves the most common calibration dispute of all.

A worked example. Resolving a disagreement with an anchor, not an argument:

```
Manager A says 3, Manager B says 2 for the same person on CRM
anchor test  →  "Does their work still need checking?"  Yes,
on complex cases
agreed  →  Level 2 — the evidence fits the described level,
not the louder voice.
```

SEE THE SESSION

A calibration session, *end to end*

Here is a calibration session laid out on a timeline, from the pre-work that makes it possible, through the structured hour in the room, to the documentation that makes it stick. The session itself is short and focused; the discipline is in the preparation and the bias checks.

before the session · in the session · after the session

60–90

minutes in the room, with the real work done in pre-work and the decisions captured after

Illustrative calibration flow on the Upleashed 0 to 5 framework. Most of the value comes from preparation and structure, not from a long meeting.

WALKING THE TIMELINE

- **Before · pre-work.** Each manager completes their ratings and gathers evidence at least a day or two ahead. The facilitator scans for flashpoints, the big disagreements to spend time on. No pre-work, no calibration.
- **In the session · align.** Set ground rules (alignment, evidence over opinion, confidentiality), then review and compare ratings, focusing on outliers. Each manager explains a score; the group challenges it against the evidence and the anchors.
- **In the session · bias check.** Before finalising, run explicit prompts: any recency, leniency, halo or central-tendency effects? This single step is what stops hidden bias surviving into the agreed data.
- **After · document.** Record the agreed levels and the rationale, update the matrix, and feed back to individuals. The decisions and reasons are kept, both to act on and to improve next cycle.

RUNNING IT

Seven steps to a productive session

This is the practical playbook, from setting it up to closing it out. Follow it and the session stays focused, fair and short, and produces data everyone trusts.

1 Require pre-work

Have every manager submit completed, evidence-backed ratings a day or two before. This is the single most important rule: the meeting is for alignment, not scoring. Sharing ratings in advance also lets the facilitator spot the disagreements worth the room's time.

> **WATCH OUT** If managers arrive unprepared, the session becomes a slow group scoring exercise and never reaches alignment. Enforce the pre-work.

2 Keep the group small and add a facilitator

Invite the managers whose ratings overlap, plus a neutral facilitator (often HR or a senior leader) to guide and keep it fair. A small, focused group aligns faster than a large one, and a facilitator stops the loudest or most senior voice from setting the standard.

> **WATCH OUT** Without a neutral facilitator, calibration drifts toward whoever is most insistent, not what the evidence shows.

3 Set ground rules and the goal

Open by stating the goal, consistent, fair, evidence-based levels, and the rules: evidence over opinion, confidentiality, respectful challenge. Reminding everyone that the aim is alignment, not judging each other's teams, sets the right tone for the discussion that follows.

> **WATCH OUT** Skipping the ground rules invites defensiveness; managers protect their scores instead of aligning them.

4 Review and compare, focusing on outliers

Work through the ratings, but spend the time where it matters: the outliers and large gaps. Compare similar roles across teams, and have each manager briefly explain a contested score and its evidence. Do not relitigate the scores everyone already agrees on.

> **WATCH OUT** Reviewing every score equally wastes the session. Triage to the disagreements; agreement needs no discussion.

5

Test against the anchors

When a score is contested, resolve it by matching the evidence to the level description, not by debate. "Does their work still need checking?" settles a 2-versus-3 faster than any argument. The behavioural anchors are the neutral arbiter that keeps it fair.

> **WATCH OUT** Resolving disagreements by seniority rather than evidence defeats the purpose and breeds distrust in the result.

6

Run an explicit bias check

Before finalising, pause and ask the bias questions directly: any recency effect, leniency, halo, or clustering at the safe middle? Making this a fixed step, not an afterthought, is what measurably reduces biased outcomes. Adjust any ratings the check exposes.

> **WATCH OUT** Treating bias as something to notice in passing lets it survive. Make the check a deliberate, scripted step.

7

Document and feed back

Record the agreed levels and the reasoning, update the matrix, and feed the results back to individuals. Keeping the rationale supports fair decisions, helps the next cycle, and means a contested score never has to be re-argued from scratch later.

> **WATCH OUT** Undocumented decisions get relitigated and forgotten. Capture what was agreed and why, every time.

AVOID THESE

Six mistakes in a calibration session

MISTAKE 01

No pre-work

Scoring from scratch in the room is the classic failure. Managers must arrive with ratings and evidence ready.

MISTAKE 02

No facilitator

Without a neutral guide, the standard is set by the loudest voice. Appoint someone to keep it fair and on track.

MISTAKE 03

Reviewing everything equally

Time spent on agreed scores is wasted. Triage to the outliers and the large gaps, where alignment is actually needed.

MISTAKE 04

Opinion over evidence

Settling disputes by seniority corrupts the data. Match the evidence to the anchored level, every time.

MISTAKE 05

Skipping the bias check

Bias survives when it is not named. Make recency, leniency and halo checks an explicit, scripted step.

MISTAKE 06

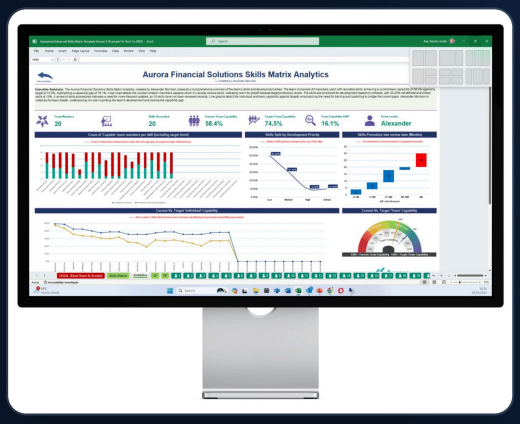
Forcing a distribution

Calibration is alignment, not fitting scores to a curve. Aim for consistent standards, not a preset spread.

CALIBRATION MADE EASY

The method is free. A ready-made matrix just gives calibration its *anchors and evidence.*

Everything here works with a spreadsheet and a meeting, and that is a fine place to start. A purpose-built template just makes calibration smoother: the 0 to 5 levels come defined as behavioural anchors, scores sit side by side for easy comparison, and the agreed results flow straight into capability and coverage, so the session has a shared reference and the outcome updates the matrix instantly.



The Advanced Excel Skills Matrix comes with the 0 to 5 levels defined as behavioural anchors, so a calibration session has a shared reference and the agreed scores update capability and coverage automatically, all on the same framework used throughout this guide.

COMMON QUESTIONS

Quick *answers*

Q What is a skills calibration session?

It is a structured meeting where managers compare and align their skill ratings so that a given level means the same across every team. They review scores against evidence, discuss the ones that disagree, check for bias, and agree adjustments, turning a set of individual judgements into consistent, comparable data.

Q How long should a calibration session take?

Usually 60 to 90 minutes. That is long enough to align standards and work through the outliers, but short enough to keep focus sharp; sessions much longer than that lose productivity. The key is that the scoring is done beforehand, so the meeting is spent on alignment, not rating from scratch.

Q Who should attend?

The managers whose ratings overlap or need comparing, plus a neutral facilitator, often an HR partner or senior leader, to guide the discussion and keep it fair. Keep the group small and focused; a large group aligns slowly, and the facilitator stops the most senior or insistent voice from setting the standard.

Q What is the most important rule?

Pre-work. Managers must arrive with their ratings completed and evidence ready, because the session is for discussion and decisions, not for scoring from scratch. When people turn up unprepared, calibration collapses into a slow group scoring exercise and never reaches its real purpose of aligning standards.

Q How do behavioural anchors help?

They give everyone a shared, observable reference for each level, so a contested score is settled by asking which described level the evidence fits, not by debate. Research credits behavioural anchors with reducing disagreement between raters by around 40%, which is why a clearly defined scale makes calibration so much easier.

Q How often should we calibrate?

Most organisations calibrate when they re-score, often quarterly or at each review cycle, with some running smaller "mini-calibrations" in between to stay aligned. What matters is consistency: making calibration a regular habit tied to your re-scoring rhythm, rather than a one-off, keeps standards aligned over time.

ABOUT THE AUTHOR

Dr Alex J. Martin-Smith

CMGR · MBA · LLM · DBA

Alex is the creator of the Upleashed capability framework that powers Skills Matrix Template, the award-winning Excel skills matrix. A Chartered Manager with an MBA, an LLM and a doctorate in business administration, he has spent more than two decades helping operations, HR and quality teams turn capability from a gut feel into something they can measure, manage and prove.

Connect on LinkedIn: linkedin.com/in/alexmartinsmith

Dr Alex J. Martin-Smith

SOURCES

Gartner. (2024). *Talent management research: Workforce skills data.* Gartner.

Martin-Smith, A. J. (n.d.). *The 0 to 5 capability framework.* Upleashed Limited. https://upleashed.com/capability-framework/

World Economic Forum. (2025). *The future of jobs report 2025.* World Economic Forum.

Make every score *mean the same.*

You now have the calibration method. The quickest way to start is to ask your managers to bring evidence-backed ratings to a 75-minute session, focus on the outliers, and run one explicit bias check before you finalise. That single hour is what turns your matrix from a set of opinions into data you can trust.