Level 3: Behavior

AI Literacy for the Modern Workforce · Kirkpatrick Evaluation Framework

Within the Kirkpatrick model, Level 3 measures whether participants apply what they learned to their on-the-job behavior. It is the tier that distinguishes knowing from doing.

It is therefore the tier that determines whether the program produced transfer. It is also the tier most AI training skips. The rest of this document is the instrument itself: the behavioral indicators, the two-instrument design, the rating scale, the interval schedule, and the analysis. The design measures what a participant does with AI, not how much they use it, sampled at 30, 60, and 90 days.

THE TRANSFER PROBLEM THIS LEVEL EXISTS TO CATCH

The L&D literature documents a persistent gap between learning and application. Philippa Hardman's synthesis of the transfer research estimates that only 10 to 20% of learning investment produces measurable on-the-job behavior change, against a global L&D spend in the hundreds of billions of dollars annually.[1] An evaluation that stops at Level 2 cannot see this leak at all; it certifies that knowledge was acquired and assumes the rest. This instrument is built to both measure the gap and create the conditions to close it: it adds indicators at 60 and 90 days that target the specific barriers (peer influence, manager reinforcement, organizational environment) where transfer tends to stall.

WHY USAGE VOLUME IS NOT BEHAVIORAL CHANGE

Through early 2026, a growing number of organizations measured AI adoption through usage volume: token consumption, login frequency, leaderboards ranking employees by how much they used a tool. The approach conflates activity with competency. A participant who generates thousands of tokens through single-turn, underspecified prompts and accepts first-generation output without verification registers as a high-usage employee on every volume metric, while demonstrating precisely the behaviors this program identifies as problematic: poor delegation, poor description, absent discernment, absent diligence. The research the program is built on shows why volume cannot proxy value: task-time savings range from 20% to 95% depending on the task (Tamkin & McCrory, 2025), and 57% of AI-touched tasks are augmentation while 43% are automation (Handa et al., 2025), so the value of an interaction depends entirely on what was delegated and how it was structured. The most competent user often shows lower consumption, because they decompose the task, prompt well enough to need fewer iterations, and route unreliable work back to their own hands.

The market reached the same conclusion the hard way. By late May 2026, Fortune declared "tokenmaxxing" over as a measure of AI return, after Uber's COO reported burning the company's entire annual AI budget in four months without a clean line to user value, Amazon deprecated an internal usage leaderboard (a senior engineering VP reportedly telling staff to "please don't use AI just for the sake of using AI"), and Meta quietly removed its own. These reversals confirm a design principle that predated them: an organization that incentivizes volume rewards the highest-risk behavior, uncritical high-volume delegation with no verification layer.

TWO INSTRUMENTS, ANCHORED TO EVIDENCE

A traditional observation checklist assumes a co-located manager who can watch the work. That assumption fails for distributed knowledge work, where structured prompting, citation verification, and task decomposition all happen invisibly at a workstation. The instrument addresses this with an evidence-informed parallel design:

- **Manager evidence review**, completed by the direct manager at each interval, rates each indicator from three sources: review of work artifacts (deliverables, documentation, shared prompt logs), a structured 15-to-20-minute check-in conversation, and any direct observation available. The manager is not assumed to have witnessed each behavior firsthand.
- **Participant self-assessment**, completed at each interval, rates the same indicators, but each rating must carry a specific recent example (situation, behavior, outcome). Fabricating a convincing, detailed example is harder than checking a frequency box, which is the primary control against self-report inflation.

Where the two agree, the change is well-evidenced. Where they diverge, the gap is the most useful signal in the instrument.

THE RATING SCALE

Both instruments use a four-point frequency scale with no neutral midpoint, which forces a directional call between "occasionally" and "consistently," and tops out at modeling for others, because peer modeling is how one person's competence becomes a team's.

RATING	LABEL	DEFINITION
1	**Not yet observed**	Not seen since completion. Not a negative judgment: may indicate insufficient opportunity, organizational barriers, or that the behavior needs more time.
2	**Occasionally**	Seen in some situations but not consistent; demonstrated when prompted or when conditions are favorable.
3	**Consistently**	Observed regularly across relevant work, applied as a standard part of the workflow without prompting.
4	**Modeling for others**	Applies it consistently and actively demonstrates it to colleagues, explains the reasoning, or coaches. Expected mainly at 60 and 90 days.

THE TWELVE BEHAVIORAL INDICATORS

The indicators map to the program's 4D competency framework (Delegation, Description, Discernment, Diligence) and to the program's action map. Each is phrased as a concrete, observable action, never a mindset: "verifies factual claims against independent sources before incorporating them" is measurable; "values accuracy" is not. Each is also classified by evidence type, which sets how much weight the manager review can carry: **artifact** (a reviewable work product, high confidence), **conversation** (elicited through the check-in, moderate confidence), or **self-report** (visible mainly to the participant, leaning on the specificity requirement). The eight core indicators run at all three intervals; transfer indicators are added at 60 days, organizational-influence indicators at 90.

ARTIFACT — HIGH CONFIDENCE **CONVERSATION** — MODERATE **SELF-REPORT** — SPECIFICITY-CONTROLLED

#	COMPETENCY	INDICATOR	EVIDENCE
B1	**Delegation**	Before using AI for a task, defines which components need human judgment and which are appropriate for AI.	CONVERSATION
B2	**Delegation**	Identifies when a task falls outside the model's reliable range and adjusts (less delegation, more verification, or manual work).	SELF-REPORT
B3	**Description**	Provides structured context, format, and quality criteria when prompting, rather than single-turn underspecified requests.	ARTIFACT
B4	**Description**	Iterates on outputs through multi-turn refinement rather than accepting or discarding first-generation results.	ARTIFACT

#	COMPETENCY	INDICATOR	EVIDENCE
B5	**Discernment**	Verifies factual claims, citations, and statistics against independent sources before incorporating them.	ARTIFACT
B6	**Discernment**	Evaluates the reasoning in an output, not just the product, checking for gaps, unsupported assumptions, and circular logic.	CONVERSATION
B7	**Diligence**	Documents AI's role in a deliverable, distinguishing AI-generated, AI-assisted, and human-authored components.	ARTIFACT
B8	**Diligence**	Takes full accountability for accuracy before sharing, applying the review standard used for a junior colleague's work.	SELF-REPORT
B9	**Transfer · 60d**	Discusses AI practices openly with colleagues, using the 4D vocabulary to describe how they work.	CONVERSATION
B10	**Transfer · 60d**	Has changed the AI approach on at least one regular task based on the program (more involvement where apt, less where risky).	CONVERSATION
B11	**Influence · 90d**	Has helped a colleague improve their AI practices, through coaching, sharing materials, or demonstrating techniques.	CONVERSATION
B12	**Influence · 90d**	Has raised a concern about output quality, data handling, or delegation appropriateness in a team setting.	CONVERSATION

Three of the twelve (B2, B8, and partly B1) are self-report by nature, because they describe cognitive processes that leave no artifact; the specificity requirement is their primary control. This graduated-confidence model is a transparency mechanism: organizations should weight artifact-anchored claims heavily and temper self-report claims accordingly, which is more honest than treating all twelve as equally observable.

PROGRESSIVE INTERVALS AND CONVERGENCE

INTERVAL	INDICATORS	FOCUS
30 days	B1–B8 (core)	Individual baseline: are the most actionable practices taking hold?
60 days	B1–B10	Transfer: is the participant adapting their workflow and discussing practices openly?
90 days	B1–B12	Influence: is the change propagating to peers and team practice?

The core indicators produce a three-point trend line. The expected pattern is progressive adoption; an indicator that rises from 30 to 60 days but falls by 90 signals a behavior adopted and then crowded out, a different problem than one never adopted, so the 90-day interval is a sustainability check rather than a delayed re-measurement. The 60-day open-discussion indicator targets the concealment dynamic directly: the Anthropic Interviewer study found 69% of professionals conceal their AI use at work, and a program that builds private skill while ignoring the social conditions for visible use produces no change an organization can measure.

The parallel design produces a manager rating and a self rating for each indicator, and the four agreement patterns each carry distinct meaning. Both high is the target. Manager-high/self-low usually means the participant has internalized the practice to the point it feels unremarkable. Manager-low/self-high is the most actionable: the behavior may be happening invisibly (a legibility problem), overstated (a self-report problem), or unrecognized by the manager (an orientation problem), and each points to a

different fix. Both low, cross-referenced against the Level 2 data, separates a transfer failure (knowledge present, behavior absent) from an upstream knowledge gap.

LIMITATIONS, AND ALTERNATIVES

The instrument has more than one weak point. The most obvious is the manager: it only works if managers spend real attention on it, reading the artifacts and the written examples rather than speed-running the rating scale, and a manager who treats it as one more compliance checkbox produces confident noise. The participant side is just as fragile. Self-assessment depends on honest reporting, and the 69% concealment finding cuts straight against it: where AI use carries social stigma, people under-report what they actually do, and a well-meaning L&D team reading the aggregate may have no idea how much pressure sits on the ground beneath the numbers. The data can look clean while the culture it came from is not.

The instrument also fails quietly when it is bolted onto already-full days. If the manager review and the self-assessment become one more task on a manager's or an employee's plate, with no protected time to reflect and engage honestly, both default to whatever rating clears the form fastest. Two further risks compound these: evaluation apprehension, where a participant shapes a self-report to look good to the person who rates them, especially if the check-in reads as appraisal rather than development; and uneven evidence, since the artifact-anchored indicators assume an organization where deliverables and prompt logs are actually reviewable, which is not every organization.

None of this is fatal, but it does mean the framework should not be deployed naively. Some mitigation is structural: protect real time for the check-in, frame it as development rather than appraisal, and lean on more than one measurement surface so no single blind spot decides the result. To that end the framework offers alternatives a deploying organization can layer in or substitute: sampled artifact audits, where an L&D reviewer or quality lead scores a random sample of AI-assisted deliverables against the same indicators; peer review, where colleagues close enough to the work can see it; and self-assessment with spot-checks, where specific examples are verified against real artifacts rather than taken on faith. No single surface is bias-free, but each fails differently, and the more independent the surfaces, the less any one of them decides the outcome.[2]

1 *The transfer estimate and the global L&D-spend figure are widely cited but should be read as directional, since both depend on how "behavior change" and "L&D spend" are defined across organizations.*

2 *This is an evaluation instrument design, not a results report. No cohort has run these instruments; the program is an independent portfolio project not yet deployed at organizational scale. The indicators, rating scale, evidence-type classifications, interval design, and convergence protocol are complete and ready for a deploying organization to administer through its own HR or performance-management infrastructure. Attribution also remains inherently limited: behavioral change at 30, 60, or 90 days cannot be definitively separated from concurrent factors such as policy changes or independent learning.*

REFERENCES

Kirkpatrick Partners — *The Kirkpatrick Model*

Anthropic — *AI Fluency: Framework and Foundations* (the 4D competency framework)

Anthropic Interviewer — *What 1,250 Professionals Told Us About Working with AI* (Dec 2025) (69% concealment finding)

Tamkin, A. & McCrory, P. — *Estimating AI Productivity Gains from Claude Conversations* (Nov 2025)

Handa, K. et al. / Anthropic — *Anthropic Economic Index* (augmentation-automation split)

Fortune — *Tokenmaxxing is over* and *Uber's AI budget burn* (May 2026)

Hardman, P. — synthesis of the learning-transfer literature (transfer rate and L&D-spend estimates)