# Level 1: Reaction

AI Literacy for the Modern Workforce · Kirkpatrick Evaluation Framework

**Within the Kirkpatrick model, Level 1 measures participant reaction: whether learners found the program relevant and applicable to their work.**

It is the earliest and most accessible evaluation tier, administered immediately after completion and before any on-the-job behavior can be observed. It is also the tier most often wasted, because the default reaction survey collects satisfaction that maps to no decision. The rest of this document is the instrument itself, from the item bank through administration to analysis. The single constraint behind every design choice is actionability: each item must produce data that drives a specific program change.

## WHAT LEVEL 1 MEASURES

Level 1 captures the participant's subjective assessment of the learning experience across four dimensions: perceived relevance to their role, confidence change in working with AI tools, content quality and pacing, and intent to apply or recommend the program. Each dimension was selected because it produces data tied to a specific improvement action. Generic satisfaction metrics ("I enjoyed the program") are excluded because they cannot be acted on.

Reaction data cannot tell whether participants learned the content (Level 2), changed their on-the-job behavior (Level 3), or produced organizational results (Level 4). A participant who rates the program highly may not have acquired the target competencies; one who rates it poorly may still have learned effectively. The correlation between satisfaction and knowledge gain is weak across the L&D literature, which is the entire reason this framework does not stop here. Level 1 diagnoses program design; it does not score individual competency.

## THE ACTIONABILITY REQUIREMENT

Every item was designed against one criterion: can the response drive a specific improvement decision? "How would you rate this program overall?" produces a number that traces to no design feature. "The scenarios in this program reflected situations I encounter in my role" maps directly to the scenario-design layer and triggers a concrete revision if it falls below threshold. That distinction separates a management-grade evaluation instrument from a compliance-grade satisfaction survey. The instrument deliberately excludes items it cannot act on: facilitator effectiveness (the program is self-paced, with no facilitator), platform satisfaction (better handled through UX research), and organizational support for AI adoption (outside the program's control).

## WHY THIS STAYS A DESIGN

This page presents the proposed item bank, the candidate questions a deploying organization would draw from, not a running survey. The instrument is intentionally not live, for two reasons.

First, the course is built for the general workforce, and a reaction survey only works when it fits the institution administering it. The items below are a calibrated starting point, but their relevance, phrasing, and scope would need tuning to a specific organization's context, roles, and reporting conventions before any response is worth collecting. Shipping a fixed, generic survey as if it were ready to run would misrepresent how reaction measurement actually works.

Second, privacy. The platform collects no personally identifiable information; all learner progress (module completion, knowledge-check responses, pre/post scores) is persisted only in the browser's local storage and never reaches a server. As a solo-built portfolio project, it has no data processing agreement, no retention policy, and no legal entity to bear data-controller liability. A

reaction survey with free-text fields, which can carry identifying detail and opinions about an employer, is precisely the data this architecture should not hold.

So the instrument is designed for external administration. A deploying organization adapts the item bank to its context and runs it through its own survey infrastructure (internal tools, a third-party platform, or an HRIS survey module), which already carries access management, retention policies, anonymization, and regional compliance. The recommended trigger is the platform's completion screen (Module 4, Section 10), where the participant sees their pre/post results: immediate post-completion is the window with the highest response rate and the freshest recall.

## SURVEY DIMENSIONS

The survey runs 8 to 13 scaled items plus 3 to 4 open-response prompts, calibrated to the deploying organization's tolerance for length (completion tends to drop past about 15 items on a voluntary post-program survey). Each dimension maps to a category of design decision:

| DIMENSION | WHAT IT MEASURES | DESIGN DECISIONS IT INFORMS | ITEMS |
|---|---|---|---|
| **Perceived relevance** | Whether the content applied to the participant's actual professional context | Scenario selection, industry representation, role targeting, example specificity | 2–3 scaled + 1 open |
| **Confidence change** | Whether participants feel more able to evaluate and work with AI outputs | Instructional depth, practice-activity design, scaffolding adequacy | 2–3 scaled + 1 open |
| **Content quality** | Whether pacing, difficulty, and structure matched expectations and needs | Module sequencing, section length, vocabulary calibration, interactive balance | 2–4 scaled + 1 open |
| **Intent to apply** | Whether participants intend to change behavior and would recommend the program | Value proposition, transfer-design effectiveness, deployment strategy | 2–3 scaled + 1 open |

## THE ITEM BANK

Scaled items use a 5-point Likert agreement scale (Strongly Disagree to Strongly Agree). Open-response items sit at the end of each dimension so they capture qualitative context without priming the scaled responses.

**Perceived relevance to role**

| # | ITEM | FORMAT |
|---|---|---|
| R1 | The scenarios in this program reflected situations I encounter in my professional role. | 5-point Likert |
| R2 | After completing this program, I can identify specific tasks in my work where the concepts apply. | 5-point Likert |
| R3 | The program addressed AI-related challenges relevant to my industry or function. | 5-point Likert |
| R4 | Which module or topic felt most relevant to your current role, and why? | Open response |

**Confidence change.** Its items map to the program's 4D competency framework: C1 to Discernment (output evaluation), C2 to Description (prompt construction), and C3 to the mechanistic understanding under both; C4 surfaces residual uncertainty that

informs supplemental resources or manager talking points for the Level 3 window.

| # | ITEM | FORMAT |
|---|---|---|
| C1 | I feel more confident in my ability to evaluate whether an AI-generated output is reliable. | 5-point Likert |
| C2 | I feel more confident in my ability to write effective prompts that produce useful outputs. | 5-point Likert |
| C3 | I feel more confident in my ability to explain to a colleague why an AI tool produced an incorrect or misleading result. | 5-point Likert |
| C4 | What, if anything, do you still feel uncertain about when working with AI tools? | Open response |

**Content quality and pacing**

| # | ITEM | FORMAT |
|---|---|---|
| Q1 | The program was paced appropriately for my level of prior knowledge about AI. | 5-point Likert |
| Q2 | The interactive elements (dashboards, simulations, sandbox activities) helped me understand the concepts better than text alone would have. | 5-point Likert |
| Q3 | The program used clear language that I could follow without a technical background. | 5-point Likert |
| Q4 | The program length was appropriate for the depth of content covered. | 5-point Likert |
| Q5 | Which section or activity, if any, felt too fast, too slow, or unclear? Please describe. | Open response |

**Intent to apply and recommend.** A1 is a leading indicator for the Level 3 behavioral evaluation; A2 and A3 give two recommendation measures (a comparable agreement item and a standard NPS item); A4 names the specific change the participant commits to, which feeds the Level 3 manager checklist.

| # | ITEM | FORMAT |
|---|---|---|
| A1 | I intend to change at least one aspect of how I work with AI tools based on what I learned. | 5-point Likert |
| A2 | I would recommend this program to a colleague in a similar role. | 5-point Likert |
| A3 | On a scale of 0–10, how likely are you to recommend this program to a colleague? | NPS (0–10) |
| A4 | What is the single most important change you plan to make in how you work with AI tools? | Open response |

## ADMINISTRATION PROTOCOL

**Timing.** Administer immediately after completion, triggered by the program closing section. Delays beyond 48 hours produce lower response rates and reconstructed rather than recalled impressions.

**Distribution.** An embedded link in the post-completion communication (an LMS-triggered email, an intranet notification, or a direct link on the completion screen). The link should appear at the moment of completion, not in a separate follow-up days later.

**Anonymity.** Administer anonymously to maximize candor, particularly on the confidence and quality dimensions where social desirability inflates scores. If an organization needs to link Level 1 reactions to the same individual's Level 3 data, use a coded identifier rather than a name or email, and disclose the linking in the survey introduction. Identified responses enable longitudinal analysis but may suppress critical feedback.

**Response rate.** Target a minimum 60% response rate for cohort-level interpretation. Below that, non-response bias becomes a real concern, since strong positive and negative reactors tend to respond while moderate reactors do not, producing a bimodal distribution that misrepresents the cohort.

## READING THE DATA

Each scaled item is analyzed for mean, standard deviation, and distribution, because a mean of 3.5 with SD 1.4 tells a different story than the same mean at SD 0.6. Directional thresholds guide interpretation rather than fixed pass/fail cutoffs:

| MEAN | SIGNAL | ACTION |
|---|---|---|
| 4.0–5.0 | **Strong** | The design feature works as intended. Monitor across cohorts; no revision needed. |
| 3.0–3.9 | **Mixed** | Acceptable, not strong. Review the open-response data for friction; revise if it does not improve across 1–2 cohorts. |
| Below 3.0 | **Concern** | Not meeting expectations. Prioritize revision; cross-reference the open responses and the Level 2 per-construct analysis to locate the cause. |

If A3 is included, the Net Promoter Score is promoters (9–10) minus detractors (0–6), with passives (7–8) excluded; above +30 is generally strong for corporate training, below 0 signals real dissatisfaction. Open responses are thematically coded (two reviewers, inductive on the first cohort, then a deductive codebook) into a frequency-ranked theme list that maps to revision priorities. The full value emerges across cohorts: when a revision is made between cohorts, the next cohort's reaction data shows whether it worked. That feedback loop is the instrument's primary use: one cohort's reaction data drives a revision, and the next cohort's data shows whether it worked.

## INTEGRATION WITH THE FRAMEWORK

Level 1 sets the context for the levels above it. A cohort that scores well on the Level 2 pre/post but rates relevance poorly has the knowledge but cannot see how to apply it: the scenarios need stronger workplace connection. A cohort that rates the program highly but shows little pre/post improvement enjoyed the experience without changing its knowledge state: engaging but not challenging enough. And items A1 and A4 bridge to Level 3, where the manager review measures whether the stated intent became behavior. The gap between intent and action is itself a diagnostic, one that informs the transfer supports (job aids, manager coaching prompts, peer accountability) that turn intention into practice.[1]

1 *This is an evaluation instrument design, not a results report. The program is an independent portfolio project that has not been deployed at organizational scale, so no Level 1 cohort data exists. The page presents the proposed item bank; the administration protocol and analysis thresholds are complete, and a deploying organization would adapt the items to its own context before administering the survey through its own infrastructure.*