Cheat Sheet: Four Levels of Evaluation

My blog post on the Four Levels of Evaluation explains how you know if your course is a success. The purpose of this sheet is to help you construct your own course evaluation beyond the first level. Use at least one of the suggestions for each level you choose; the level you choose will likely reflect your role in the course (see page 4).

I'll walk through Levels 1–4 using prompts, then I'll give you a tiny starter move so this doesn't feel overwhelming.

Level 1: Reaction (the "smile sheet")

What it tells you

How learners experienced the training: clarity, relevance, usability, pacing.

What counts as evidence

Perception. Useful diagnostic input, not proof of learning or performance.

What is commonly measured

A post-course survey can be designed to capture more than reaction, but most aren't.

- ❒ Satisfaction ratings
- ❒ Instructor ratings
- ❒ Perceived relevance
- ❒ "Would recommend"
- ❒ Perceived confidence

Methods for measuring

- ❒ Post-course survey (sometimes dubbed a "smile sheet")
- ❒ Pulse poll
- ❒ Open comments
- ❒ "One thing to keep/one thing to change."

Most common mistake

Treating Level 1 as the conclusion. A course can be pleasant and still ineffective.

Level 2: Learning (can they do it now?)

What it tells you

Whether learners gained the knowledge, skill, or judgment the course promised.

What counts as evidence

A demonstration of competence immediately after training—what they can do, decide, or explain.

What is commonly measured

- ❒ Appropriate decisions in short scenarios.
- ❒ Accurate interpretation of key indicators, alerts, or outputs.
- ❒ Adherence to the critical steps in a process (including safety steps).
- ❒ Appropriate selection of settings/parameters based on a defined case.
- ❒ Sound troubleshooting choices using the approved pathway.

Methods for measuring

- ❒ Scenario-based multiple-choice questions (MCQs).
- ❒ "Spot the error" items using a realistic case or device setup.
- ❒ Short-answer prompts: "What would you do next, and why?"
- ❒ Mini-case interpretations.
- ❒ Simulation exercises.
- ❒ Checklist-based return demos when a procedure matters.

Most common mistake

Using verbs you can't measure. "Understand," "know," "be aware of," "be familiar with," and "feel confident" *cannot be measured*. Verbs such as "discuss," "describe," and "list" don't move any dial in the real world. If the goal is performance, Level 2 must require application—decision-making, problem-solving, and choosing the next best step—not a vocabulary recital. Replace vague verbs with observable actions, or your Level 2 data will be mush.

Level 3: Behavior (learning transfer)

What it tells you

Whether learners apply what they learned in real work, under real constraints. *Competence shows up in performance*. Level 2 can tell you what learners can do at the end of training. Level 3 tells you whether they do it when it counts.

What counts as evidence

Observable, verifiable behavior in the workflow—what people actually do when it's busy.

What is commonly measured

- ❒ Reduced workarounds and more consistent use of the standard process.
- ❒ Adherence to Instructions for Use (IFU) or Standard of Practice (SOP) steps when it matters most.
- ❒ Appropriate settings/parameters selected in real cases, not just in training.
- ❒ Troubleshooting follows the approved pathway before escalation.
- ❒ Documentation reflects correct action and decision-making.
- ❒ Escalation occurs based on criteria, not on hunches.

Methods for measuring

- ❒ Direct observation with a short checklist.
- ❒ Documentation or chart audits tied to the behaviors you trained.
- ❒ Workflow compliance reports.
- ❒ Device or system logs that reveal real-world choices and patterns.
- ❒ Manager or mentor verification at 30/60/90 days, using defined criteria.
- ❒ Support ticket tagging by reason for call or root cause.

Most common mistakes

Not recognizing Level 3 as competence in action. In competency-based education (CBE), competence isn't just "passed the test."

The other big mistake is assuming behavior changed without validating it. Pick a method e.g., observation, audit, logs, verification.

Level 4: Results (did it change anything that matters?)

What it tells you

Whether training influenced outcomes the organization cares about: quality, safety, time, return on investment (ROI), and risk.

What counts as evidence

A meaningful metric shifts over time in the expected direction, plausibly linked to the behaviors the training targeted.

What is commonly measured

- ❒ Reduced errors, incidents, or near-misses.
- ❒ Reduced rework and fewer "do-overs."
- ❒ Faster time-to-competence/time-to-independence.
- ❒ Fewer escalations, fewer support requests, and fewer repeat requests.
- ❒ Reduced downtime and fewer workflow interruptions tied to user error.

- ❒ Improved quality metrics tied to correct use and correct decisions.
- ❒ Reduced risk—including incident risk and legal exposure tied to misuse, non-adherence to IFU, or inconsistent practice.
- ❒ ROI: cost avoidance, efficiency gains, revenue protection, reduced support burden.
- ❒ For medical devices: improved adoption and usage rates, increased utilization of key features, and (when training is part of rollout) stronger uptake of a new model—plausible outcomes when paired with Level 3 evidence.

Methods for measuring

- ❒ Operational dashboards and Key Performance Indicator (KPI) trends.
- ❒ Quality and safety reports.
- ❒ Incident reporting trends.
- ❒ Support analytics and call reason analysis.
- ❒ Time-to-competence tracking.
- ❒ ROI reporting: cost avoidance, support cost reduction, efficiency gains, revenue protection.
- ❒ Adoption and usage analytics (product, platform, or device telemetry where available).

Most common mistake

Attribution fantasy. Training rarely acts alone. The cleanest story combines Level 3 behavior evidence with Level 4 trend data: behavior changed, and the downstream metric moved in the same direction.

Tiny starter move: Build a simple plan in five minutes

If you want a lightweight version of training evaluation beyond satisfaction surveys, do this.

Name your role:

- ❒ owner, director, product manager, education manager
- ❒ instructional designer
- ❒ presenter
- ❒ sponsor
- ❒ stakeholder

Write one sentence: "This course is successful if _____."

Now pick anchors for Levels 2, 3, and 4.

Level 2 anchor (Learning):

What will learners be able to *do* immediately after the training that they could not do before?

Examples:

- ❒ Choose the best next action in this scenario (and justify your choice).
- ❒ Prioritize what you would do first, and what you would do next.
- ❒ Identify the misuse in a device setup.
- ❒ Choose the next best step—and explain your reasoning.

Level 3 anchor (Behavior)

What on-the-job behavior should change in the real work setting? What should we be able to observe or verify?

Examples

- ❒ Follow IFU steps without skipping critical steps.
- ❒ Use the approved troubleshooting pathway before escalating.
- ❒ Document the setting choice and rationale consistently.

Level 4 anchor (Results):

What measurable outcome should that behavior influence over time—quality, safety, time, return on investment (ROI), or risk?

Examples

- ❒ Fewer support requests tagged "setup error."
- ❒ Reduced rework.
- ❒ Fewer near-misses and reduced legal exposure tied to misuse.