McLean Performance Group

Insights · Foundational

Beyond the Checkbox

Why most training programs measure the wrong thing

By Adam J. McLean, PhD — Founder & Principal, McLean Performance Group

Service-Disabled Veteran-Owned Small Business · Madison, Alabama

The moment usually comes about forty-five minutes into a compliance review. The auditor — federal contracting officer, Joint Commission surveyor, insurance-carrier underwriter, plaintiff's counsel in deposition — sets aside the attendance roster and asks the question that has no good answer.

Can you show me evidence that this training changed how your people actually do the work?

The file folder produces sign-in sheets. It produces exam scores from the post-test. It produces a slide deck with a copyright notice on the cover. It does not produce evidence that the trained behavior surfaced on the operations floor on a Tuesday afternoon. It does not produce a measurement of whether the patient outcome shifted, the franchisee variance closed, or the CPARS narrative ratcheted a level. The folder is full. It is also empty.

This essay is about the gap between those two facts, and what it costs.

The four levels, briefly

Donald Kirkpatrick described four levels of training evaluation in 1959. Sixty-seven years later, his model is still the operative framework in every serious training-compliance conversation — federal, healthcare, franchise, fitness, manufacturing, public safety. The model is short enough to state in a paragraph.

Level 1 is reaction. Did the learner like it? (Smile sheets.) Level 2 is learning. Did the learner acquire the knowledge or skill the program intended? (Exam scores.) Level 3 is behavior. Is the learner now doing the work differently — on the job, in conditions of normal pressure? Level 4 is results. Did business outcomes change — productivity, quality, safety, retention, revenue, cost — in a way that traces back to the training?

Levels 1 and 2 are inputs to the system. Levels 3 and 4 are outputs of the system. The audit defense lives at Levels 3 and 4. So does the return on investment.

Why most organizations stop at Level 2

The reasons are predictable. Levels 1 and 2 are cheap. The instruments are off-the-shelf. A learning management system handles smile sheets and post-test scoring without configuration. The data exports cleanly. A quarterly board deck full of Level 1 and Level 2 charts looks like measurement, because it is measurement — of the wrong things.

Levels 3 and 4 are expensive in a particular way. They require instrumentation in the work, not in the classroom. Direct observation. Workplace audit checklists. Manager verification with a calibrated rubric. Patient or customer outcome data linked back to the training cohort. Pre-post comparisons of operational metrics on a defensible attribution model. None of this comes free, and none of it comes from the LMS vendor.

The third reason is more comfortable than either of the first two: a quiet organizational belief that Level 2 is enough. That if the learner passed the post-test, the program worked. That if the file folder is full, the audit will pass. That belief is wrong, and the cost of being wrong is the subject of the next section.

What stopping at Level 2 actually costs

The cost shows up in four places. Each one is invisible until it isn't.

The audit you fail

The compliance authorities that govern serious training programs — DCMA surveillance against a federal performance work statement, Joint Commission HR.01.06.01 in a hospital, a state HCBS audit in senior care, a franchise insurance carrier's annual renewal questionnaire, an OSHA inspection in a multi-unit fitness operation — do not stop at the post-test score. They ask for evidence that the trained behavior actually surfaced. When the file folder cannot produce that evidence, the finding is not "no training." The finding is worse. The finding is "documented training, no demonstrated capability." That finding is what drives a CPARS downgrade, a conditional accreditation, a non-renewal, a citation, a deposition exhibit.

The turnover you absorb

In acute care nursing, NSI's 2026 *National Health Care Retention & RN Staffing Report* put bedside RN turnover cost at $60,090 per departed nurse, with average hospitals absorbing $4.2M to $6.2M annually. Training that does not change behavior on the unit does not retain nurses. Training that does — that produces a calibrated, supported, accountable practitioner — does. The financial line moves. The same arithmetic applies, with different denominators, in federal program staffing, franchise unit-operator retention, and fitness studio instructor longevity. The Kirkpatrick gap is also a financial gap.

The regulatory drift you absorb without knowing it

When a training program measures only attendance and exam scores, it cannot detect skill decay. It cannot detect divergence between what the standard says and what the workforce actually does. Drift accumulates silently between audits, then surfaces as a finding pattern, then surfaces as a sanction. The organization's defense is the same flat file folder. The standard moved; the file folder did not.

The return on investment you cannot prove

Training is one of the largest discretionary lines in most operating budgets. It is also one of the least defensible at the budget meeting, because the case for it terminates at attendance and exam scores. When the CFO asks what the program returned, the Level 1 / Level 2 dashboard cannot answer. The program survives at the political margin until the first budget pressure, and then it does not. The training that survives is the training that proves its results.

What Level 3 and Level 4 look like in practice

The good news is that Levels 3 and 4 are not aspirational. They are measurable on a cadence with the right instruments. Four methods carry most of the Level 3 weight.

Direct observation. A calibrated observer watches the trained worker perform the trained task under operational conditions, scoring against a rubric the training program produced. Cadence varies by setting — monthly in high-risk healthcare specialties, quarterly in most federal programs, semi-annually in franchise field operations. The observer is calibrated against a known standard, not improvised.

Workplace audit. The work product, the chart, the operations log, the customer interaction — sampled and scored. The sample frame is defensible. The scoring criteria are written. The trend line is reviewed at a fixed interval and the deviations are escalated through a defined loop.

Manager verification. The first-line supervisor signs a competency confirmation against a structured rubric on a fixed schedule. The rubric is not the manager's opinion; it is the program's behavioral checklist,

with explicit criteria, anchored examples, and a calibration session at the start of each cycle.

Patient, customer, or operational-outcome correlation. The downstream metric — readmission rate, customer satisfaction, defect rate, time-to-resolve, safety event frequency — is linked to the training cohort, the trainer cohort, and the calendar. The attribution is conservative; the trend is honest; the cycle is repeated.

Level 4 follows from Level 3. If the behavior changed, the business result either changed too or the link was not what the program claimed. Level 4 instrumentation builds the financial line from the behavioral signal: turnover reduction in dollars, defect reduction in dollars, audit-finding reduction in dollars, time-to-competency reduction in dollars. The Level 4 case is conservative by design — it does not claim everything, because claiming everything is the failure mode that loses credibility. It claims what the data supports, and it does so with the supporting math attached.

A different way to think about training

Most organizations treat training as a delivered product. You buy it; you deliver it; you log the delivery; you move on. That framing is what produces the Level 1 / Level 2 trap. The program is judged by whether the delivery happened, not by whether the delivery did anything.

A different framing produces different programs. **Training is a measured intervention.** The point of the intervention is to change a behavior under operational conditions, and the change is verifiable on a cadence. The post-test is not the end of the program; it is a checkpoint on the way to the Level 3 measurement, which is on the way to the Level 4 outcome. Every artifact in the file folder traces forward to that outcome and backward to the program standard. The folder is no longer a paper trail. It is an instrument of defense.

The organizations that adopt this framing do not abandon Levels 1 and 2. They keep them. They keep them because compliance authorities still require them, and because they remain useful as early signals. What changes is where the program's center of gravity sits. The center moves down the model, from the classroom to the work. The dashboard moves with it.

The architecture under the model

McLean Performance Group designs the training compliance architecture that satisfies external audit and proves measurable ROI at Kirkpatrick Levels 3 and 4. The seven-document MPG Framework — master program index, master training syllabus, instructor certification, instructor utilization policy, evidence-based improvement framework, training effectiveness measurement program, industry tailoring guide — was built simultaneously against compliance authorities by vertical and against the Level 3 / Level 4 evaluation model. Not sequentially. Simultaneously. That is the design choice that distinguishes a defensible program from a documented one.

The entry product is the Program Audit. The audit reviews the existing training program against the seven-document framework and the compliance authorities that govern your vertical, identifies the gap between attended training and demonstrated capability, and produces a written engagement plan that closes the gap. It is a fixed-fee, principal-delivered engagement, scoped in a short discovery conversation.

If you can already prove, on a cadence, that your training drove behavior change and produced business results — you do not need McLean Performance Group. If you cannot, the next audit is coming, and the

file folder is still full.

> *Most training programs measure attendance. The right architecture measures whether training drove behavior change and produced business results — that's where ROI lives, and that's where audit-defensibility lives.*

— **Adam J. McLean, PhD** Founder & Principal, McLean Performance Group Adam@McLeanPerformanceGroup.com · 256-270-6582