From Instrument to Iteration: How Phase 1 Evaluation Drove Program Design

A companion piece to "Designing for Behavior Change: A Cohort-Based Leadership Development Framework"

Chris Bailey
Founder & Learning Development Consultant
Bailey Learning Works, LLC

Why This Document Exists

Evaluation instruments are common in leadership development programs. What's less common is evidence that anyone did anything with the results.

This document walks through the evaluation design decisions behind the Phase 1 assessment used in the Regional Community Leadership Program's third cohort – what the instrument was built to capture, how responses were systematically analyzed, and specifically what changed in program design as a result.

The instrument itself is included as an appendix. What follows is the thinking behind it.

The Evaluation Design Problem

Kirkpatrick's model is widely cited in L&D. Kirkpatrick Level 3 – measuring whether participants actually changed their behavior because of a program – is rarely implemented well. The reason isn't ignorance. It's design sequencing.

Most programs build evaluation instruments after the program is designed, treating measurement as an administrative task rather than an instructional one. The result is end-of-program satisfaction surveys that confirm participants enjoyed themselves and reveal nothing about whether the program worked.

Measuring behavior change requires answering a specific question *before the program begins:* what would it look like for a participant to lead differently because of what this program taught? If you can't define that in advance, you can't capture it afterward.

The Phase 1 evaluation instrument was built around that question from the start.

What the Instrument Was Designed to Capture

The assessment served three simultaneous purposes, each informing a different design decision:

Participant self-awareness – prompting participants to identify and articulate what they were actually doing differently, not just what they had learned. Skills confidence ratings without behavioral examples are Level 2 measurement at best. The instrument required both.

Program feedback for real-time iteration – structured questions about what was working, what wasn't, and what participants needed more of going into the next phase. This wasn't post-program evaluation. It was formative assessment administered at the phase transition point specifically to allow curriculum adjustment before Phase 2 began.

Baseline data for longitudinal comparison – confidence ratings and behavioral change examples collected at the end of Phase 1 established a reference point for Phase 2 and Phase 3 assessments. Progress becomes visible only when you have something to measure it against.

The instrument was administered at the Phase 1 transition session – the final session of the September–November sequence – with 40 minutes allocated for completion. Participants were explicitly framed that honest, critical feedback served the program's improvement; evaluator notes emphasized not influencing responses.

Three Design Decisions Worth Examining

1. Behavior change examples as primary evidence, not supplementary data

Most evaluation instruments treat open-ended responses as qualitative texture around quantitative ratings. The Phase 1 instrument inverted that relationship. Section 2A asked participants to describe *one specific example* of something they were doing differently – in their work, their community involvement, or their personal life – with enough specificity that someone who wasn't in the room could evaluate whether real change had occurred.

This produced the evidentiary layer that satisfaction scores cannot. Participants weren't asked whether the program was valuable. They were asked to demonstrate

that it had changed something. The responses that resulted – structured weekly team check-ins implemented within 30 days, difficult conversations with struggling staff members initiated, proactive one-on-ones with supervisors that hadn't existed before – are what Kirkpatrick Level 3 looks like in practice.

The instrument design made that evidence collectable. Without the prompt structure, those data points don't exist.

2. Skills application with required examples, not self-report alone

Section 1B asked participants to check which skills they had applied AND provide a brief example for each. The conjunction is the design decision. Checking a box is confirmation bias. Providing a concrete example is evidence.

This structure also revealed something the confidence ratings obscured: participants who reported high confidence in a skill area sometimes struggled to produce a specific application example, while participants reporting moderate confidence provided detailed, concrete evidence of consistent practice. The examples were more informative than the ratings.

3. Forward-looking questions as formative data, not courtesy

Section 4 – asking what participants wanted to develop in the December–May phase and what concerns they had – was not a courtesy question. It was formative assessment designed to inform Phase 2 curriculum before that phase began.

The responses were treated as design inputs, not participant wishes. What emerged from the analysis shaped specific Phase 2 content decisions.

What the Analysis Revealed

Responses from ten participants were analyzed individually before being synthesized at the cohort level. Individual analyses examined application patterns, confidence trajectories, behavioral change examples, and program feedback. Cross-cohort synthesis identified patterns, outliers, and design implications.

Three findings shaped the most significant Phase 2 design adjustments:

Finding 1: Conflict resolution had emerged as the primary development need.

Conflict was mentioned across multiple participant responses – not as a program topic, but as an active challenge they were navigating without sufficient tools. Two participants identified conflict resolution as their single most important development need. One participant named it in her skills application, her goal progress, her application challenges, her program improvement suggestions, her session-specific feedback, and her forward-looking priorities – six separate places across one assessment. That's not a preference. That's a diagnostic signal.

Design response: Conflict resolution was elevated from a supporting element in the February session to its organizing focus, with explicit practice scenarios added to create what one participant had specifically requested: "a smooth conflict resolution experience" that could build confidence through successful application before the stakes were higher.

Finding 2: The VUCA framework needed different scaffolding for different learners.

One participant found the VUCA framework integrative and named it as her most valuable November learning. Another was confused by it despite pre-reading the materials, specifically noting: "Explain where you are going with the VUCA. I was a little lost even with reading ahead of time." Both responses were valid. The framework was being introduced abstractly before participants had generated their own examples of volatility, uncertainty, complexity, and ambiguity from their actual work contexts.

Design response: The November session sequence was restructured to move from participant experience to framework rather than framework to participant experience. Participants would generate examples from their own leadership contexts before the VUCA acronym was introduced, grounding an abstract model in concrete situations before naming it. Facilitator notes were updated to flag this as a discovery sequence, not a teaching sequence.

Finding 3: Participants were navigating widely varying workplace contexts, including toxic ones.

The analysis revealed something the program design hadn't explicitly accounted for: several participants were applying leadership development skills in actively

dysfunctional organizational cultures. One was building what she called a "workplace oasis" for her team within an institution where administrators modeled the opposite of what the program was teaching. Another was managing in a high-turnover environment with chronic understaffing and peer culture that undermined new hires. A third was navigating job insecurity and overt mistreatment.

These participants were not failing to apply what they'd learned. They were applying it under conditions that required significantly more skill, not less. The program design had implicitly assumed organizational contexts that were at least neutral, if not supportive.

Design response: Phase 2 content was adjusted to explicitly address leading within organizational systems that resist the behaviors the program was developing. This wasn't a new module. It was a reframing of existing content on trust-building and influence that named the context in which many participants were actually operating.

Three Participant Snapshots

The following examples are drawn from actual participant assessments, anonymized to protect participant identity. They illustrate the range of application patterns the evaluation instrument was designed to surface.

Participant A applied all seven Phase 1 skills and provided detailed examples for each. Her most significant behavioral change: a cognitive reframe about staff turnover that shifted her from experiencing departures as personal failure to understanding them as culture-fit information. "I have learned that it is ok when people leave as they may not be a good fit for the position or the culture. If they stay it can bring down morale." She translated this reframe into a three-part action plan: hire more staff to build organizational buffer, conduct individual meetings with current staff proactively, and build rapport as a retention strategy. Her October goal – improving conflict resolution – was the subject of repeated requests across every section of her assessment. Five concrete staffing and team development goals were listed in a single response.

This participant's assessment is what comprehensive Phase 1 integration looks like: frameworks internalized, applied to real challenges, generating specific behavioral change and forward-looking strategy.

Participant B also applied all seven skills, with a distinctive pattern: she was using the program's frameworks as tools for self-compassion as much as leadership practice. The VUCA framework specifically gave her language for her organizational context – what she described as "big, weighty work" – and permission to "be patient with myself" in navigating it. Her most significant behavioral change was boundary-setting: closing the computer, protecting family time, taking vacation days. She rated relationship management at 5/5, then annotated: "worries me – I feel like I am good at this but worry I'm not." She rated resilience at 5/5, then added: "but I'm not."

This participant's assessment surfaced something the confidence ratings alone would have obscured: high self-assessed confidence coexisting with genuine self-doubt is not a measurement error. It's imposter syndrome operating in a highly self-aware leader. The instrument's annotation design – allowing participants to add qualifiers to their ratings – made this visible.

Participant C was newer to her professional role and navigating confidence-building in two simultaneous contexts: the program and her job. Her most significant behavioral change was in response pattern: "I always take a moment before responding. I want to ensure my responses are filled with meaning." Her October goal – presenting a development plan to her supervisor – was assessed as incomplete. But the analysis noted what she hadn't: she had successfully presented to her supervisor and created a plan, which she was dismissing as insufficient because execution hadn't begun. Her self-assessment revealed perfectionist tendencies that prevented recognition of genuine progress. "Not enough hours in the day – need to prioritize better" appeared alongside "I am new and still building relationships and self-confidence in my role."

This participant's assessment informed the Phase 2 framing decision to explicitly normalize the gap between awareness and confidence – naming it as a predictable stage of development rather than a sign of insufficient progress.

What the Data Produced at the Cohort Level

Cross-cohort synthesis of all ten assessments produced the following indicators:

- **90%** of participants rated all three Phase 1 sessions 4/5 or 5/5
- **80%** reported leadership confidence "somewhat" or "significantly" higher than at program start

- **70%** applied four or more skills from Phase 1 content with documented examples
- **100%** identified the cohort relationship and psychological safety as valuable

Behind those aggregate numbers, the analysis identified meaningful variation: two participants applying all seven skills comprehensively, several applying skills in crisis contexts where application required significantly more effort than the numbers reflected, one participant with philosophical resistance to goal-setting frameworks that pointed toward a gap in program design, and one participant in sufficient distress that one-on-one check-ins were warranted alongside programmatic support.

Aggregate satisfaction data would have shown a highly successful program. The individual analyses showed a more complex picture – and a more useful one for design iteration.

The Principle Behind the Practice

Evaluation built in from the start of a program produces evidence. Evaluation bolted on after the fact produces testimony.

The distinction matters because testimony and evidence support different conclusions. Testimony tells you participants valued the experience. Evidence tells you whether the experience changed anything. For a program designed explicitly around behavior change, only one of those answers is sufficient.

The instrument, the analysis process, and the design adjustments that resulted are the same methodology described in the behavior change case study – applied to the evaluation function specifically. Measurement isn't a separate activity from program design. It's an extension of the same design logic.

Note: *The Phase 1 Participant Assessment instrument is included as an appendix to this document.*

Chris Bailey is the founder of Bailey Learning Works, LLC, a leadership development consultancy specializing in cohort-based program design, executive facilitation, and behavior change measurement. He holds an M.Ed. in Learning, Leadership, and

Organization Development from the University of Georgia and brings certifications in LEGO® Serious Play facilitation and Executive Change Coaching. He works with clients across manufacturing, energy, and professional services, and is available for corporate leadership development and organizational development engagements.

Connect at hello@baileylearningworks.com or linkedin.com/in/chrisbaileyworks

Personal Leadership Phase Assessment

September - November Reflection & Program Feedback

Your Name: _____

Purpose

Today marks the end of our personal leadership phase (September-November). This assessment helps you:

- Reflect on your growth over the past three months
- Identify what you're learning and actually using
- Give me feedback to improve the program for you and future cohorts

This is not a test or evaluation of you. I'm trying to understand what's working in the program and what you're able to apply in your real life. Your honest feedback helps me serve you better.

PART 1: Skills Learned & Applied

A. Most Valuable Learning

What have you learned in September-November that's been most valuable to you? (This could be a concept, skill, framework, insight, or practice)

Why has this been valuable? How has it helped you?

B. Skills You've Actually Used

Check all that apply and provide a brief example:

☐ **Self-awareness practices** from September
Example of how I used it:

☐ **Managing my emotions** effectively (September)
Example:

☐ **Values-based decision making** (October)
Example:

☐ **Vision development** for my leadership (October)
Example:

☐ **Goal-setting** approaches (October)
Example:

☐ **Resilience practices** (November)
Example:

☐ **Navigating uncertainty** (November)
Example:

☐ **Other:** _____
Example:

C. Confidence Self-Assessment

Rate your current confidence in each area (1 = Not confident, 5 = Very confident):

Self-awareness (understanding my emotions, triggers, strengths):	1 2 3 4 5
Self-management (managing emotions, staying focused under pressure):	1 2 3 4 5
Social awareness (reading others, understanding different perspectives):	1 2 3 4 5
Relationship management (building trust, communicating effectively):	1 2 3 4 5
Vision & goal-setting (knowing my direction and setting meaningful goals):	1 2 3 4 5
Resilience & adaptability (staying effective when facing challenges):	1 2 3 4 5

Compared to September, my overall leadership confidence is:
☐ Significantly lower ☐ Somewhat lower ☐ About the same ☐ Somewhat higher ☐ Significantly higher

---

PART 2: Behavior Change & Application

A. Specific Changes You've Made

Describe ONE specific example of something you're doing differently in your work, community involvement, or personal life because of this program. Be as concrete as possible: What's the situation? What do you do differently now? What's been the result?

B. October Goal Progress

Quick Win Goal: (What was your 14-21 day goal from October?)

Progress made: ☐ Completed successfully ☐ Made significant progress ☐ Made some progress ☐ Struggled to make progress ☐ Did not attempt

What helped you make progress?

What obstacles or challenges did you encounter?

What do you need to continue working toward this goal?

C. Application Challenges

What obstacles have you encountered in applying what you've learned? (Select all that apply and add details)

☐ Time constraints - Details:

☐ Workplace culture doesn't support it - Details:

☐ Lack of opportunity to practice - Details:

☐ Still building confidence - Details:

☐ Unclear how to apply to my specific context - Details:

☐ Need more practice/support - Details:

☐ Other: _____

PART 3: Program Feedback

A. What's Been Most Effective

Which learning activities have been most effective for you? (Rank your top 3: 1 = most effective)

___ Small group discussions and sharing

___ Individual reflection time

___ Practical exercises and simulations

___ Hearing from community leaders

___ Learning frameworks and models

___ Applying concepts to real scenarios

___ Accountability partnerships

___ Self-assessment tools

___ Other: _____

Why have these been most effective for your learning?

B. What Would Improve the Program

What would make the program more valuable for you?

What do you need more of in the remaining sessions (December-May)?

What do you need less of?

C. Session-Specific Feedback

Rate each session (1 = Not valuable, 5 = Extremely valuable):

September (Self-Awareness & Emotional Intelligence): na 1 2 3 4 5

Most valuable part:

Could be improved:

October (Personal Vision & Goal Setting): na 1 2 3 4 5

Most valuable part:

Could be improved:

November (Resilience & Adaptability): na 1 2 3 4 5

Most valuable part:

Could be improved:

PART 4: Looking Ahead

A. December-May Focus

As we move into workplace leadership (December-February) and community leadership (March-May), what do you most want to learn or develop?

What concerns or anxieties do you have about the next phase?

PART 5: Additional Comments

Is there anything else you'd like me to know about your experience so far, your learning, or the program?