

Program Evaluation

A Leadership Session

Haverford College | Senior Staff

Before We Begin

Do not answer these out loud. Just notice what comes to mind

- 1 Think about the terms evaluation, assessment, and monitoring. In your mind, are they the same thing? If not, what is different?
- 2 Think of one program under your area. If someone asked you right now whether it is working, what evidence would you point to?
- 3 When was the last time a program decision you made was changed by evidence rather than confirmed by it?
- 4 What assumptions are you making about your programs that you have never tested?

Hold onto whatever comes to mind. We will come back to these

Session Overview

You Will

- Learn the difference between monitoring, assessment, and evaluation
- Identify untested assumptions in your programs
- Apply a five-category leadership question framework
- Know when and how to coach your teams through evaluation
- Practice real coaching conversations
- Understand what your teams are learning in the course

You Will Not

- Design evaluations or write evaluation questions
- Learn data collection or analysis methods
- Build logic models or theories of change

The Evaluation Mindset Shift

Meet Greenfield College

Greenfield College is a fictional liberal arts institution similar to Haverford. Their Office of Student Affairs runs a Student Leadership Development Program with four components:

- 1 Leadership Retreat** Intensive off-campus weekend experience
- 2 Peer Facilitation Training** Semester-long skills development series
- 3 Community Service Projects** Team-based service learning
- 4 Speaker Series** Monthly guest lectures and panels

The program has been running for four years. You are about to see their dashboard

- Open the Greenfield College dashboard on your screen
- Explore the data across all three views: Summary, Cost & Resources, and Demographics
- As a leadership team, you must make a recommendation for this program: Invest More, Maintain As-Is, Redesign, or Divest
- Be ready to explain your reasoning and which data points drove your decision
- You have 15 minutes

<https://www.turningdataintowisdom.com/programevaluation/>

What Did You Decide?

What data did you rely on most?

How confident are you in your recommendation?

What Was Missing?

You just made a resource decision about a \$287,400 program. Here is what the dashboard never told you

1

No defined outcome. The program never stated what "leadership development" looks like in observable terms

2

No evidence of student change. Every metric is about activity, attendance, or satisfaction. Not one measures behavior change

3

No comparison. We have no idea whether participants are different from non-participants

4

No evaluation question. The dashboard answers "how much did we do?" but never "did it work?"

5

No portfolio comparison. The program has four parts with very different costs and reach, but no way to compare which ones produce the most impact

This is the difference between reporting and evaluation. The dashboard told you what happened. It never told you what changed

The Data Had Gaps. . .

If the data did not answer the question, what did?

Your assumptions did. Let's look at two kinds

Two Kinds of Assumptions

Think about a routine doctor's visit

Kind 1: Observational

How you read the evidence

Your blood pressure is normal. Weight is stable.
Cholesterol is in range. The doctor says "everything looks good." You leave feeling healthy.

*But the doctor only measured what they measured.
The numbers created confidence that may not be warranted.*

This is the Greenfield dashboard

Kind 2: Design

The untested bet in the program

You start taking a supplement because you read it "supports heart health." You assume it will work. That is a hypothesis. You have not tested it. You just believe the design will produce the result.

Believing the design will produce the change and testing whether it does are different things.

This is every program you launch

We are going to explore both. First, the observational assumptions you just made with the Greenfield dashboard

Observational Assumptions

Kind 1: Observational Assumptions

Assumption: High attendance means the program is reaching the right students

Reality: *68% of participants already hold leadership roles*

Assumption: High satisfaction means students are benefiting

Reality: *Satisfaction measures the experience, not the outcome*

Assumption: Growing participation means the program is effective

Reality: *Participation rate is actually declining as enrollment grows*

Assumption: A program called "leadership development" is developing leadership

Reality: *The program never defined what leadership development looks like*

Every program is built on assumptions about what drives the intended outcome. Evaluation is how you test them

Every Program is a Hypothesis

Kind 2: Design Assumptions

When you invest in a program, you are placing a bet. You are betting that what you designed will produce the change you intended. That bet is a hypothesis. It is not a fact until you test it.

The implicit bets inside the Greenfield program

Leadership Retreat: *"An intensive weekend of reflection and bonding will produce lasting leadership behaviors"*

Never tested

Peer Facilitation Training: *"Practicing facilitation skills will transfer to other campus leadership contexts"*

Never tested

Community Service Projects: *"Team-based service will develop collaborative leadership capacity"*

Never tested

Speaker Series: *"Exposure to leaders will inspire students to develop leadership skills"*

Never tested

Each of these sounds logical. Each might be true. But none has been validated

This Applies to Every Program You Run

Every program has this structure, whether you have named it or not:

We will do ***[activity]*** and it will produce ***[outcome]*** for ***[population]***

The link between the activity and the outcome is your hypothesis. It is the bet. It is an assumption.

Student Affairs: **Peer mentoring program** → Increased sense of belonging among first-year students

Academic Support: **Tutoring center expansion** → Improved academic performance in gateway courses

Library Services: **Reading engagement program** → Increased reading behavior among patrons

When your team proposes a new program, ask: "What is the hypothesis? What are we betting will happen?"

Naming Assumptions is a Leadership Skill

One of the most valuable questions you can ask your team

"What are we assuming about this program that we have not tested?"

This question does not require evaluation expertise. It requires the willingness to name the assumption.

Common assumption patterns in programs

If we build it, they will come

Are we reaching the students who need this most?

If they attend, they will change

What specific change should we see, and when?

If they like it, it must be working

Can a program be satisfying and ineffective at the same time?

Evaluation is Decision Support

The goal of evaluation is not to prove a program works. The goal is to reduce uncertainty enough to make a better decision and determine what to do next.

The Greenfield Dashboard

- How many attended?
- Were they satisfied?
- How many events did we run?

What Evaluation Would Ask

- What changed for students?
- Which parts drove that change?
- What should we do differently?

Quality comes from focus, not volume

Activity vs. Outcome: The Greenfield Example



What Greenfield Reports

- 847 participants across 64 events
- 4.4/5 satisfaction score
- 92% would recommend

Activities and reactions



What Evaluation Would Require

- Students demonstrate leadership behaviors in campus roles
- Participants show measurable skill development
- Leadership capacity is distributed more equitably

Observable changes in students

When your team describes what they do rather than what changes, that is your coaching moment. Ask: What does success look like for the student?

"But We Cannot Measure That"

When a program director tells you an outcome cannot be measured, it usually means the obvious method does not work. It does not mean measurement is impossible. Here are five practical approaches:

The outcome is hard to measure directly: Use a proxy variable

A library cannot watch patrons read. But they can track borrowing frequency as a proxy for reading behavior.

You are not sure the design will work: Run a pilot

Test a redesigned version with one cohort before full rollout. Lower risk, real evidence.

Automated data does not exist: Use manual collection for a bounded period

Structured observations, interviews, or artifact review for one semester. It does not need to be permanent to be useful.

You cannot compare to a control group: Use pre/post or matched comparison

Measure before and after, or compare to a similar group. Not perfect, but far better than no comparison.

The timeline is too long for immediate results: Identify leading indicators

Short-term markers that predict longer-term outcomes. Early engagement patterns, skill demonstration, behavior shifts.

What Your Teams Are Learning

The evaluation pipeline and where your coaching matters

The Evaluation Pipeline

Your team's work

Modules 1-4

Foundation

Systems thinking, theory of change, logic models

Modules 5-6

Design

Evaluable outcomes, questions, judgment criteria

Module 8

Evidence

Indicators, data collection, triangulation

Module 9

Diagnosis

Interpret findings, root cause analysis

Modules 10-11

Action

Improvements, recommendations, briefings

Your coaching moments

"Develop student leaders" is not an outcome. What does it look like?

Are we measuring satisfaction again, or actual behavior change?

Does this give us enough to decide about the Retreat budget?

If you wait until the final briefing to engage, you have missed your three best opportunities to shape the work

The Architecture of Evaluation Thinking

What was missing from the Greenfield dashboard? Every single layer.

Outcomes

What should change?

*Never defined
what "leadership
development"
means*

Questions

What do we want to learn?

*No evaluation question, just
satisfaction surveys*

Components

Where will we look?

*Never identified which parts of the
program matter most*

Dimensions

How will we judge?

No criteria for what "good" looks like

Zooming In: The Leadership Retreat

Why this component?

- Consumes 49.7% of the total program budget (\$142,800)
- Costs \$1,831 per participant — 34x more than the Speaker Series
- Serves only 78 students — the smallest component by far
- Participation has been flat for four years
- Has the highest satisfaction score (4.6) but no outcome evidence

High cost. High satisfaction. Zero evidence of impact. This is exactly the kind of program that evaluation is designed for

From Vague to Evaluable

What Greenfield Claims

"The retreat develops student leaders"

Assumes success. No observable change defined.

An Evaluable Outcome

"Leadership behaviors among first-year Retreat participants in their campus roles during the following semester"

Specific. Bounded. Can fail.

Three tests your team applies

- 1 Specific Change** "leadership behaviors in campus roles" names the observable change
- 2 Bounded Scope** "first-year Retreat participants" and "following semester" bound population and time
- 3 Judgement Ready** The outcome defines what to examine, not what will happen. The evaluation question asks whether it occurred.

The Evaluation Question

To what extent do students who participate in the Leadership Retreat demonstrate increased leadership behaviors in their campus roles during the following semester?

This is where your coaching matters most. When your team drafts this question, ask:

Feasibility: Can we actually observe leadership behaviors? How?

Utility: If we answer this, does it change the \$142,800 budget decision?

Focus: Are we asking one clear thing, or three things jammed together?

You do not need evaluation expertise to ask these questions.

What This Makes Possible

With an evaluable question, the decision logic becomes clear:

If: Retreat produces measurable leadership behaviors

Then: The \$1,831/student cost may be justified. Consider expanding.

If: Behaviors increase but only for certain students

Then: Redesign access or targeting. The retreat works, but not equitably.

If: No measurable change despite high satisfaction

Then: The retreat is a positive experience, not a development program. Redesign or reallocate.

This is what the Greenfield dashboard could not give you. Not just data, but decision logic

- Each group has received a draft evaluation plan from a Greenfield program director. The draft is not obviously wrong. Parts are reasonable. But it has weaknesses.
- Read the draft. Discuss what is strong and where the thinking needs sharpening.
- Formulate 2-3 coaching questions you would ask to help the director strengthen their plan. Do not rewrite it for them.
- Be ready to share: what was strong, what concerns you, and the coaching questions you would ask.

Three Patterns You Will See from Your Teams

Every draft your program directors bring you will fit one of these patterns. Recognizing the pattern tells you which coaching question to reach for.

1

Vaguely Right but Imprecise

The draft uses the right language but lacks specificity. It sounds like an outcome but you cannot tell what success actually looks like.
Students will develop a stronger sense of belonging.

Your coaching move:

"What would I see a student doing differently if there sense of belonging were stronger?"

2

Two Things Jammed Together

The draft mixes an activity with an outcome, or a question tries to evaluate multiple things at once. It is doing too much.
Did the leadership retreat increase student confidence and improve retention?

Your coaching move:

"Which one matters most for the decision you face? Let's start there."

3

Technically Correct but Possibly Meaningless

The draft passes every test on paper but sets a bar so low it cannot fail. Awareness. Satisfaction. Self-report.

Your coaching move:

"Is this the real change you care about, or a proxy for something deeper?"

The third pattern is the one you are least likely to catch on your own. Something can be technically evaluable and still not worth evaluating

Pattern 3 in Practice

A career services office runs a resume workshop. Here is their evaluation plan:

Outcome: *"Students who attend the resume workshop will report increased confidence in their ability to write a resume."*

Specific change? Yes. **Bounded?** Yes. **Judgement ready?** Technically, yes.

But will it ever fail?

You just spent an hour teaching someone about resumes. Of course they feel more confident afterward. You could run the worst workshop in history and students would still report increased confidence because they were exposed to information they did not have before.

The real problem

Confidence in writing a resume ≠ Writing a better resume ≠ Getting an interview ≠ Getting a job

The coaching question: "If every student reports increased confidence but none of them actually improve their resumes, would you call this program successful?"

Leadership Questions to Interrogate Programs

A coaching and questioning framework you can use immediately

These Questions Serve Two Purposes



Interrogate the Program

Assess whether your programs are aligned, effective, and worth the investment.

Think: the Greenfield dashboard. These questions surface what was missing.



Coach Your Team

Use the same questions when your team comes to you mid-evaluation. The question that challenges a program also sharpens an evaluation.

Five Ways to Interrogate Programs

1

Purpose & Alignment

Are we solving the right problem?

2

Outcomes & Impact

Is the program producing meaningful outcomes?

3

Program Design & Effectiveness

Which parts are actually working?

4

Resource Use & Value

Is the program worth the resources invested?

5

Learning & Improvement

What are we learning and how are we improving?

1. Purpose & Alignment

Are we solving the right problem?

Questions to ask:

- What problem was this program originally designed to address?
- Is that problem still relevant today?
- How does this program contribute to institutional priorities?
- If this program disappeared tomorrow, what would we lose?

At Greenfield...

A program called "leadership development" that mostly serves students who are already leaders. Is that solving the right problem?

Coaching Moment

When your team starts an evaluation, ask this category first. It prevents wasted effort on a misaligned program.

2. Outcomes & Impact

Is the program producing meaningful outcomes?

Questions to ask:

- What outcomes are we expecting from this program?
- What evidence do we have that those outcomes are occurring?
- Are the outcomes meaningful for students?
- Which students benefit most and least?

At Greenfield...

Greenfield had no outcome data at all. Satisfaction is not an outcome. What is actually changing for students?

Coaching Moment

If your team cannot name the outcome, they are not ready to evaluate. Push them to get specific.

3. Program Design & Effectiveness

Which parts are actually working?

Questions to ask:

- Which components contribute most to outcomes?
- Are there parts adding little value?
- Are we delivering the program as intended?
- If we redesigned this today, what would we change?

At Greenfield...

*The Retreat consumes half the budget.
The Speaker Series reaches 6x more
students at 3% of the cost. Which is
driving impact?*

Coaching Moment

*This question gives your team permission
to be honest about what is not working.*

4. Resource Use & Value

Is the program worth the resources invested?

Questions to ask:

- How much time, staff effort, and funding does this require?
- Are outcomes proportionate to the resources invested?
- Are there lower-cost ways to achieve similar outcomes?
- Are there programs that continue mainly due to history?

At Greenfield...

\$1,831 per student for the Retreat vs. \$54 for the Speaker Series. Without outcome data, you cannot compare value.

Coaching Moment

When findings arrive, this category turns evidence into budget conversations.

5. Learning & Improvement

What are we learning and improving?

Questions to ask:

- What have we learned from evaluating this program?
- What improvements have been made based on evidence?
- What questions remain unanswered?
- How often should we re-examine this program's design?

At Greenfield...

Greenfield has run this program for four years and never asked what it produces. What would they know now if they had started evaluating in year one?

Coaching Moment

After each evaluation cycle, ask your team: what did we learn, and what should we look at next?

Coaching Your Team Through Evaluation

How to shape the work, not just receive the report

The Tool You Are Taking With You

You have spent the last 90 minutes building the muscle on Greenfield. Now it is time to connect it to your own programs.

The Leadership Evaluation Reference Guide is organized by nine situations you will actually face:

1. A director proposes a new program
2. A team member is mid-evaluation
3. A team member presents findings
4. A director resists evaluation findings
5. Budget season: a director requests more funding
6. A legacy program has never been questioned
7. A director reports strong satisfaction numbers
8. A program is underperforming and you must decide
9. Annual review: reviewing the program portfolio

Find your situation. Ask the questions. Coach the conversation.

What Your Team Will Deliver

1

Finding

A clear, evidence-based statement about what happened

2

Root Cause

A diagnosis of why, using a structured framework

3

Recommendation

An improvement strategy aligned to the cause, not the symptom

The quality of this output depends on the conversations you have along the way

The Greenfield Retreat: Evaluation Results

Imagine Greenfield had done a proper evaluation. Here is what they might have found:

Finding: Students who completed the Retreat showed no measurable difference in leadership behaviors compared to matched non-participants. However, 78% reported high satisfaction and personal growth.

Root Cause: The retreat design emphasizes personal reflection and team bonding but does not include structured skill practice, application planning, or post-retreat follow-through. Students enjoy the experience but do not transfer behaviors.

Recommendation: Redesign the retreat to include structured leadership skill practice and a post-retreat application plan with a 60-day check-in. Pilot with one cohort before full rollout.

This is the same program you defended two hours ago. The recommendation targets the cause (no skill transfer), not the symptom (low impact)

Five Questions to Ask When You Receive Findings

1

Is the finding about an outcome or an activity?

Confirms your team evaluated change, not effort

2

Does the root cause point to something we can change?

Ensures the diagnosis is actionable

3

Does the recommendation target the cause, not the symptom?

Prevents surface-level fixes

4

What would we do differently if the finding were reversed?

Tests whether the evaluation informed a real decision

5

What does this evaluation not tell us?

Names the blind spots so you are not overconfident

Practice the Coaching Conversation

10 min

- Your program director brings you the Greenfield Retreat evaluation results. Despite the finding of no measurable behavior change, they push back: "Students love this program. Satisfaction is 4.6. Mentors say it is transformative. I do not think this evaluation captures what the Retreat really does."
- Using the leadership question framework, how do you coach this director through the findings without dismissing their perspective or ignoring the evidence?
- What questions would you ask to move the conversation from belief to evidence?

How This Connects to Your Teams

Your Role as Leaders

- Ask the right questions about programs
- Coach your team at key evaluation moments
- Challenge assumptions about what is working
- Demand evidence, not just satisfaction data
- Act on findings, not just receive reports

Your Staff's Role

- Build program logic models and theories of change
- Define evaluable outcomes and evaluation questions
- Collect and analyze evidence systematically
- Diagnose root causes using structured frameworks
- Deliver findings with actionable recommendations

Your questions drive their evaluations. Your coaching sharpens their work. Their evidence informs your decisions

Key Takeaways

- 1 Every program is trying to create change. Evaluation shows whether it actually does
- 2 Program evaluation reduces guessing. It helps you decide what to do next
- 3 If you don't define success, everything looks like success
- 4 The question is not whether your program is running. It is whether it is making a difference
- 5 Better data doesn't replace judgment. It improves it