Measuring Leadership Capability Beyond Participation

Six patterns from Australian leadership development practice

Beth Hall FAITD

Joanne Spriggs GAICD

May 2026

Foreword

This research began with a broader line of inquiry, how L&D is measuring impact, and whether what we measure is genuinely telling us anything meaningful about capability. Across the profession, there is increasing pressure to demonstrate value. Yet much of what is measured continues to focus on activity, attendance, completion and participant feedback, rather than what has changed. This raised a more fundamental question for us: what should we be measuring, and why?

As we explored this through case studies and practitioner input, a consistent theme emerged. While there is no shortage of frameworks or evaluation models, the challenge sits in how evaluation and impact are designed, embedded and used in practice. Measurement is often treated as something that follows delivery, rather than something built into how capability is developed and observed over time.

Within this, leadership capability became a natural point of focus. It is an area that continues to emerge as a priority for organisations, and one where practitioners consistently report difficulty in evidencing impact in a meaningful way. The case studies included in this report reflect where we were able to access both strong practice and credible evidence, allowing us to examine this challenge in a more applied context.

What the research shows is a clear shift. The organisations generating the most useful evidence are not those with the most complex approaches, but those that are deliberate in how they define success, structure data capture, and use insights to inform decisions. Evaluation is designed upfront, evidence is anchored in behaviour and performance in role, and findings are actively used to adapt, refine, scale or stop investment.

This points to something more fundamental than measurement technique. It reflects how leadership development and evaluation are operationalised within organisations. Where measurement is embedded into the normal rhythm of leadership, supported by governance, systems, manager involvement and clear ownership, capability becomes more visible and more actionable over time.

This work also reinforces an important shift in how we think about evidence. Participation and satisfaction remain useful signals, but they are not indicators of capability. What matters is what leaders do differently in their roles over time, and how that connects to outcomes the organisation can see and act on.

At AITD, this aligns with how we support the profession, leading, connecting and sharing practice that strengthens capability in a practical and applied way. Research like this helps make that visible, grounded in what organisations are doing. Thank you to Beth and Joanne for the depth and rigour of this work, and for bringing forward insights that will help practitioners think more clearly about how capability is built, measured and strengthened over time.

Ben Campbell

Chief Executive Officer

Copyright

Version 1.0 2026

Executive Summary

Leadership capability is one of the differences between an organisation that adapts and one that stalls. As complexity increases and the demands on leaders grow, the ability to build, track, and strengthen leadership capability over time is becoming a strategic must have, not just an L&D priority.

This research draws on in-depth case studies with three Australian organisations to examine what leadership capability measurement looks like in practice, what enables it, and what gets in the way. These cases do not represent the full state of Australian practice. They offer practical examples of what becomes possible when organisations move beyond activity reporting and begin treating leadership capability measurement as part of how leadership is managed.

Six patterns emerged consistently across organisations measuring leadership capability shift:

1. Measurement needs to sit inside the normal operating rhythm
2. Organisations with weak systems struggle to produce strong evidence
3. Transfer is designed in rather than diagnosed after
4. The useful evidence changes decisions
5. Organisations are borrowing from multiple frameworks rather than adopting one clean model
6. Someone must own the measurement cadence and follow-through

Across the cases, the strongest evidence shared three characteristics:

- A clear success chain - what should change, and how it will be evidenced
- Evidence anchored in work - application, not perception
- Active use of evidence - informing decisions to adjust, target, scale, or stop

These patterns point to an operating model and prioritisation problem.

Measurement is often treated as a post-program activity, rather than embedded into the design, transfer, and follow-up of leadership development. As a result, organisations optimise for what is easy to measure, rather than what is meaningful.

Moving beyond attendance requires a shift in what counts as evidence, and a commitment to tracking leadership capability as it is applied and evolves over time.

Otherwise, organisations keep making investment decisions based on participation and perception. A stronger evidence standard makes leadership capability more visible and more governable. The central finding is that leadership capability measurement is a design, governance and operating rhythm problem, not a reporting problem.

The Case for Measuring Leadership Capability

Leadership capability continues to be a top priority for organisations across Australia and ANZ, representing a multi-billion-dollar annual investment in leadership development globally (Association for Talent Development, 2023; McKinsey & Company, 2021).

> **Yet many organisations still struggle to answer a fundamental question:**
>
> How do we know if leadership capability has improved?

At the same time, leaders are operating in increasingly complex environments shaped by AI, shifting workforce expectations, and organisational ambiguity. Leadership capability is no longer a "nice to have" - it is a critical lever for performance, adaptability, and organisational resilience (McKinsey & Company, 2021).

Measuring the activity of leadership development and measuring leadership capability are not the same thing. Programs, experiences, and structured learning are the inputs. Leadership capability observed in role over time, the behaviours, judgements, and decisions leaders demonstrate under real conditions, is the outcome.

Many organisations are still using program activity as a proxy for capability, that is a poor substitute.

Across Australian organisations, participation rates and post-program surveys remain the most common indicators of success. While useful for tracking attendance and completion, these measures offer limited insight into whether leadership capability has genuinely shifted. Much of the evaluation literature cautions against treating satisfaction data as evidence of capability change, noting that it captures how participants feel about an experience, but not whether development has influenced how they think, behave, or perform (Leaman, 2016; Cendros, 2025). As Leaman argues, organisations have become focused on activity-based metrics because they are easy to collect rather than because they are useful to make decisions. Without more meaningful evidence, it becomes difficult to make informed decisions about where to invest, scale, or redesign leadership development efforts.

Behaviour change consistently emerges across the literature as the most critical and most under-measured indicator of development effectiveness. Empirical evidence suggests that while learning outcomes are moderately predictive of behaviour change, participant reactions show weak and inconsistent relationships with downstream impact (Alliger et al., 1997; Sitzmann et al., 2008). Despite its importance, behaviour change remains difficult to measure due to attribution challenges, limited access to performance data, and reliance on self-report measures. For leadership specifically, this challenge is compounded: leadership behaviour is relational, visible only through others, and its impact on teams and culture accumulates slowly over months and years. The outcomes that would constitute genuine proof of leadership capability shift, succession depth, culture change, team performance, are co-produced with many other organisational factors, making direct attribution unrealistic.

Across the literature development effectiveness is shaped by factors extending well beyond program design. Managerial support, opportunity to apply, reinforcement, and the broader organisational environment significantly influence whether development translates into sustained capability change (Holton et al., 2000; Yaqoot et al., 2017). Evaluation approaches

that isolate development from the organisational context in which it is expected to deliver value consistently underestimate these conditions as determinants of impact.

While organisations increasingly seek to demonstrate business value from leadership investment, measurement of organisational impact remains elusive. Garavan et al. (2019), in a review of 217 empirical studies, identify widespread validity threats in research examining the development-performance relationship, calling into question the strength of many causal claims. For leadership capability specifically, contribution logic is usually more honest and defensible than attempting to isolate direct ROI, unless the evaluation design has been built to support that level of attribution.

The gap between evaluation theory and practice is well documented. Academic research advocates for longitudinal, multi-level, and context-sensitive approaches. Organisations face practical challenges including limited evaluation capability, data access issues, and competing priorities. This gap is not primarily a lack of awareness. It is a practical response to organisational realities (Cendros, 2025; Garavan et al., 2019). So, the task is to design an evaluation model that an organisation can sustain, is practically feasible, and is designed to support decisions rather than satisfy compliance.

The frameworks below represent the most referenced lenses for measuring leadership capability and development effectiveness. None is designed to be adopted wholesale. Each brings a different question into focus. Use this table to understand what each framework is optimised for, then draw selectively based on your context, maturity, and the decisions you need to make.

Table 1: Evaluation Frameworks and Lenses: A Practitioner Guide

Framework	What it brings into focus	Limitations and best fit
Kirkpatrick & Kirkpatrick (2016)	A simple structure for thinking beyond satisfaction - from reaction to learning, behaviour and results. Creates shared language with stakeholders.	In practice, often applied narrowly and stalls at Levels 1–2. “Results” can be claimed without robust linkage. Limited attention to context or transfer conditions. Best when you need a starting point or common language to elevate the evaluation conversation.
LTEM (Thalheimer, 2018)	The quality of evidence collected - Prioritises evidence of transfer and application in role, with a focus on observed performance rather than perceptions or participation data.	Requires clarity on expected performance. Harder to apply without access to performance data. Best when you want to strengthen credibility and reduce reliance on satisfaction or self-report measures.
LTSI (Holton et al., 2000)	The conditions that enable or block transfer - manager support, opportunity to apply, reinforcement and environment.	Can become diagnostic-heavy without action. Full inventories can feel complex. Best when behaviour change is inconsistent and you suspect systemic barriers.

Success Case Method (Brinkerhoff, 2003)	What worked, for whom, and why - by exploring high- and low-impact cases in depth. Surfaces enabling conditions.	Does not estimate average impact. Risk of biased case selection if not handled carefully. Best when you want compelling evidence stories to inform redesign or scale decisions.
Contribution Analysis (Mayne, 2011)	Whether development plausibly contributed to outcomes in complex environments where attribution is unrealistic.	Can become documentation heavy. Still relies on strength of assumptions and evidence. Best when multiple initiatives influence performance and you need a contribution story.
Realist Evaluation (Pawson & Tilley, 1997)	What works, for whom, in what context - explaining variation through context–mechanism–outcome thinking.	Conceptually heavier. Requires thoughtful interpretation of findings. Best when impact varies across teams, roles or environments and you want to understand why.
Theory of Change / Outcomes Chain (Weiss, 1995)	The pathway from activity → capability → performance → organisational outcome. Clarifies leading and lagging indicators.	Can become over-engineered. Risk of spending more time mapping than measuring. Best when you need alignment between development activity and strategic business priorities.
Capability / Skills-Based Lens	Capability progression anchored to role expectations and proficiency standards. Links development to workforce readiness.	Depends on having clear capability definitions. Risk of becoming taxonomy work. Best when you are aligning development to workforce planning or skills-based strategy.
Talent Readiness & Mobility Metrics	Whether capability can be deployed - succession depth, internal fill, time-to-readiness. Positions impact in business terms.	Outcomes are lagging and attribution is shared. Requires clean definitions of readiness. Best when you want to connect development to workforce strength and enterprise resilience.

The research that follows examines what this looks like in practice, drawing on in-depth case studies with three Australian organisations to surface what leadership capability measurement requires.

Research Approach

Data Collection

This study draws on semi-structured interviews (45–60 minutes) with representatives from three Australian organisations: GO.FARM, Hickory Construction, and the Institute of Public Administration Australia (IPAA). The research was designed to explore how leadership capability measurement is enacted in practice across different organisational contexts.

Interviews were conducted via video, recorded, and transcribed verbatim. Individual case studies were developed from the interview data to capture each organisation's approach to measuring leadership capability, with written approval obtained from all participating organisations prior to publication.

Participants were senior leaders responsible for people, culture, and capability, including:

- GO.FARM: Richard Bligh, Head of People & Performance
- Hickory: Pete Howell, Chief People Officer
- IPAA: Kate Fraser, Executive Director, Program Delivery

Organisations were intentionally selected to ensure diversity across sector (agriculture, construction, and public administration), workforce composition (blue-and white-collar), and organisational size and maturity. This approach was designed to capture a broad representation of the Australian organisational landscape and enhance the transferability of findings across contexts.

This research is based on a small number of in-depth case studies. The findings should be read as practice patterns rather than sector-wide benchmarks. The value of the research lies in surfacing transferable principles, not prescribing a single measurement model.

Measurement in Practice

Table 2 summarises three organisations measuring leadership capability in practice across different stages of maturity. Together, they demonstrate a shift beyond participation and satisfaction toward behavioural change, capability progression, and organisational impact. Each organisation draws on different elements of evaluation frameworks, applied in line with its context, priorities, and stage of maturity.

Table 2: Organisational Snapshot

Organisation	Development Focus	Primary Measurement Approach	Key Evidence Sources	Success Indicator	Evaluation Frameworks & Logic
Hickory Construction ~1,050	PM and site manager pipeline	Foundations-first: behavioural framework, simplified performance signals, quarterly cadence, competence verification	Promotion readiness assessments; three-point ratings; graduate competency assessments	80% internal fill rate	Outcomes chain · LTEM · Readiness/mobility · Transfer conditions
GO.FARM Agriculture ~160 (targeting 400+)	Multi-level leadership pipeline	Psychometric progression with multi-rater evidence at transition points	Saville Wave Professional Styles; Saville Wave 360; behavioural commitments; action plans	Behavioural consistency; capability uplift; pipeline readiness	Outcomes chain · LTEM · Readiness/ progression · Transfer conditions
IPAA Victoria Public sector association 90+ member orgs; 350k+ reach	Leadership and AI capability	Lean measurement with work-anchored application evidence	Pre/post competency ratings; project rubrics; intentions worksheet; 3-month follow-up	Competency uplift; observable application; cohort demand	Outcomes chain · LTEM · Contribution logic · Transfer conditions

Discussion

What Measuring Leadership Capability Actually Requires

Before drawing out what the evidence shows, it is worth being precise about terms. Leadership development is the intervention: the programs, experiences, coaching, and structured learning that organisations invest in to grow their leaders. Leadership capability is what is observed: the behaviours, judgements, and decisions that leaders demonstrate in role, over time, under real conditions. In practice, this means evidence must be anchored in observed behaviour in role over time, not in-program indicators or immediate post-program feedback. The distinction matters because it shapes what measurement is trying to do. We need to understand whether leadership capability in the organisation is shifting in ways that matter. Keeping that distinction clear is the first discipline of measurement.

The organisations documented here have already invested in clearer expectations, stronger follow-up, and better evidence. They were selected because they are taking practical steps beyond attendance data. They are not the average baseline. Many Australian organisations are still operating primarily at Tier 1, and the patterns here reflect what becomes possible when leadership capability measurement is treated as a strategic priority. The intention is not to set an intimidating standard, but to make visible what practice looks like so that practitioners can identify their own next step.

What the evidence confirms

Six patterns emerged across the three case studies in organisations generating credible evidence of leadership capability shift.

Pattern 1: Measurement needs to sit inside the normal operating rhythm

The strongest cases in this research are not evaluating programs after the fact. They are running leadership capability measurement as the way they manage leadership, built into pipelines, transition points, performance cadence, and board reporting. That is a fundamentally different starting point. For these organisations, measurement is not an activity that follows delivery. It is part of how leadership is run.

This matters because leadership capability does not shift at the pace of a program. Habits shift through deliberate practice, reinforcement, failure, and adjustment across months, not at the end of a workshop. That requires a longitudinal view that most L&D functions are not currently resourced or structured to maintain. The profession is reasonably good at tracking promotions and exits. What remains hard to track is the behavioural change happening in between, because it is incremental, contextual, and visible only to the people working alongside those leaders every day. Most L&D teams have moved on to the next delivery by the time the more meaningful evidence could be collected.

From there, the evidence standard is that leadership capability evidence needs to be anchored in observed behaviour in role, over time. Commitments tracked through action plans and reviewed at regular intervals. Pipeline data that connects capability progression to readiness for the business ahead. Follow-up conversations at 60 or 90 days. The story senior stakeholders need is not "here is what leaders said they would do at the end of a workshop." It is "here is how leadership capability has shifted across this cohort over eighteen months, and here is where the pipeline is stronger than it was." Building the capacity to tell that story requires treating measurement as part of how leadership is managed, not as a post-program report.

Pattern 2: Organisations with weak systems struggle to produce strong evidence

Before leadership capability evidence is possible, foundations must exist, clear behavioural expectations anchored to observable standards, a workable performance cadence that creates regular touchpoints, managers equipped to participate in goal-setting and structured follow-up, and data infrastructure capable of holding evidence across cohorts and time. These are the prior conditions the frameworks assume but rarely address.

The practical consequence is that organisations investing in sophisticated measurement before they have these foundations in place are wasting effort. If there is no agreed definition of what good leadership looks like in behavioural terms, there is nothing to measure against. If performance conversations are irregular or avoided, there are no natural moments to surface development signals. If managers are not equipped or expected to assess capability in role, the observational data simply does not exist. Frameworks and tools cannot compensate for structural gaps in how leadership is expected and managed day to day.

This is not a counsel of perfection. Organisations can build measurement capability and system foundations at the same time, and several of the organisations in this research are doing exactly that. But it does mean being honest about what the evidence can and cannot say at any given stage of maturity. Credible leadership capability evidence reflects the system it sits inside. The question is not just "what should we measure?" It is "what are we actually capable of measuring well, and what do we need to build first?"

Pattern 3: Transfer is designed in rather than diagnosed after

The measurement approaches that generate the most credible evidence in this research are inseparable from the conditions that enable transfer. Nomination requirements that signal accountability before a program begins. Spaced delivery across months rather than compressed into days. Applied projects with real stakes and real audiences. Manager calibration of goals at the start, with structured check-ins across the program. Peer accountability through group coaching or cohort touchpoints. These are transfer design choices. They are also what makes leadership capability measurement meaningful, because they create the conditions under which application evidence can be collected. If a program is not designed for transfer, measurement will consistently confirm that transfer is weak.

There is a persistent gap worth naming honestly. Most of what is described as leadership capability evidence, even from well-designed programs, is still proxy evidence. Documented commitments, manager ratings, and self-assessed competency are all meaningfully better than satisfaction surveys. But they are not the same as verified behaviour change observed over time in role. A leader who commits to "offering responsibility rather than tasks" is expressing intent. Whether that intent becomes embedded capability is a different question, and most organisations have limited capacity to answer it consistently. A strong behavioural commitment is meaningful leading data. It is not confirmed capability change. The two are often conflated in how investments are evaluated and reported and closing that gap honestly requires more systematic follow-up than most teams currently have capacity for.

The implication is that measurement design and transfer design are not separate activities. The embed phase, the structured period of reinforcement, practice, observation, and adjustment that determines whether what was learned becomes what is done in role, needs to be built into the program architecture, resourced, and followed through. The programs generating the most credible leadership capability evidence in this research were still actively engaged with participants at three, six, and in some cases eighteen months after formal delivery concluded. That is not yet the norm in the profession. It needs to become one.

Pattern 4: The useful evidence changes decisions

Every strong case in this research uses leadership capability evidence to adjust, target, scale, or stop. The shift from "did this work?" to "what do we do differently?" is the defining characteristic of good measurement. Where that shift has happened, evidence is not produced for a report that gets acknowledged and filed. It is used: to redesign a module where transfer was weak, to target support for individual leaders, to inform a case for scaling what is working, or to redirect investment that is not producing the intended shift.

This has a direct implication for how organisations frame the outcomes they are trying to demonstrate. The hard organisational outcomes that would constitute proof of leadership capability impact, succession depth, culture shift, team performance, internal fill rates, are produced over years, in complex systems, shaped by many factors beyond any single leadership investment. Attempting to prove direct ROI from leadership development is measuring in the wrong way, or the wrong timeframe, or both. The more credible and more honest framing is contribution logic: assembling the data you have into a plausible evidence chain, testing alternative explanations, and presenting what the evidence can and cannot support.

That framing is not a concession. It is more persuasive to senior stakeholders than a contested ROI calculation, and more defensible when challenged. Stretching evidence further than it can credibly reach does not strengthen the case for investment in leadership capability. It undermines it. The organisations generating the most decision-useful evidence in this research are the ones that are clear about what their evidence can and cannot say, and that use it to make specific decisions rather than to justify a position already held.

Pattern 5: Organisations are borrowing from multiple frameworks

None of the organisations in this research adopted a single framework wholesale. All drew selectively from multiple sources, shaped by their context, maturity, and the decisions they needed to make. One reading of this is pragmatic: organisations use what is available and adapt it to what they need. But there is a more substantive interpretation. No single framework has yet fully solved the problem of measuring leadership capability in practice. Each brings a different question into focus. Kirkpatrick creates shared language but stalls at levels one and two. LTEM sharpens evidence quality but requires access to performance data. LTSI surfaces transfer conditions but can become diagnostic-heavy without a clear action pathway. The cases suggest that fidelity to any single framework matters less than having a coherent measurement chain that decision-makers can understand and that L&D teams can run.

The practical implication is to treat the frameworks as a toolkit rather than a methodology. The question is not which framework to adopt. It is which elements, assembled in which sequence, produce the leadership capability evidence the organisation needs to make the decisions that matter. That assembly should be deliberate rather than opportunistic: start with the decision, work backwards to the evidence required to support it, then identify which framework lenses are most useful at each point in the chain. An outcomes chain gives you the overall logic. LTEM helps you assess whether your evidence is worth collecting. LTSI tells you where transfer conditions are working against you. Contribution logic shapes how you present what you find. Used together and selectively, they produce something more useful than any of them alone.

Pattern 6: Someone must own the measurement cadence and follow-through

In every strong case in this research, specific ownership for follow-up, embed, and measurement cadence is named and held. Someone is responsible for tracking what was committed to, holding the review rhythm, and acting on what the evidence shows. That ownership sits close to the work, not in a separate reporting function. When measuring leadership capability is treated as a shared responsibility across a team or program, it defaults

to no one's priority. The organisations generating the most credible leadership capability evidence treat measurement governance as a structural decision, made before delivery begins, not a coordination problem to be resolved afterward.

Ownership within L&D is only part of the picture. In every strong case, the measurement approaches that work are those where the broader organisation has committed to leadership capability as a business imperative. That means managers and senior leaders participating in goal setting, observation, and structured follow-up. It means executive visibility of capability data, not just program activity. Without that broader organisational commitment, data collection becomes an administrative exercise, and findings are acknowledged without action. The measurement function cannot carry this alone. Where it has to, the evidence rarely changes anything.

This is the governance question that is easy to overlook in measurement design: who tracks what, at what intervals, who is accountable for acting on what the evidence shows, and who has the authority to change course when findings indicate an investment is not producing the intended shift. These are not coordination details. They are structural decisions that belong in the program design, not in its aftermath. Making them explicitly, before delivery begins, is one of the most reliable differences between organisations where measurement produces decisions and organisations where it produces reports.

Tensions worth naming

Three tensions run through the evidence that the literature does not fully resolve.

The first is the tension between speed and rigour. There is a real risk that in trying to make leadership capability measurement feel accessible, we remove too much of the evidential strength. Simplifying the complex means making robust methodology understandable and actionable. Diluting it means losing the evidential weight behind it entirely. We need to find language and formats that make rigorous measurement practical, without stripping it of credibility.

The practical question is: what is the lightest evidence set that remains defensible?

The second tension is between standardisation and contextualisation. The profession has become meaningfully better at contextualising leadership development. What is still underdeveloped is allowing participants to build and demonstrate leadership capability through their own context, applying thinking through their own lens, practising against their actual challenges rather than generic scenarios. Standardising for a group or an organisation has genuine value in building shared language about what good leadership looks like. But the most powerful contextualisation happens when participants are not just receiving tailored content but generating leadership capability evidence through their own real practice.

The practical question is: what must be common across the cohort, and where do leaders need to demonstrate capability through their own work?

The third tension is between evidence ambition and evidence usefulness. There is a version of measurement ambition that becomes its own problem: pursuing data with a tenuous link to the leadership capability being developed, building a body of evidence that cannot withstand scrutiny, and in doing so creating reputational damage rather than credibility. The risk is not just weak measurement, but loss of credibility. When evidence cannot be clearly connected to leadership capability, it undermines confidence in the findings, weakens decision-making, and over time erodes trust in L&D's ability to provide commercially meaningful insight. The practical correction is to ask, before collecting any data point, whether there is a direct and defensible line between that evidence and the leadership capability the investment is designed to shift. If the honest answer is no, or only loosely, that data point is more likely to undermine the measurement case than strengthen it.

The practical question is: will this data point change a decision?

These tensions do not have clean resolutions; the profession is still working through them. What the evidence in this research suggests is that the most useful response is to make the tension visible when designing measurement. Name which trade-off you are making and why. A deliberate choice between speed and rigour is defensible.

What credible leadership capability measurement requires

Leadership behaviour is relational. It is only visible through others. A technical skill can be assessed in isolation against a standard. Leadership capability shows up in how it lands on a team, how it shapes decisions under pressure, how it cascades through a culture over months and years. This is why the strongest leadership capability evidence in this research includes multi-source data. It is structurally necessary, because leadership capability cannot be seen from a single vantage point. Self-assessment alone is insufficient. Manager rating alone has its own blind spots. Multi-rater evidence, at the transitions and levels where leadership behaviour has the greatest consequence, is one practical way to strengthen visibility of what is happening.

The strongest leadership capability evidence in the research is collected at points of consequence, the transitions, decisions, and levels where leadership behaviour carries the greatest organisational weight. The practical question is "where does leadership capability matter most, and are we investing our strongest evidence there?" Signal evidence and intent data can sit at lower-consequence points in the chain. Multi-rater feedback, structured development planning, and longitudinal capability tracking belong where the stakes are highest.

> We need to treat leadership capability measurement as an operating system issue, not a program evaluation exercise.

When measurement is built into a multi-year pipeline with defined transition points, it becomes an operating rhythm for tracking capability. When it is anchored to a cultural destination with regular checkpoints, it becomes a change management architecture. When it connects readiness data to board-level risk reporting, it positions leadership capability as a business continuity variable. The measurement is not separate from how the organisation runs. It is woven into it.

What needs to fundamentally change to get there is not primarily a skills question, though skills matter. It is a prioritisation question. As a profession, we love delivery. It is visible, energising, and immediate in its feedback. Measuring leadership capability is slower, less glamorous, and requires us to stay present long after the exciting part is done. But it is ultimately what keeps L&D in the room when investment decisions are made. We need to spend as much time and energy on leadership capability measurement strategy as we do on delivery design, and resource it accordingly. Blaming low survey response rates, unavailable managers, and hard-to-track participants as reasons measurement doesn't happen is understandable. But it also keeps us stuck. Across the cases, measurement worked because it was designed early and treated as part of the project.

What the field has not yet solved

Even in the strongest cases documented here, hard organisational outcomes from leadership capability investment remain under-specified or rely on lagging indicators with shared attribution. This is an honest reflection of the measurement problem at the heart of leadership development: the outcomes that would constitute proof of leadership capability impact are

produced over years, in complex systems, through hundreds of small behavioural decisions that are never individually observable. What can be done is building coherent contribution stories: chains of plausible leadership capability evidence that connect development inputs to intermediate outcomes and ultimately to indicators that matter at an enterprise level. That is not the same as proof. But it is credible, it is decision-useful, and it is honest about what it is. Naming what the evidence can and cannot support is itself a professional standard worth holding.

The field would also benefit from more public examples of what happens when leadership capability evidence indicates an investment is not producing the intended shift, and where the decision is to stop or substantially redirect. The cases in this research reference adjustment and iteration, but none document a decision to cancel. That gap may reflect selection bias in which stories get told, or it may reflect a genuine tendency for sunk cost and stakeholder expectation to override data. The capacity to act on difficult evidence remains the most underdeveloped part of the leadership capability measurement skillset in the profession, and one of the most important to build.

Recommendations

The following recommendations are drawn from what the evidence demonstrates works in practice for measuring leadership capability. They are sequenced deliberately, starting with the foundations that must exist before meaningful measurement is possible, moving through measurement design, and finishing with how evidence should be used. Not all will apply at your current stage of maturity. Use them as a diagnostic as much as a to-do list. If your foundations are not yet in place, start with recommendations 1 and 2. If they are, recommendations 3 through 6 are your highest-leverage next moves.

Set the foundations

1. Be clear on what you are measuring and the decision it needs to support

The first discipline is maintaining the distinction between the intervention and the outcome. Programs, workshops, coaching, and cohort experiences are the inputs. Leadership capability observed in role over time is what measurement is trying to understand. Design your evidence chain around indicators of capability change, not indicators of program completion or satisfaction.

Before selecting an evaluation approach, ask: what decisions does this evidence need to support, and who needs to make them? The clearest leadership capability measurement approaches in this research all flow from a commercial or strategic question. Do we have the leadership pipeline to meet the business we are forecasting? Is our leadership culture shifting in the direction the organisation needs? Are we developing leaders fast enough to reduce our dependence on external hiring? When your measurement chain starts with a question like that, every evidence point you collect has a reason to exist. Without a clear decision at the end of the chain, measurement defaults to reporting. And reporting, however well presented, rarely changes what gets invested in or how programs are designed.

The question that should anchor your measurement design is: what is different about how leaders in this organisation think, decide, and behave, and what evidence do we have for that?

2. Build the foundations before the programs

If your organisation does not yet have clear leadership capability expectations anchored to observable behaviours, a workable performance process that surfaces development signals, managers equipped to have capability conversations, and basic data infrastructure to hold

evidence over time, address these first. Programs built on unstable foundations cannot produce credible evidence of leadership capability change. The foundations that matter most are: a capability framework that describes what good leadership looks like at each level in behavioural terms; a performance cadence that creates regular touchpoints where capability evidence can surface naturally; manager readiness to participate in goal-setting, observation, and structured follow-up; and a data system capable of comparing evidence across cohorts and time. All of them are prerequisites.

Design the evidence system

3. Make measurement sustainable: assign ownership, build the embed phase in, and keep it repeatable

When measuring leadership capability is everyone's responsibility, it becomes no one's. Before a program launches, name who owns each evidence point, when it will be collected, and who is accountable for reviewing and acting on it. That ownership sits close to the program, not in a separate reporting function.

Ownership alone is not enough. One of the most consistent gaps between well-designed programs and credible evidence of capability change is the embed phase: the structured period of reinforcement, practice, observation, and adjustment that determines whether what was learned becomes what is done in role. Build it into the architecture, resource it, and stay with the cohort. The programs generating the most credible leadership capability evidence in this research were still actively engaged with participants at three, six, and in some cases eighteen months after formal delivery concluded. That is not yet the norm in the profession. It needs to become one.

The final check is whether your measurement system can actually run. The most useful leadership capability evidence in this research is not the most sophisticated. It is the most repeatable. A follow-up call that the team actually makes is more valuable than a 90-day survey that sits unanswered. When designing your evidence system, ask: can we realistically do this consistently, every cohort, for the next two years? If the answer is no, simplify until it is. Repeatable evidence that can be compared across time and cohorts is the foundation of a credible leadership capability story.

4. Design for transfer before you design for measurement

The organisations with the strongest evidence of leadership capability change are also the ones with the most deliberate transfer architecture. Nomination requirements that signal accountability before a program begins. Spaced delivery across months rather than compressed into days. Applied projects with real stakes and real audiences. Manager calibration of goals at the start, with structured check-ins across the program. Peer accountability through group coaching or cohort touchpoints. These are transfer design choices, and they are also what makes leadership capability measurement meaningful, because they create the conditions under which application evidence can be collected. If a program is not designed for transfer, measurement will consistently confirm that transfer is weak.

5. Invest your strongest evidence at points of consequence

You cannot measure leadership capability everywhere, and trying to do so produces administrative burden without proportional insight. Identify where leadership behaviour carries the most organisational weight, where promotions are made, where culture is set, where the most critical delivery decisions happen, and concentrate your highest-quality evidence there. Multi-rater feedback, structured development planning, and longitudinal capability tracking belong at those points. Signal evidence, confidence ratings, and end-of-

session intentions can sit usefully elsewhere in the chain. Ask yourself: where in this organisation does leadership capability have the most consequence? That is where the investment in stronger evidence is justified and most credible.

6. Use three tiers of evidence quality as a practical diagnostic

To assess where your current measurement sits and what to build next, consider three tiers. These are not a strict quality hierarchy, a well-run pre/post competency assessment can produce stronger evidence than a poorly designed 360. They are a guide to the type and proximity of evidence, and where to invest as maturity grows.

- Tier 1 is signal evidence: participation rates, satisfaction ratings, confidence self-reports, end-of-session intent. Useful as leading indicators of engagement, insufficient as standalone evidence of leadership capability change.
- Tier 2 is application evidence: competency assessments with manager and self-rating, documented behavioural commitments with structured follow-up, applied work outputs reviewed against criteria, transfer check-ins at 60 or 90 days, impact plan reviews. This is where the majority of the most decision-useful leadership capability evidence in this research sits.
- Tier 3 is consequence evidence: multi-rater 360 at key leadership transitions, internal promotion and succession readiness data, culture indicators tracked over time, operational metrics connected to leader performance, risk-level reporting to executive or board.

The aim is not to reach Tier 3 across every initiative. It is to operate at Tier 2 as a consistent baseline and reach Tier 3 deliberately where the leadership stakes are highest.

7. Build manager involvement into the measurement design structurally

Every credible transfer story in this research has manager accountability embedded structurally. Development goals co-developed with managers expected to set stretch targets. Nomination processes that make managers accountable for a participant's development before it begins. Operational leaders positioned as assessors of leadership capability, not passive observers. Manager involvement does not have to be complex. A structured goal-setting conversation at program entry, a midpoint check-in against stated commitments, and a final assessment of capability in role are sufficient. But they must be designed in, expected, and followed up.

Use evidence to govern investment

8. Adopt contribution logic as your default position on organisational outcomes

As practitioners we might want to rethink trying to prove that leadership development caused business results. Build the most credible contribution story you can: a chain of plausible data points and observations that connects what the development did, to what changed in individual leadership capability, to what shifted in team performance, to what moved at the organisational level. That story is more persuasive to senior leaders than a contested ROI calculation and more defensible when challenged. Present it honestly, including what it cannot prove. Stretching evidence further than it can credibly reach does not strengthen the case for investment in leadership capability. It undermines it.

9. Document what you decided because of the evidence

At the end of every measurement cycle, identify the specific decisions the evidence supported. Did findings change the next module? Trigger targeted support for specific leaders? Surface a facilitation gap? Inform a recommendation to scale or sequence differently? If evaluation of leadership capability consistently produces reports rather than decisions, it is functioning as

compliance activity. Over time, documenting what changed because of measurement findings builds one of the most compelling cases you can make to senior stakeholders: not "our programs work" but "our measurement makes us smarter about developing leadership capability."

10. Build the capability to act when evidence says something isn't working

The hardest evaluation skill is not designing a measurement architecture. It is using evidence to stop, substantially redesign, or redirect investment when findings indicate that leadership capability in the organisation is not shifting as intended. If your evaluation system is only ever used to justify continuing, it is not functioning as a genuine decision tool. Evidence that challenges the current approach is valuable, not inconvenient. The organisations that will build the most credible leadership capability measurement reputations are the ones known for changing course when the data says they should.

Collectively, these practices reposition leadership capability measurement from a retrospective evaluation activity to an operating system embedded in how the organisation builds, tracks, and governs leadership capability over time.

Leadership Capability Measurement Readiness Check

Before you design or commission leadership capability measurement, work through these seven questions. A no, or an honest "not yet" tells you where to focus first.

1. Have we defined what good leadership looks like in observable behavioural terms?
2. Do we know what evidence would confirm those behaviours are shifting in role?
3. Have we designed for transfer before we designed for measurement?
4. Are managers part of the measurement process, not just the delivery?
5. Is evidence reviewed inside an existing operating rhythm, not as a separate reporting exercise?
6. Do we know what decisions this evidence will inform, and who will make them?
7. Is someone named and accountable for follow-through?

If you can answer yes to all seven, your measurement approach has the foundations it needs. If not, the gaps here are more important to close than any choice of framework or tool.

Building our own Capability to do this work

The capabilities this research asks of practitioners sit inside the Evaluation and Impact domain of the AITD Practitioner Capability Framework. The capabilities at the centre of this work are evaluation design, data capture and management, data analysis and insight generation, and impact measurement and decision-making. Across the cases, these are the capabilities most consistently identified as underdeveloped, and the ones that most directly determine whether leadership capability evidence gets used to make decisions or produced to satisfy compliance.

Two other capability domains are worth noting. The Operations domain, particularly L&D governance and building L&D team capability, reflects the system conditions Pattern 2 and Pattern 6 identify as foundations that must exist before meaningful measurement is possible.

The Strategy and Partnering domain, specifically organisational capability building and business partnering, reflects the positioning work Pattern 4 requires: framing leadership capability evidence as a business question, not an L&D activity.

Practitioners looking to develop in these areas can use the AITD Practitioner Capability Framework as a diagnostic for where to invest their own development first. For practitioners, the implication is clear: evaluation capability can no longer sit at the end of the learning lifecycle. It needs to be part of diagnostic practice, program architecture, stakeholder contracting, data design and governance. The practitioner capability required is both technical and relational: the ability to frame capability as a business question, design evidence that can survive scrutiny, and facilitate decisions from what the evidence shows.

Conclusion

A consistent pattern emerges; organisations are not short on frameworks, models, or ways to measure leadership capability. What they are short on is the deliberate decision to treat measurement as a core part of how leadership capability is built, rather than an activity that follows delivery.

Where leadership capability measurement is strongest, it is not because organisations have adopted a particular methodology in full. It is because they have made a series of conscious choices. They have defined what leadership capability looks like in their context. They have anchored evidence in real work rather than participant perception. They have built follow-up and transfer into the architecture, not as an afterthought. And critically, they have used evidence to make decisions - to adjust, to target, to scale, and, where necessary, to stop.

Moving beyond attendance is not about adding more data. It is about changing what counts as evidence and being honest about what that evidence can and cannot say. Participation and satisfaction will always have a place. But they are signals of engagement, not indicators of capability. Treating them as proof of impact is not a measurement limitation, it is a decision.

The deeper challenge is one of time and discipline. Leadership capability does not shift at the pace of a program. It shifts through application, reinforcement, and adjustment over months and years. Measuring it credibly requires staying with that process - building longitudinal evidence, engaging managers, and holding attention beyond the point where most programs end.

That is not an easy shift. It requires different resourcing, different ownership, and a willingness to prioritise work that is less visible than delivery but ultimately more consequential.

Without credible leadership capability evidence, organisations are left making investment decisions based on participation, perception, and assumption. With it, leadership capability becomes something that can be understood, shaped, and governed - not just delivered.

References

Alliger, G.M., Tannenbaum, S.I., Bennett, W., Traver, H. and Shotland, A. (1997) 'A meta-analysis of the relations among training criteria', *Personnel Psychology*, 50(2), pp. 341–358.

Association for Talent Development (2023) *State of the Industry Report*. ATD.

Brinkerhoff, R.O. (2003) *The Success Case Method: Find Out Quickly What's Working and What's Not*. San Francisco: Berrett-Koehler Publishers.

Cendros, R. (2025) 'Beyond learner reaction: Measuring the impact of leadership development at The Ivey Academy', *International Journal of Advanced Corporate Learning*, 18(1), pp. 54–63.

Deloitte (2022) *The skills-based organisation: A new operating model for work and the workforce*. Deloitte Insights.

Garavan, T.N., McCarthy, A., Sheehan, M., Lai, Y., Saunders, M.N.K., Clarke, N., Carbery, R. and Shanahan, V. (2019) 'Measuring the organisational impact of training: The need for greater methodological rigor', *Human Resource Development Quarterly*, 30(3), pp. 291–309.

Holton, E.F., Bates, R.A. and Ruona, W.E.A. (2000) 'Development of a generalized learning transfer system inventory', *Human Resource Development Quarterly*, 11(4), pp. 333–360.

Kirkpatrick, J.D. and Kirkpatrick, W.K. (2016) *Kirkpatrick's Four Levels of Training Evaluation*. Alexandria, VA: ATD Press.

Leaman, C. (2016) 'Measuring what matters most in your training', *TD Magazine*, March, pp. 76–79.

Mayne, J. (2011) 'Contribution analysis: Addressing cause and effect', *Evaluation*, 17(4), pp. 363–380.

McKinsey & Company (2021) *Building leadership capability for the future*. McKinsey & Company.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: Sage.

Sitzmann, T., Brown, K.G., Casper, W.J., Ely, K. and Zimmerman, R.D. (2008) 'A review and meta-analysis of the nomological network of trainee reactions', *Journal of Applied Psychology*, 93(2), pp. 280–295.

Thalheimer, W. (2018) *The Learning-Transfer Evaluation Model (LTEM)*. Available at: Work-Learning Research website (Accessed: 1 Feb 2026).

Weiss, C.H. (1995) 'Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives', in Connell, J.P. et al. (eds.) *New Approaches to Evaluating Community Initiatives*. Washington, DC: Aspen Institute.

Yaqoot, E.S.I., Wan Mohd Noor, W.S. and Mohd Isa, M.F. (2017) 'Factors influencing training effectiveness: Evidence from public sector in Bahrain', *Oeconomica*, 13(2), pp. 31–44.

Appendices

Case Study: Richard Bligh, GO.FARM

Measuring Leadership Capability Across a Pipeline Built for Scale

Richard Bligh is Head of People & Performance at GO.FARM, responsible for building the leadership and people capability required to support rapid organisational growth.

He is known for a practical, systems-focused approach that connects leadership development to business performance, culture and long-term scalability. His focus is on creating clear expectations, structured development pathways and stronger readiness for progression into larger roles.

Bligh leads the design of leadership capability across multiple levels of the business, from graduates and emerging leaders through to farm managers and executives. His remit spans leadership development, succession readiness, behavioural standards, coaching and the systems that help capability grow in step with the business.

His work reflects a clear belief: sustainable growth depends on building capable leaders early, consistently and at scale.

Organisation Snapshot

GO.FARM is an Australian-owned agricultural investor, developer and manager focused on sustainable growth across Australian agriculture. The business has grown from around 10 people to 160 over the past decade and is forecasting 400+ people and $2.5b AUM by 2030. Richard Bligh, Head of People & Performance, leads capability development across that growth trajectory.

Leadership expectations at GO.FARM are anchored in "The GO.FARM Way" a set of values and behaviours all team members are held accountable to, including Curiosity, Courage, Collaboration and Care, alongside standards such as being solutions-focused, driven and fun.

"We're scaling fast, and leadership capability & consistency is the constraint. If we don't build the bench deliberately, our system breaks down. Standards drift, decision quality varies, and culture becomes inconsistent across sites."

As GO.FARM's headcount and operational complexity rise, the leadership bench becomes the critical risk. Richard's priority is building capable, consistent leaders across farms and the broader business, leaders who can execute under pressure, empower their teams, and role-model The GO.FARM Way as the organisation scales toward 400 people.

Success is defined as three things:

1. Behavioural consistency: leaders and team members living The GO.FARM Way in day-to-day decisions and interactions.
2. Capability in role: visible uplift in the leadership capabilities that drive execution, decision quality, collaboration, influence, change leadership, and problem solving.
3. Pipeline strength: clearer readiness for progression into larger roles over time, supported by repeatable evidence at each level.

A Pipeline Approach to Leadership Development

Rather than treating leadership development as a single program, GO.FARM has built a pipeline across multiple levels, each with distinct measurement and reinforcement mechanisms. The logic is deliberate: different levels of leadership need different measures, not just different content.

The Graduate Program serves as the capability-building entry point, with two pathways: a 12-month farm-based stream and an 18-month Agribusiness stream. The program builds self-awareness, resilience, commercial awareness and stakeholder management, providing graduates an opportunity to understand themselves, early level leadership concepts, on-ground practical experience and networks across the business.

Future Farm Leaders (FFL) is a two-year program for emerging farm leaders, with modules spanning self-awareness and personal impact, team dynamics, change, applied operational problem solving, and leading with courage and care. Modules are delivered twice yearly, with practical projects between sessions that embed learning directly into farm operations. "Participants are nominated and endorsed by their leaders," Richard notes, "which means there's already a built-in layer of accountability before the program even begins."

The High-Performing Leadership Program (HPLP) operates at the next level of leadership responsibility. Content progresses to judgement, execution through others, commercial discipline, and leading across farms and partners, mirroring FFL's emphasis on application but at greater scope and complexity.

At the General Manager stage, multi-rater evidence comes into play. Richard describes this as the point where leadership behaviour is most visible and consequential, and therefore where the strongest evidence of impact is both possible and necessary. Saville Wave 360 provides multi-source evidence of observed leadership, followed by structured development planning and twice-yearly Multipliers group coaching.

At the Executive stage, GO.FARM also uses Saville Wave 360. Here the purpose extends beyond individual development to provide a senior-level read on culture stewardship, alignment to The GO.FARM Way as the organisation scales toward 400 people, all through the lens of Legacy.

Measures and Cadence

In short, GO.FARM's measurement chain is: capability expectations (The GO.FARM leadership framework) → psychometric insight and development focus (Saville Wave Professional Styles at FFL and HPLP) → transfer intent and follow-through (post-session commitments, action and development plans) → observed behaviour and impact at a critical transition point (Saville Wave 360 at General Manager and Exec) → reinforcement to sustain change (development planning and twice-yearly Multipliers coaching).

Three mechanisms hold this together.

A consistent psychometric across levels. Saville Wave is embedded through the pipeline with Focus Styles at Graduate level, Professional Styles is used at both FFL and HPLP levels, creating a coherent evidence base as leaders progress through the pipeline. At General Manager and Exec level, Saville Wave 360 provides multi-rater feedback on leadership behaviour and impact. " The goal is to create a consistent view of development focus over time, leading to targeted development and career growth," Richard explains, "not a one-off assessment that tells you where someone sits on a single day."

Application evidence built in, not optional. Post-session surveys capture what leaders will start, stop, or continue, along with specific behavioural commitments with timeframes. Action and development plans record longer-term commitments that are followed up through normal performance and development rhythms.

Early data signals strong engagement: 94% of participants (30 of 32) completed the recent post-session HPLP survey. All reported being clear on what good leadership looks like at GO.FARM, and a very high proportion felt confident they could apply at least one tool or approach within two weeks.

The qualitative evidence is equally telling. When asked what specific behaviour they would change in the next seven days, participants gave grounded, concrete responses: *"I will offer responsibility, not tasks"; "Stop going straight to providing the solution when team members ask how to approach something"; "Start doing brief daily check-ins with the team each morning."* When asked what might get in the way, they named time pressure, harvest demands, and the pull of old habits, practical transfer barriers, not disengagement. "That kind of evidence gives us a far richer picture of learning impact than satisfaction scores alone," Richard says.

These commitments provide useful leading evidence of transfer intent. They become stronger evidence of capability shift only when followed up through manager observation, action plan review, or multi-rater feedback.

Reinforcement and transfer designed into the system. At General Manager level, the 360 provides multi-source evidence, leaders complete development planning to translate insights into behaviour change commitments, and Multipliers coaching creates peer accountability for sustained change. "The benefit we're already seeing is clearer expectations and better leadership conversations," Richard reflects. "People are more explicit about what 'good' looks like, what they'll do differently on-farm, and how they'll follow through."

Building and Measuring at the Same Time

One of the distinctive challenges Richard is navigating is that GO.FARM is not evaluating a finished program, it is building and refining one in real time, against a backdrop of rapid scale. "Rather than designing the full program upfront, we're building and refining it iteratively across cohorts," Richard explains, "using emerging on-farm pressures and post-session feedback to shape the next module."

Richard shared "the post-session surveys and commitment data keep me honest. If someone writes down that they'll 'offer responsibility, not tasks' and nothing changes on farm six weeks later, that's on us to follow up, not ignore."

This iterative design has a direct implication for how evidence is used: measurement has to be reliable enough to track progress and flexible enough to evolve. Findings from each cohort actively shape the next module, reinforcing what's working, adjusting where transfer is weak, and targeting support for individual leaders before patterns become problems. Richard is clear "The thing I've had to accept is that we can't measure everything perfectly while we're still building the program. So we've had to be intentional about where we invest our strongest evidence, and comfortable that some signals are directional rather than definitive."

How Evidence Will Be Used

Within this approach, the evidence closest to leadership behaviour and impact in role carries the most credibility: multi-rater evidence of observed behaviours at General Manager level, repeatable psychometric insights that inform development planning across the pipeline, and documented application commitments with follow-through tracked via action and development planning. These carry more weight than attendance or satisfaction data because they provide clearer line-of-sight to actual leadership behaviour.

In practice, evidence is used to adjust module emphasis and reinforcement where cohorts show weak transfer; to target support for individual leaders using Professional Styles, action

plans, and 360 data where applicable; and to strengthen the pipeline by clarifying what readiness looks like at each level and what development is required to progress.

Connecting to the Frameworks

GO.FARM's approach aligns with several of the evaluation and capability frameworks explored in this paper:

- Outcomes chain thinking: The logic is explicit — capability expectations (GO.FARM framework) → psychometric insight (Professional Styles) → transfer intent (commitments, action plans) → observed behaviour and impact (360 at General Manager and Exec) → reinforcement to sustain change (development planning and Multipliers coaching).
- LTEM-style evidence quality: The approach deliberately moves away from attendance and satisfaction towards higher-value signals — documented application commitments, follow-through artefacts, and multi-rater evidence of behaviour in role at General Manager and Exec levels.
- Readiness and progression lens: Repeated use of a consistent assessment suite creates a coherent development language across the pipeline, while multi-rater feedback, action planning and follow-up provide stronger evidence of leadership behaviour over time.
- Transfer conditions thinking: Transfer is supported through explicit application expectations (projects and commitments), structured follow-up through action and development planning, and reinforcement via coaching — consistent with what LTSI-related research identifies as the critical conditions for learning to stick.

What it takes to measure leadership at scale:

- Build measurement into the pipeline, not onto it. If evaluation is an afterthought, it measures the wrong things. By embedding psychometrics, application commitments, and 360 at specific levels, GO.FARM creates measurement that is structural not an evaluation bolt-on that arrives after the fact.
- Use the same tool across levels to track progression, not just position. A single psychometric snapshot tells you where someone is. Repeated use across a multi-year pipeline tells you whether they're moving and in what direction.
- Treat what people say they'll do as data. Post-session commitments are not just engagement signals. When a participant commits to "offering responsibility, not tasks," that is a measurable behavioural intent. Capture it, follow it up, and let it shape what you do next.
- Invest your strongest evidence at the point of highest consequence. Multi-rater feedback, structured coaching, and development planning belong where leadership behaviour matters most. For GO.FARM, that's General Manager level. Know your equivalent.
- Nomination and endorsement are not just admin. Requiring leaders to nominate and endorse participants builds accountability before the program begins. It signals that development is a leadership responsibility, not an HR activity.
- Design for the organisation you're becoming. "At 160 people, a leadership pipeline feels ambitious," Richard reflects. "At 400, it's the only thing standing between a strong culture and a noisy system." The time to build it is before you need it.

Case Study: Kate Fraser, IPAA Victoria

Measuring Capability Impact Across Leadership and AI

Kate Fraser is an organisational development and program delivery leader known for designing and scaling capability initiatives that deliver measurable impact. Across her career, she has consistently positioned learning and development as a strategic lever, underpinned by clear evaluation frameworks, defined success metrics and a strong focus on return on investment. Her approach moves beyond participation and satisfaction, emphasising behaviour change, capability uplift and organisational outcomes.

As Executive Director, Program Delivery at IPAA Victoria, Fraser leads a portfolio spanning professional development, flagship events and thought leadership for the Victorian public purpose sector. Delivering professional development services to a workforce of more than 350,000 professionals, she is focused on delivering scalable, evidence-based capability solutions that respond to shifting sector needs.

Leading a small, high-performing team, Fraser oversees a comprehensive public training calendar, bespoke in-house programs and large-scale events that engage thousands of participants each year. Central to her approach is the integration of customer insights, sector trends and capability frameworks to inform program design, alongside robust measurement practices that track engagement, learning application and impact over time.

With increasing demand for leadership development and AI capability, Fraser works closely with senior stakeholders to design programs that are both practical and future-focused. Her work demonstrates how disciplined evaluation, combined with responsive design, can strengthen workforce capability and deliver tangible value to organisations and the communities they serve.

This case study focuses on the measurement approach Fraser has used most successfully in practice (the de-identified leadership program) and shows how she is applying the same measurement logic to an emerging AI pilot at IPAA.

Fraser's approach to measuring impact has been shaped across two contexts: a leadership program designed and delivered within a Victorian public sector entity, and an AI capability pilot currently being developed at IPAA. Though the settings differ, the measurement thinking that connects them is consistent: start with the problem you are solving, define what success looks like early, and resist the temptation to claim more than the evidence supports.

A Leadership Program (De-identified)

Before her role at IPAA Vic, Fraser led the design of an eight-month leadership program for people leaders in the public sector. The program combined Gallup Clifton Strengths profiling accompanied with pre- and post-program competency assessments rated by both participants and their managers. Rather than compressing content into an intensive bootcamp, the program was deliberately spaced across distinct modules, each followed by an applied group project presented to the cohort and a member of the executive team. Participants also received three coaching sessions with accredited Clifton Strengths coaches.

This design generated multiple forms of evidence. The gap between self-reported and manager-reported competency provided a useful signal of growth and self-awareness. A "confidence worm" tracked confidence ratings across each module, mapped how participants' self-reported confidence moved over time. Executive presentations created visibility for senior leaders to observe application of learning, while a basic rubric assessed whether participants

were drawing on module content and demonstrating critical thinking. Evidence from the pilot informed course redesign for subsequent rollouts, identified growth and support areas for the facilitation team, and built stakeholder confidence through transparency, what Fraser describes as "showing our work along the way."

Success was defined as measurable uplift in leadership competency (self and manager-rated), observable application through applied projects, improved workplace collaborations and evidence that the program was credible enough to influence redesign decisions and build leader demand for future cohorts.

An AI Capability Pilot (IPAA Victoria)

At IPAA Victoria, Fraser is applying similar thinking to a new context. In response to member demand, IPAA is piloting two AI courses designed specifically for the non-tech workforce: a foundations course addressing practical skills such as prompting and responsible use, and a leadership of AI course aimed at leaders sitting on governance boards without technical backgrounds. As Fraser explains, "Now that AI is in the hands of the non-tech workforce, our focus is on how we bring them the skills and the information to thrive and add value"

The courses are anchored to the Victorian Public Sector Commission (VPSC) Capability Framework, which provides a sector-wide reference point for expected capability levels. The planned measurement approach includes pre- and post-program self-ratings on capability and confidence, an in-session intentions worksheet where participants commit to specific actions over the following three months, and a follow-up phone call at the three-month mark to check progress against those stated intentions. Fraser also sees potential in convening learning leaders across departments to develop a broader picture of AI capability across the sector.

In short, Fraser's measurement chain is: define success upfront → establish a baseline (self plus manager where possible) → create work-anchored application evidence (projects, rubrics, observed practice) → track a small number of signals over time (confidence and follow-up) → use findings to redesign and strengthen delivery.

Measures and Cadence

Across both initiatives, Fraser keeps measurement deliberately lean, focusing on a small number of evidence points that show progress beyond participation:

- Baseline and post measures (leadership program): competency ratings captured from both participants and managers to indicate perceived capability uplift and calibration gaps.
- Confidence tracking (leadership program): a simple "confidence worm" to track confidence shifts across modules over time.
- Application evidence (leadership program): applied group projects presented with executive visibility and assessed against a basic rubric (use of module content, critical thinking, practicality).
- Pre/post self-ratings (AI pilot): capability and confidence self-ratings anchored to the VPSC Capability Framework.
- Transfer follow-up (AI pilot): an in-session intentions worksheet (commitments for the next three months) plus a three-month follow-up check to test progress against those intentions.

How Evidence Will Be Used

Fraser uses evidence primarily as a decision tool rather than a reporting artefact. In the leadership program, early measurement helped redesign subsequent rollouts, identify facilitation development and support needs, and build credibility with stakeholders by "showing the work along the way". In the AI pilot, measurement is designed to be light enough to run consistently while still generating follow-up evidence of transfer.

In practice, evidence is used to:

- Stop/adjust: redesign content or facilitation where confidence dips or applied work shows weak transfer.
- Scale/sequence: prioritise cohorts and rollout order based on where uplift is strongest and demand is highest.
- Target support: identify who needs additional coaching or different supports using manager-rated gaps and application evidence.
- Build advocacy: amplify signals that matter (participant pull and manager endorsement) rather than relying on satisfaction scores alone.

What Needs to Be True for Learning to Stick

Across both initiatives, Fraser is attentive to the conditions that enable transfer. For the leadership program, enablers included executive signalling (a personal invitation from the organisational head), manager permission for protected asynchronous learning time, spaced delivery rather than compressed bootcamps, and a pilot cohort who became champions, presenting at all-hands briefings and filming testimonials that created peer-to-peer advocacy.

For the AI foundations course, Fraser's view is pragmatic: participants need time and space away from emails and messaging to practise, feel safe experimenting, receive guidance from an expert and learn from peers. "They just haven't carved out time to sit down with it, be uncomfortable with it, play around with it," she notes. "They tell us that they need the time and space and guidance to be able to get started."

Connecting to the Frameworks

Fraser's approach aligns with several of the evaluation and capability frameworks explored in this paper:

- Outcomes chain thinking: define the problem and success measures → establish a baseline (self and manager where possible) → generate work-anchored evidence of application (projects assessed with a rubric and executive visibility) → run a lean follow-up check (AI pilot) → use findings to refine design, delivery, and sequencing.
- Job performance evidence quality: emphasises evidence closest to job performance and application (manager ratings, observed project outputs, follow-up against intentions) rather than relying on reaction or satisfaction data.
- Contribution logic: explicitly frames impact as contribution, separating what the data shows from the broader narrative and avoiding over-claiming attribution.
- Transfer conditions thinking: treats executive signalling, protected time, spaced delivery, and psychological safety to practise as prerequisites for application in role, not optional extras.

What it takes to measure leadership without over-engineering:

- Workshop a list of high priority challenges with decision makers, design to solve for those challenges to see biggest rewards.
- Treat imperfect data as useful data. It doesn't need to be an "A+ report card" to be decision useful.
- Use manager-rated uplift as a credibility anchor. The self vs manager gap is a stronger signal than sentiment alone.
- Build application into the design. Applied projects and executive visibility create evidence of implementation, not just knowledge and visibility to the leadership team.
- Measure transfer without chasing people forever. Use in-session intention setting plus a three-month check-in to see what stuck.
- Never be the only advocate. If participants and managers are asking "when's the next cohort?", that's a stronger signal than satisfaction scores.

Fraser's approach shows how to build credible impact evidence without over-engineering measurement, by focusing on a small number of work-anchored signals that leaders can use

Case Study: Pete Howell, Hickory

From Activity Reporting to Decision-Useful Capability Evidence

Pete Howell is Chief People Officer at Hickory, leading the People agenda during a period of rapid commercial expansion and organisational scale-up. Known for his pragmatic, commercially grounded approach, Howell focuses on building people systems that directly support delivery performance.

Drawing on experience in operational environments and Agile ways of working, Howell brings a practical focus on clarity, accountability and disciplined execution. His philosophy is clear: capability development should deliver measurable business value, strengthen leadership pipelines and reduce delivery risk.

Leading a lean people function of five, Howell is responsible for designing the workforce foundations required to support Hickory's next phase of growth. His remit spans leadership capability, succession planning, graduate pathways, performance systems and the infrastructure needed for a billion-dollar business to scale successfully.

His work reflects a foundations-first belief: when capability systems are clear, practical and commercially aligned, growth becomes more achievable and sustainable.

About Hickory

Hickory is a privately owned Australian construction company with a workforce of approximately 1,050 people. Around 350 are salaried management and professional staff, with the remainder in construction and manufacturing roles across Data Centers, high-rise developments, vertically integrated manufacturing and associated operations. The business is headquartered in Melbourne and operates predominantly across commercial and residential construction sectors.

Revenue has doubled in two years, from approximately $500 million to over $1 billion, and the business is targeting $2.5 billion within the next three years, driven by significant activity in data centre construction, government partnerships, and institutional housing. Pete Howell, Chief People Officer, leads a people function of five.

Hickory's growth trajectory has created a single, defining challenge: ensuring the business has a pipeline of capable project managers, contracts managers, and site managers to deliver the volume and complexity of work ahead. As Howell explains: "The pipeline of work isn't the problem. What we've identified is the big risk to our business is not having the capability."

The challenge is compounded by two structural factors. First, the small-to-medium projects that once served as developmental stepping stones no longer exist. Where a project manager might previously have cut their teeth on an $80–90 million build before taking on a $200 million project, those mid-range jobs have effectively disappeared. Second, Hickory has a strong preference for internal promotion over external recruitment, not as sentiment, but as risk management. "The risk is the culture we've built in Hickory gets diluted the more externals you bring in." The business needs to grow its own leaders, faster, and with greater rigour.

Defining Success

The primary indicator Howell has set is concrete and quantifiable: fill approximately 80 per cent of vacancies through internal promotion. If the talent pipeline and development approach are working, this target should be achievable. If gaps appear in succession readiness, it signals the development architecture needs attention.

This metric is deliberately commercial. It connects capability investment directly to business continuity and growth execution, rather than measuring training activity. As Howell puts it: "You can spend a lot of money in this space, turn around, and go, so what did we achieve?"

The Initiative: Foundations Before Programs

Rather than launching a leadership program first and measuring it second, Hickory has taken a deliberate "foundations first" approach. Howell describes this as building the measurement plan before layering programs on top of it. The sequence is intentional:

1. Behavioural Framework

A new behaviour framework has been developed from scratch, specific to Hickory, covering three levels: leading self, leading others, and leading the enterprise. This is designed to make expectations visible and assessable across the entire workforce, a prerequisite for any meaningful evaluation of whether development is shifting behaviour.

2. Simplified Performance Signals

The existing performance review process, a five-point scale across fifteen competencies, has been deemed unworkable. Howell is candid about why: "We're asking people to do a process that HR designed, that looks like HR designed for HR, and it doesn't help deliver operational outcomes." The replacement is radically simpler: three ratings (exceptional year, successful year, challenging year) combined with a promotion readiness assessment. The readiness assessment will use three stages: well-placed for now, promotion in the next 12 to 18 months, or significant advancement in the next three to five years, combined with the performance rating to form a standard nine-box for decision-making. The aim is to generate decision-useful data rather than compliance-driven paperwork.

3. Quarterly Execution Cadence

Howell is moving toward a quarterly objective cycle: what did you say you would do, did you do it, and what are you doing next quarter? This cadence, informed by his background in Agile methodology, creates a recurring feedback loop that connects individual execution to business priorities.

4. Competence Verification

A critical gap has been identified in the graduate program: there is currently no consistent formal assessment of whether participants have developed the expected competencies. Graduates progress through modules and complete a two-year program without structured verification. The refresh will introduce assessments, using AbilityMap to assess alignment with Hickory's way of working behaviours, subject matter experts to assess technical competence, and managers to assess overall progress against the pre-defined capabilities expected across the two-year program. Hickory also builds its graduate pipeline through a cadet program, with an RPL process in place to support a faster pathway for high achievers. Howell is unequivocal on what these assessments should measure: "I have no interest in whether they enjoyed the day or whether they had fun. I want to know, did this help you do your job?"

How Evidence Will Be Used

The data architecture Hickory is building is designed to serve specific decisions:

- Fast-tracking and development: Identifying who has had an exceptional year and needs accelerated development or promotion, and who has had a challenging year and needs targeted support or a different conversation.

- Pipeline planning: Quantifying how many project managers are needed over the next three years, assessing whether internal candidates are ready, and identifying gaps before they become delivery risks.
- Board and executive reporting: Moving from qualitative assertion ("we don't have the capability") to quantified risk ("we need ten project managers, we have four ready, and here is the development plan for the next six").
- Manager accountability: Asking operational leaders directly whether development interventions have improved their team's capability and delivery and being prepared to change course if they haven't.

The Hardest Part

Howell is transparent about what Hickory cannot yet do. The business does not currently have a human resources information system capable of capturing and collating people data in a usable format. Implementing one is the immediate priority for the first half of 2026, and it represents a significant undertaking. Without it, the people function cannot automate processes, generate reliable reporting, or support the business as it scales.

The transformation is planned as a two-year build. Howell acknowledges the discomfort: "The hardest thing has been being comfortable that this could take two years." But the sequencing is deliberate. Getting the foundations wrong, or skipping them entirely, would mean programs are built on unstable ground. As Howell reflects: "Without that foundation, everything else would have failed."

In the interim, the business is not standing still. A partnership with Udemy provides immediate access to AI and professional development content. The graduate and cadet programs continue to run. Operational managers retain discretion to invest in their teams. But none of this is yet connected to the measurement plan that Howell is building. That connection is the next phase.

Connecting to the Frameworks

Hickory's approach aligns with several of the evaluation and capability frameworks explored in this paper:

- Outcomes chain thinking: The logic model is explicit, even if not formally articulated as such behavioural expectations → role clarity → quarterly execution reviews → promotion readiness → internal fill rate → growth delivery. Each element builds on the one before it, and the 80 per cent internal promotion target serves as the measurable outcome at the end of the chain.
- LTEM-style evidence quality: The emphasis on assessing whether participants can actually perform, rather than whether they completed a module or enjoyed it, reflects a focus on decision-making competence and real-world application as the meaningful indicators of learning.
- Readiness and mobility lens: The 80 per cent internal fill target, succession planning pipeline, and promotion readiness assessments align with a talent mobility approach to capability measurement.
- Transfer conditions thinking: The deliberate sequencing, behavioural clarity, simplified reviews, manager accountability, and HR system enablement, addresses the organisational conditions identified as critical to whether learning transfers into sustained performance.

The foundations-first playbook:

Howell's advice for people leaders entering a high-growth business:

- Understand the business before designing solutions. "Take the time to really understand the business before you bring in solutions that you think will be the answer." Howell's first 100 days were spent listening, not building.
- Get the foundations right before the exciting programs. Behavioural expectations, role clarity, and a workable performance process must come before leadership academies or development frameworks.
- Simplify measures so they drive action. If the data from a process cannot be used for decisions within weeks, the process is too complex.
- Let interventions run their course. "You've actually got to let it run" avoid jumping at shadows before a program has had time to demonstrate impact.
- Build a peer network. "Find some people that you can network and vent with and ask questions with, because it's a lonely role."

About AITD

The Australian Institute of Training and Development (AITD) is the leading membership association for professionals in training, learning and development, organisational development and related roles.

AITD provides a range of professional development opportunities including courses, conferences, communities of practice, networking events, online learning and other activities.

Visit www.aitd.com.au for more information.

aitd.

aitd.com.au

PO Box 4093
Macquarie Centre
Macquarie Park NSW 2113

02 9211 9414
enquiries@aitd.com.au

ABN: 52 008 516 701