FOR

program directors, evaluators, data staff, grant writers, executive directors, and anyone deciding what to track

INCLUDES

a six-criteria measure evaluation framework, a measure evaluation table, a behavioral health worked example, a say/cannot-say table, and a 15-minute team exercise

Choosing Measures That Fit the Work

How to evaluate whether a specific measure is worth tracking — and what makes a measure genuinely useful

Your organization has a list of things it could measure. Maybe it is a long list — pulled from reporting requirements, staff suggestions, a new database, and what similar programs track. Maybe a planning document on someone's screen is asking for "at least three outcome measures."

Most organizations have too many candidate measures and too little clarity about which ones deserve the effort to collect, clean, and act on. Every measure costs staff hours, participant time, system complexity, and attention. The work is to choose the measures that earn those costs.

The short answer

A useful measure is one your organization can collect reliably, interpret honestly, and connect to a real decision. A measure that fails on several of those fronts is probably a poor use of effort, even when it appears in a reporting template or another program's logic model.

A useful measurement plan contains a small set of indicators that help real people make better choices about real work.

Why this is harder than it looks

Choosing measures feels like it should be straightforward. You name the change you want to see, find a number that tracks it, and start collecting. But in practice, three problems get in the way.

First, the measures that sound the most important are often the hardest to collect well. “Housing stability” is a meaningful concept, but defining it precisely, collecting it consistently, and interpreting it fairly is harder than it appears.

Second, people anchor on the first measure someone proposes. The first suggestion sticks, and the team moves to logistics rather than asking whether it is the best available indicator for the decision at hand.

Third, there is pressure to measure everything. Boards want dashboards. Public reporting systems and accrediting bodies have indicator lists. Without a shared framework for evaluating candidates, the result is a bloated measurement plan that nobody fully trusts or maintains. Data collected without a clear purpose tends to be reported and filed but rarely used.

Six qualities of a useful measure

When you are deciding whether a specific measure belongs in your plan, test it against six criteria. A strong measure does well on most of them. A measure that fails on several is a warning sign.

1. VALID

Does the measure actually capture what you think it captures?

This is the most fundamental question, and it is easy to get wrong. “Housing stability” measured by “still at the same address after twelve months” may miss people who moved to better housing — and flag people as stable who are trapped in unsafe situations. “Job placement” counted at the moment of hire may miss whether employment lasts or pays a living wage.

If this number goes up, does that reliably mean the underlying condition improved? If you are not confident, look for a better indicator or pair this one with something that fills the gap.

2. OBSERVABLE

Can you actually collect this measure reliably with the staff, systems, and access you have?

A measure may be conceptually perfect but practically impossible. Tracking participants’ health outcomes six months after they leave your program requires follow-up capacity many organizations do not have. Measuring community-level change requires population data you may not be able to obtain.

Observable does not mean easy. It means achievable with your current team, data systems, and participant relationships.

3. TIMELY

Will the data be available when the decision it supports needs to be made?

An annual outcome survey that produces results in September cannot inform a budget decision made in June. Timeliness depends on alignment between when information arrives and when choices happen.

If a measure arrives too late to be useful, it may still be worth collecting for learning — but it is not a decision-support measure.

4. HARD TO GAME

Could staff or participants easily inflate or deflect this number without actually improving the underlying condition?

When a measure is tied to performance expectations or resource decisions, there is natural pressure to make the number look good. A workforce program that counts "job placements" without requiring a minimum duration may find staff prioritizing quick placements over lasting ones. A housing program that counts "referrals made" may see high numbers that do not reflect whether anyone was actually housed.

Examine whether the measure's structure makes it easy for the number to look better than the reality, even through routine decisions and honest shortcuts.

5. NOT TOO BURDENSOME

Does the cost of collecting this measure — in staff time, participant time, and system complexity — justify the value it provides?

Every measure has a cost. A pre-and-post survey takes participant time and goodwill. A six-month follow-up call takes staff hours and may reach only a fraction of people. Burden falls unevenly, so programs serving people in crisis, with limited English, or with low trust in institutions should be especially careful about what they ask participants to do. Good measurement must be accurate, proportionate, and respectful.

A measure can score well on every other criterion and still not be worth collecting if the burden outweighs the value.

6. CONNECTED TO A DECISION

Would a change in this number cause someone in your organization to do something different?

This is the test that eliminates the most waste. A useful measure is one where you can name the person who would respond to a significant change and describe what they would do. If enrollment drops below a threshold, the outreach coordinator adjusts strategy. If referral completion falls, the case management supervisor investigates. If participants report feeling unsafe, leadership acts immediately.

If you cannot identify what would change, the measure is probably serving compliance, not learning.

A measure evaluation table

Before committing to a measure, rate each candidate against the six criteria. A simple assessment is enough to compare options and surface weaknesses.

CRITERION	QUESTION TO ASK	STRONG	WEAK
Valid	Does it capture the real condition?	Tracks what actually matters	Could go up while the real situation worsens
Observable	Can we collect it with available resources?	Fits current staff, systems, and access	Requires capacity we do not have
Timely	Will it arrive when decisions are made?	Available before the relevant choice	Arrives after the decision window
Hard to game	Is the number resistant to inflation?	Difficult to improve without real change	Easy to inflate through reporting choices
Not too burdensome	Is the cost justified?	Low-to-moderate effort relative to value	Heavy staff or participant burden for limited use

CRITERION	QUESTION TO ASK	STRONG	WEAK
Connected to a decision	Would a change cause someone to act?	A named person would respond to a shift	Nobody would do anything different

A measure that is strong on all six is rare. One or two weaknesses can be managed. Three or more weaknesses deserve serious reconsideration, including when the measure appears in an external requirement.

Worked example: a behavioral health program chooses an outcome measure

A community behavioral health program serves adults with co-occurring mental health and substance use conditions. It provides peer support, care coordination, and connection to treatment. The team needs one primary outcome measure for a new grant report. Three candidates are on the table:

Candidate A: Standardized symptom scale (PHQ-9 for depression)

- *Valid:* Yes — well-established and clinically meaningful.
- *Observable:* Partially. Some participants have literacy barriers or decline to answer.
- *Timely:* Can be administered at regular intervals.
- *Hard to game:* Moderately. Self-report can be influenced by context or desire to please.
- *Not too burdensome:* Moderate. Nine questions, but adds to an already long intake process.
- *Connected to a decision:* Unclear. No one has specified what would change if scores improved or did not.

Candidate B: Treatment engagement (attended at least three appointments with a referred provider within 60 days)

- *Valid:* Partially. Engagement does not guarantee improvement, but it is a necessary step the program directly facilitates.
- *Observable:* Yes — tracked through care coordination records the program already maintains.
- *Timely:* Available within 60 days of referral.
- *Hard to game:* Moderately hard. Requires documented attendance at an outside provider, not just an internal referral.
- *Not too burdensome:* Low. Uses existing data.
- *Connected to a decision:* Yes. If engagement rates drop, the care coordination team investigates barriers and adjusts their approach.

Candidate C: Self-reported recovery progress (a brief recovery capital scale)

- *Valid:* Yes for the concept, but the team is unsure which version of the scale to use.
- *Observable:* Partially. Requires a new instrument the team has not piloted.
- *Timely:* Depends on a follow-up schedule not yet established.
- *Hard to game:* Similar to any self-report measure.
- *Not too burdensome:* Unknown. Staff would need training.
- *Connected to a decision:* Not yet. The team has not discussed what they would do with the results.

The team's choice: Candidate B — treatment engagement — as the primary outcome measure. It offers the best combination of clinical relevance, reliable collection, honest interpretation, and practical use. Engagement is a

leading indicator: it signals early whether the program's core service — connecting people to treatment — is actually working.

They decided to pilot the PHQ-9 with a subset of willing participants to see whether it is feasible, before committing to it program-wide. This is what a disciplined measurement decision looks like: usefulness over impressiveness.

Leading indicators: measures that signal early

Some of the most useful measures are not final outcomes at all. They are leading indicators — early signals that things are moving in the right direction (or not).

Leading indicators matter because final outcomes often take months or years to materialize. If you wait for the twelve-month employment retention number to learn whether your workforce program is effective, you have waited too long to adjust.

Examples of leading indicators across program types:

- **Workforce:** Participants complete their first week of employment. Participants report satisfaction with job fit at 30 days.
- **Housing:** Households complete move-in within 30 days of voucher issuance. Tenants make their first three rent payments on time.
- **Youth development:** Participants return after the third session (a common drop-off point). Youth report having at least one trusted adult at the program.
- **Health:** Patients attend their second appointment (indicating engagement, not just initial contact). Referrals to specialty care result in a scheduled visit within two weeks.

- **Food access:** Households return after their first visit. Participants report using the food they received.
- **Transit:** Riders use their pass more than once per week. Riders report reaching their primary destination on time.

Leading indicators work alongside outcomes as an early-warning system. Movement in the wrong direction signals that something needs attention before the final outcome arrives.

What you can say and what you cannot say

WHAT YOU CAN SAY	WHAT YOU CANNOT SAY YET
"We chose this measure because it is the best available indicator we can collect reliably and act on."	"This is the only measure that matters."
"We evaluated several candidates and selected the one that balances validity, feasibility, and decision relevance."	"Our measure perfectly captures the outcome we care about."
"This is a leading indicator — it tells us early whether our core service is working as intended."	"Because the leading indicator looks good, the final outcome will too."
"We decided not to track this measure because the cost of collecting it exceeds the value for our current decisions."	"We do not care about that outcome."
"Our measurement plan is designed to be useful, not comprehensive."	"We have measured everything that matters."

A 15-minute team exercise: evaluate three measures

Choose three measures your organization currently tracks or has been asked to track. For each one, answer together:

1. **Valid:** Does this measure capture the condition we care about? Could the number improve while the real situation stays the same?
2. **Observable:** Can we collect this reliably with the staff and systems we actually have?
3. **Timely:** Is this data available before the decisions it should inform?
4. **Hard to game:** Could someone make this number look better without improving the underlying condition — even unintentionally?
5. **Not too burdensome:** Is the cost of collecting this justified by the value it provides?
6. **Connected to a decision:** If this number changed significantly, who would respond and what would they do?

Then ask: Is there a measure we are not tracking that would score higher than something we already collect? If so, consider the swap.

Review whether the measure still represents the mission

It is easy to start optimizing for the measure rather than for the condition the measure represents. When a workforce program tracks job placements, staff may steer participants toward any available job rather than sustainable employment. When a housing program tracks “days to placement,” the team may prioritize speed over fit.

This is a predictable consequence of measurement, sometimes called Goodhart's Law: when a measure becomes a target, it can lose its value as a measure.

Treat each measure as a partial view of the underlying condition. Check periodically whether your measures still reflect the outcomes you care about and whether the way staff respond to them is producing the behavior you actually want.

When to bring in help

Consider bringing in outside support when:

- you need to choose among several candidate measures and are unsure which best fits your population or decision context;
- you are adopting a standardized instrument and need guidance on administration, scoring, or interpretation;
- an external requirement specifies a measure and you want to evaluate whether it is feasible and appropriate before committing;
- your measures seem to be driving the wrong staff behavior and you need help redesigning incentives without losing accountability;
- you are tracking outcomes across multiple sites or service models and need a measurement framework that holds together;
- you want to evaluate whether your current measurement plan provides enough useful information for the decisions you actually face.