

Can These Programs Be Compared?

Decide whether the purpose calls for ranking, grouping, or learning across difference

FOR

funders, boards, public agencies, portfolio teams, intermediaries, and organizations comparing programs, grants, providers, or sites

USE WHEN

a common dashboard or outcome is being proposed for oversight, allocation, public reporting, or shared learning

INCLUDES

a four-part comparability test, use-first framework, worked example, synthesis options, and a 15-minute team exercise

A board asks for one view of a portfolio. Staff gather outcome rates from every program and place them in a table. The numbers look comparable because they share a column. Yet one program measures employment at exit, another at six months, and a third reports any paid work during the year. Some serve people already close to employment. Others serve people facing substantial barriers.

A common number can create the appearance of fairness while concealing differences that matter.

The short answer

Start with what the comparison will be used to do. Ranking and funding allocation require much stronger alignment than shared learning or descriptive grouping. Check outcome meaning, population, period, and source before placing results on the same scale. When those do not align, use a structured narrative or grouping rather than a false league table.

Programs can be worth learning from together without being fair to rank against one another.

Name the use before the metric

Comparisons usually serve one of four purposes.

Oversight

A board or public body wants to see whether programs are operating, reaching intended groups, managing risk, and responding to problems. Comparable outcome rates may be only one part of the view.

Allocation

A funder or agency wants to choose among programs or direct more resources toward some than others. Measurement differences become consequential because the comparison affects money and access.

Learning

Programs want to identify shared barriers, delivery practices, or useful adaptations. They may learn from one another even when outcomes cannot be pooled.

Public account

An institution wants to describe what a portfolio did and what changed. Aggregation must preserve the meaning of the underlying measures rather than producing one impressive but unstable total.

Do not use the lightest standard because the data is available. Use the standard required by the action attached to the comparison.

The four-part comparability test

Outcome meaning

Are programs measuring the same thing? “Stable housing,” “job placement,” and “engagement” often hide different definitions.

Population

Do programs serve people with similar starting conditions, eligibility rules, or levels of need? A lower raw outcome can reflect a population facing greater barriers rather than poorer performance.

Period

Are outcomes observed over the same time? Exit, three-month, and twelve-month results answer different questions.

Source

Do the numbers come from sources with similar coverage and incentives?

Administrative records, participant surveys, staff judgment, and self-reported provider counts have different strengths and gaps.

If any dimension differs, write down whether the difference can be repaired, adjusted, grouped, or must remain visible.

A worked example

A foundation funds six youth programs and wants to compare “school success.” Two report attendance, one reports course completion, two report graduation, and one reports participant confidence. Follow-up periods range from program exit to two years. The programs also serve different age groups.

A single average would have no stable interpretation. Standardizing everyone on graduation would exclude younger participants and distort the work of short-term programs.

The foundation instead creates three groups: attendance and engagement, progression and completion, and longer-term attainment. Within each group, it aligns definitions where feasible and reports coverage. Across the full portfolio, it synthesizes common delivery questions and participant barriers rather than ranking outcome rates.

For allocation, it combines the evidence with need, strategic fit, cost, quality, and the consequences of reducing support. The portfolio view becomes more honest and more useful because it stops asking one number to carry every decision.

Alternatives to a league table

Group like with like. Create defensible clusters based on population, program model, outcome, and period.

Use common questions, not always common measures. Programs can answer how they understand access, retention, participant experience, or delivery constraints using evidence suited to their work.

Show ranges and coverage. When aggregation is meaningful, preserve variation and the number of people represented.

Use structured narrative synthesis. Summarize where evidence converges, differs, or remains absent. A disciplined narrative can be more informative than an unstable score.

Separate threshold checks from ranking. A minimum safeguarding or reporting requirement can apply across programs without implying that every outcome rate is comparable.

What you can say

- “These three programs use the same outcome definition, period, and source for broadly similar populations.”
- “The programs cannot be fairly ranked, but they reveal a common access problem.”
- “This portfolio result describes activity across all grants; outcome evidence is reported in three groups.”
- “Allocation considered need, evidence, cost, strategic fit, and transition consequences.”

What you cannot say yet

- “Program A performs better,” based on unadjusted outcomes for different populations.
- “Everyone measures the same outcome,” because the column carries the same label.
- “Standardization creates fairness,” when the standard fits some program models better than others.
- “The portfolio achieved a 64 percent success rate,” when unlike outcomes were added or averaged.

What to do next

Minimum useful step: Write the purpose of the comparison and complete the four-part test for the measures already proposed.

Stronger step: Group programs where alignment is defensible, document differences, and select a synthesis method suited to oversight, allocation, learning, or public account.

Before using a ranking: Examine population differences, missingness, source incentives, uncertainty, and the consequences of error. Give affected programs an opportunity to verify definitions and provide context without allowing the process to become advocacy by spreadsheet.

A 15-minute team exercise

Choose two programs currently shown in the same dashboard.

1. State what the comparison is expected to change.
2. Write each program's outcome definition word for word.
3. Add the population, period, and source.
4. Mark each dimension aligned, repairable, or meaningfully different.
5. Decide whether the pair supports ranking, grouping, shared learning, or no comparison.
6. Rewrite the dashboard label to match that decision.
7. Name one contextual fact that should travel with the number.

If the team disagrees about the intended use, stop there. The disagreement is more important than choosing another common metric.

A note on fairness

Raw rankings can punish programs for serving people with greater needs, operating in places with fewer services, or reporting more honestly. Statistical adjustment can help in some settings, but it does not remove value judgments about which differences should count.

Make those judgments visible. Ask who could lose resources because of the comparison and whether the evidence is strong enough for that consequence. Fairness requires attention to the decision, not only the formula.

When to bring in help

Outside support is useful when a portfolio comparison will affect allocation, measures require technical adjustment, programs dispute the definitions, or the institution needs a defensible synthesis across unlike evidence. Independent review may be warranted when the party designing the ranking also benefits from its conclusion.

The aim is not to make everything comparable. It is to identify which comparisons are honest enough for the use and where difference must remain visible.

Questions to carry forward

- What will the comparison be used to do?
- Do outcome meaning, population, period, and source align?
- Would grouping or narrative synthesis answer the real question better?
- Who bears the consequence if the comparison is wrong?