NSW Treasury

Outcome Evaluation Design

TPG 22-22 Policy and Guidelines: Evaluation – Technical Note

June 2023

Acknowledgement of Country

We acknowledge that Aboriginal and Torres Strait Islander peoples are the First Peoples and Traditional Custodians of Australia, and the oldest continuing culture in human history. We pay respect to Elders past and present and commit to respecting the lands we walk on, and the communities we walk with.

Artwork:
Regeneration by Josie Rose

Contents

Abstract

Technical Note: Outcome evaluation design	
Background	Outcome evaluation is a process that measures the achievement of intended outcomes of an initiative. It assesses the extent to which the initiative has led to its intended outcomes, whether in the short, medium or long term. The timing of the evaluation is determined by the expected realisation period for outcomes and the availability of information required to answer evaluation questions.
Scope	This document summarises the three main groups of outcome evaluation designs including their key features, uses, strengths and limitations for producing evidence. These include: A. Experimental design B. Quasi-experimental design C. Non-experimental design.
When to use this technical note	This document is recommended for use during evaluation planning. Evaluation design requires early planning to set up a comparison group, collect baseline data and establish data linkages. Adjustments can be made over time, if necessary, to ensure that the initiative is on track to achieve its outcomes.
Potential implications	Evaluation designs differ in their ability to manage bias in their results. Using multiple evaluation designs and methods (via triangulation) can mitigate the inefficiencies of any individual evaluation design and can build on the accuracy of evidence generated. A non-experimental outcome evaluation design often involves challenges in determining attribution as it does not include a control or comparison group that acts as an actual counterfactual. This technical note addresses methods to investigate attribution when there is no clear counterfactual.
Keywords	Attribution, triangulation, evidence, evaluation design
Associated Resources	Evaluation Workbook I. Foundations for evaluation & Workbook VIII. Complex initiatives, TPG23-08 CBA Guide, and Technical Notes on Sampling strategy & Evidence in evaluation.

Context

Outcome evaluations demonstrate the contribution of an initiative towards achieving intended outcomes and benefits. To do this effectively, outcome evaluations should be designed to demonstrate that the initiative caused the observed outcomes1. The evaluation design determines required data collection and analysis, and the strength of any conclusions.

Evaluation designs differ in their ability to manage bias. Experimental designs, followed by quasi-experimental designs, reduce potential for bias and are better at demonstrating causal relationships. Non-experimental designs alone do not provide strong evidence of causal relationships.

When designing an evaluation, choose designs that:

- will provide strong evidence
- are suitable to answer the evaluation questions
- are relevant to the initiative's context
- are feasible to implement in given contexts, provide value for money, and are within available time and resources.
- An outcome evaluation may include more than one design if one method is insufficient in generating evidence to address the evaluation questions. The use of two or more evaluation designs and methods to produce evidence, known as triangulation, can strengthen and check the accuracy of evidence.

Triangulation should be used where a single data source alone will not provide strong evidence of causal links. See *Technical note: Evidence in evaluation* for further information.
For outcome evaluation this could include:

- qualitative analysis to support experimental and quasi-experimental designs findings
- a combination of non-experimental methods to rule out alternative explanations for change when experimental or quasi-experimental designs are not feasible
- triangulation with findings from other reliable sources, such as previous evaluations and research projects.

See *Technical note: Evidence in evaluation* for further information.

Evidence hierarchies can help select the right evaluation method or weigh different pieces of evidence. This is particularly useful when using multiple methods. NSW Department of Education and NSW Department of Communities and Justice) provide examples of evidence hierarchies for evaluators.

The quality of an evaluation's evidence determines whether its findings may be generalised. An evaluation that achieves **internal validity** can conclude that the initiative caused the intended outcomes, ruling out other plausible explanations. An evaluation that achieves **external validity** can generalise its findings to different contexts, such as different locations or populations.

[1] Note the difference between causation (referred to here) and correlation. Correlation refers to a linear relationship between variables, while causation implies a direct relationship. Correlation does not necessarily imply causation, as there may other factors involved in the relationship.

Types of outcome evaluation

Experimental design is the strongest for establishing causal attribution

Experimental design, or randomised control trials (RCTs), randomly assign participants to either an intervention group or a control group to test for the causal attribution of an initiative to its outcomes. Participants allocated to the control group are not exposed to the initiative, and may receive an alternative intervention, a later intervention, or services as usual.

The randomised selection of participants for the different groups reduces bias that may affect the outcomes of an initiative. Because the allocation of the intervention is random, both identifiable and unknown factors that may also affect change are more likely to be evenly distributed between groups.

Following initiative implementation, the average outcome for the intervention group is compared to that for the control group. If there is a statistically significant difference, then it is reasonable to conclude that this is attributable to exposure to the initiative, excluding alternative explanations.

A high quality quasi-experimental design can demonstrate causal attribution

Quasi-experimental designs use non-randomised methods to assign participants to intervention and comparison groups. It is also known as non-randomised experimental design or natural experimental design.

A quasi-experimental design tests if the average outcome for the comparison group and for intervention groups are statistically different. Selection bias can, however, be a challenge. It is difficult to determine whether the observed changes are due to the intervention or other factors, without controlling for other factors. Quasi experimental designs also may not be generalizable due to differences in the characteristics between the intervention and the control groups.

Non-experimental design can provide context and narrative

Non-experimental designs can provide a context for change, as well as a narrative for quantitative data. Unlike experimental and quasi-experimental designs, non-experimental designs do not include a control or comparison group that acts as a counterfactual. Identifying attribution can be challenging, particularly many factors are likely to drive change.

Methods

Experimental design

Experimental design is feasible when:

- random allocation is concealed (i.e. measures are in place to ensure that the participants are allocated to their designated group before the intervention)
- an adequate sample size to detect differences between control and intervention groups is available.

Experimental designs are most feasible when established before the initiative is implemented, so that relevant baseline data are collected, and processes established.

Experimental designs are often appropriate for evaluating pilot initiatives or phased rollouts before wider implementation. When an initiative is in its early stages, the number of participants may be limited, and random sampling can minimise bias and enhance the generalizability.

Experimental designs produce stronger results when participants do not know whether they are in the treatment or control groups. This is called blinding. Blinding prevents participants from changing their behaviours or leaving a placebo trial. Ethical or legal issues should be considered prior to implementation. For example, where randomisation means initially withholding a potentially

beneficial new intervention or delivering an untested intervention. Randomisation and blinding may not be practical or ethical in such situations.

Experimental design may be pre-post with observations before and after the intervention, or post-test with observations after the intervention. Use:

- **Pre-post design** to measure the change between before and after the intervention – this is the strongest design for assessing attribution.
- **Post-test design** when a pre-post evaluation is not possible (for example when limited resources do not allow for multiple measurements over time, or the intervention is already underway) or suitable (for example, where a pre-test would influence participants behaviour and therefore the results).

For more information on pre-post and post-test designs, see Table 2.

Establishing a control group may be challenges for complex initiatives (see *Workbook VIII: Complex initiatives*). Experimental designs can still be used at sub-initiative level, provided that the sub-initiative is independently responsible for outcomes (i.e., can be evaluated as a standalone activity) and a control group can be selected.

Table 1: Experimental design examples – methods, characteristics, strengths, and limitations

EXPERIMENTAL DESIGN (Randomised Control Trials)		
Methods	Characteristics	Strengths and limitations
Randomisation methods	Suitable for well-defined (discrete) initiatives and controllable settings. Can be undertaken as pre-post or post-test control group (see Table 2 below). Variations to RCTs include: • **Factorial RCT** — independently randomise participants to multiple interventions, when evaluating more than one intervention (for example participants are randomly allocated to receive neither intervention, one or more interventions) • **Cluster RCT** — randomise groups of participants (for example schools or families) rather than individuals • **Stepped-wedge RCT** — apply an intervention sequentially and at random to groups of participants, until all participants receive the intervention.	**Strengths** • Minimise potential bias due to random assignment. • Provide strong evidence of attribution. **Limitations** • RCTS can be difficult to implement because of strict conditions (for example potential challenges in keeping group membership hidden; awareness of group membership can impact participant behaviour and results). • RCTs may present ethical and legal concerns when initially withholding new services from participants who could benefit from receiving them. Phased implementation of an intervention or the control group receiving business as usual service may address this concern. • Generalising findings from RCTs to a wider population may be limited due to the controlled setting of some studies (e.g., in laboratories). • Larger sample sizes may be required for RCTs to allow for random assignment.

Table 2: Pre-post and Post-test designs – characteristics, strengths, and limitations

Design	Characteristics	Strengths and limitations
Pre-post design	Observes participants in a control and intervention group before and after the intervention.	**Strengths** • Pre-test data can confirm that both groups have similar characteristics. • Before and after data supports understanding of how both groups changed over time. **Limitations** • Contains risk of attrition (i.e. participants dropping out) • Pre-testing may influence results (for example, children given a pre-test may study more for the post test, regardless of the group they are in).

Design	Characteristics	Strengths and limitations
Post-test design	Observes participants in a control and intervention group after the intervention.	**Strengths** • minimises the risk of participants changing behaviour • decreases the risk of participant drop out from the data collection, due to lack of pre-test. • reduces complications associated with complications relative to other methods **Limitations** • lacks comparable before and after information.

Quasi-experimental design

Quasi-experimental design should be used when randomisation is not feasible or ethical, but some form of comparison group, such as a naturally arising comparison group, is possible. When pre-existing or pre-selected comparison groups are available, a quasi-experimental design is likely to be quicker and less expensive than a RCT. When matching techniques are used to select the comparison group, more complex statistical techniques may be required.

Quasi-experimental designs generally require collection of baseline data and are most feasible when established before the initiative is implemented.

Methods to establish comparison groups include:

- **Matching**: selecting a comparison group that is a good match to the intervention group. Examples include:
  - Judgemental matching — relies on the evaluator's judgement on which characteristics to match.
  - Propensity score matching — based on participant readiness to participate.
- **Criterion based allocation:** establishing a criterion to allocate participants. Examples include:
  - Sequential allocation — assigns participants to groups using a sequence (for example every third person on the list).
  - Regression discontinuity — assigns participants based on their position below or above a pre-defined cut-off point.
- **Natural comparison groups:** allowing a comparison group to arise naturally based on the way the initiative is designed or implemented. Examples include:
  - Phased introduction — uses comparison groups that are naturally formed from earlier phases of initiative implementation.
  - Accidental delays — uses comparison groups that are formed when the intervention proceeds more rapidly for some groups than others.

Table 3 provides further information about quasi-experimental evaluation methods. Provides examples of different quasi-experimental evaluation methods, for measuring and analysing outcomes.

Table 3: Quasi-experimental design examples – methods, characteristics, strengths, and limitations

QUASI-EXPERIMENTAL DESIGN		
Methods	Characteristics	Strengths and limitations
Pre and post studies with matched comparison group	Observes participants in the comparison and intervention groups before and after the intervention. Matching technique construct a comparison group using characteristics of the intervention group.	**Strengths** • may be relatively simple to undertake. **Limitations** • unobservable differences between groups may impact results.

QUASI-EXPERIMENTAL DESIGN		
Multiple baselines / phased introduction	Exposes one group to the intervention, while other groups remain at baseline conditions. All groups receive the intervention at different times. Measures key indicators multiple times, before and after the intervention. Can use phased introduction to establish a comparison group.	**Strengths** • takes advantage of projects that are rolled out gradually • reduces selection bias by exposing the different groups to the intervention at random. • provides strong evidence, if similar change is registered after the intervention in all groups • supports analysis between and within groups over time **Limitations** • can be complex to design and analyse • can be challenging to determine the optimal number of key indicators for the intervention and control groups
Interrupted time series (ITS)	Measures outcomes at different points in time before and after the intervention, using a single group. Calculates trends before and after the intervention. Measures the effect of the initiative by changes in the slope or level of the trend lines.	**Strengths** • may be cheaper and easier to implement relative to other methods, provided data is available • may be useful for initiatives introduced to the whole population • may be useful when there is no viable comparison group • may improve understanding of change by presenting multiple measures • may allow for control of external factors affecting outcomes by adding a comparison group differentiated by exposure to external factors. **Limitations** • may not be possible to attribute changes to the initiative without a comparison group. • may require many indicators • may not account for changes due to external factors.
Difference in differences (DID)	Compares outcomes between intervention and comparison groups after the intervention, controlling for baseline differences between the groups. Can use propensity score matching or regression discontinuity to establish a comparison group.	**Strengths** • may address (in calculations) pre-existing differences between groups before the intervention. • may be simple and easy to explain relative to other methods. **Limitations** • may not consider differences between intervention and comparison groups that arise over time. • substantial data requirements • quality of results dependent on the quality of data.

Non-experimental design

Non-experimental design should be used when experimental or quasi-experimental designs are not ethical or feasible. For example, when baseline data or control or comparison groups are not available. This may include initiatives where the whole population of interest is exposed. Non-experimental methods can test the theory-of-change or logic model and to explore alternative explanations for change (for example by using expert opinion).

Non-experimental designs can be useful if the initiative is new, not well understood or not being implemented as intended. They may also help address issues of contribution in complex initiatives and identify areas for further research through experimental or quasi-experimental designs.

Table 4 provides further information on non-experimental methods.

Table 4: Non-experimental design examples – methods, characteristics, strengths, and limitations

NON-EXPERIMENTAL DESIGN		
Methods	Characteristics	Strengths and limitations
Pre and post studies without a control or a comparison group	Measures change for a single group before and after being exposed to the intervention	**Strengths** • may be relatively easy to implement **Limitations** • does not control for other factors affecting change and, therefore, may not demonstrate causal link between initiative activities and outcomes.
Contribution analysis	Offers an approach intended to reduce uncertainty about the contribution an intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not), and the roles played by the intervention and other internal and external factors.	**Strengths** • useful where it may not be possible to establish an experimental design testing cause and effect • useful where there is limited scope or opportunity to affect roll out of a programme (to allow for experimental methods) • useful for confirming or revising a theory of change. **Limitations** • The quality of the analysis depends on the quality of the theory of change to explain the logic behind the attribution of the initiative to outcomes. • Contribution analysis does not provide definitive proof that the intervention has had a causal effect on outcomes but rather it gives an evidenced logical line of reasoning. • Works on average effects, therefore, should not be used if there is a large degree of variance between how the program has been implemented or an expectation of different outcomes for different groups.
Process tracing	Examines a single case of change and tests whether a hypothesised causal mechanism, such as that proposed by a theory of change, explains the outcome.	**Strengths** • This method assists in understanding and testing causal hypothesis in 'real world' situations that can be used in ex-post evaluation of a single case • It can account for more complexity and nuances (such as the impact of omitted variables) better than other qualitative methods • It may spot emergent influences more easily than other methods because it is based on chronological assessment of events. **Limitations** • If not applied effectively, it may lead to inferential errors. To avoid this, the sequence of events leading up to findings must be established and alternative explanations must be carefully considered. Equifinality (the support of one causal relationship between variables may not preclude others) should also be considered.
Retrospective and prospective cohort studies	**Retrospective cohort study** uses historical data to compare the extent of change between and among groups of participants. **Prospective cohort study** follows participants for a period of time and assesses the level of change between and among groups. Selects participants based on their exposure to the intervention.	**Strengths** • may be used to identify trends and indicate potential relationships • may be used to study multiple outcomes. **Limitations** • useful only when relevant data is available • requires a large sample size • may not demonstrate causal links.

NON-EXPERIMENTAL DESIGN		
Cross sectional studies / Longitudinal studies	**Cross sectional study** collects information at a single point in time from a sample. A **longitudinal study** collects data from the same sample at different points over a period. Ongoing data collection supports comparison of the extent of change between participants who have had different levels of exposure to the intervention.	**Strengths** • may be relatively easy and inexpensive • may be useful for generating hypotheses for further investigation • may be used to strengthen the results of a study by repeating it with a different group of participants • may provide profile of a population, which can then be used to identify a comparison group. **Limitations** • offers limited insight into how and why an initiative may or may not have worked • does not control for other factors affecting change and, therefore, may not demonstrate a causal link between initiative's activities and outcomes.
Causal link monitoring	Involves using a logic model to **describe the process** that will lead to one point in each component of the logic model leading to the next. Describing this process involves identifying assumptions underlying the logic model. **Steps** in causal link monitoring: 1. Build a logic model. 2. Identify assumptions about causal links. 3. Enhance the logic model with diverse perspectives and contextual factors. 4. Prioritize areas of observation. 5. Collect monitoring data. 6. Interpret and use monitoring data for adaptive management. 7. Revise the logic model.	**Strengths** • encourages the use of a robust logic model with consideration of assumptions and causal links • provides a structure in using the logic model to identify causal links and scrutinise assumptions • emphasises the value of collecting and using monitoring data. **Limitations** • does not provide a clear approach where there is a lack of available monitoring data.

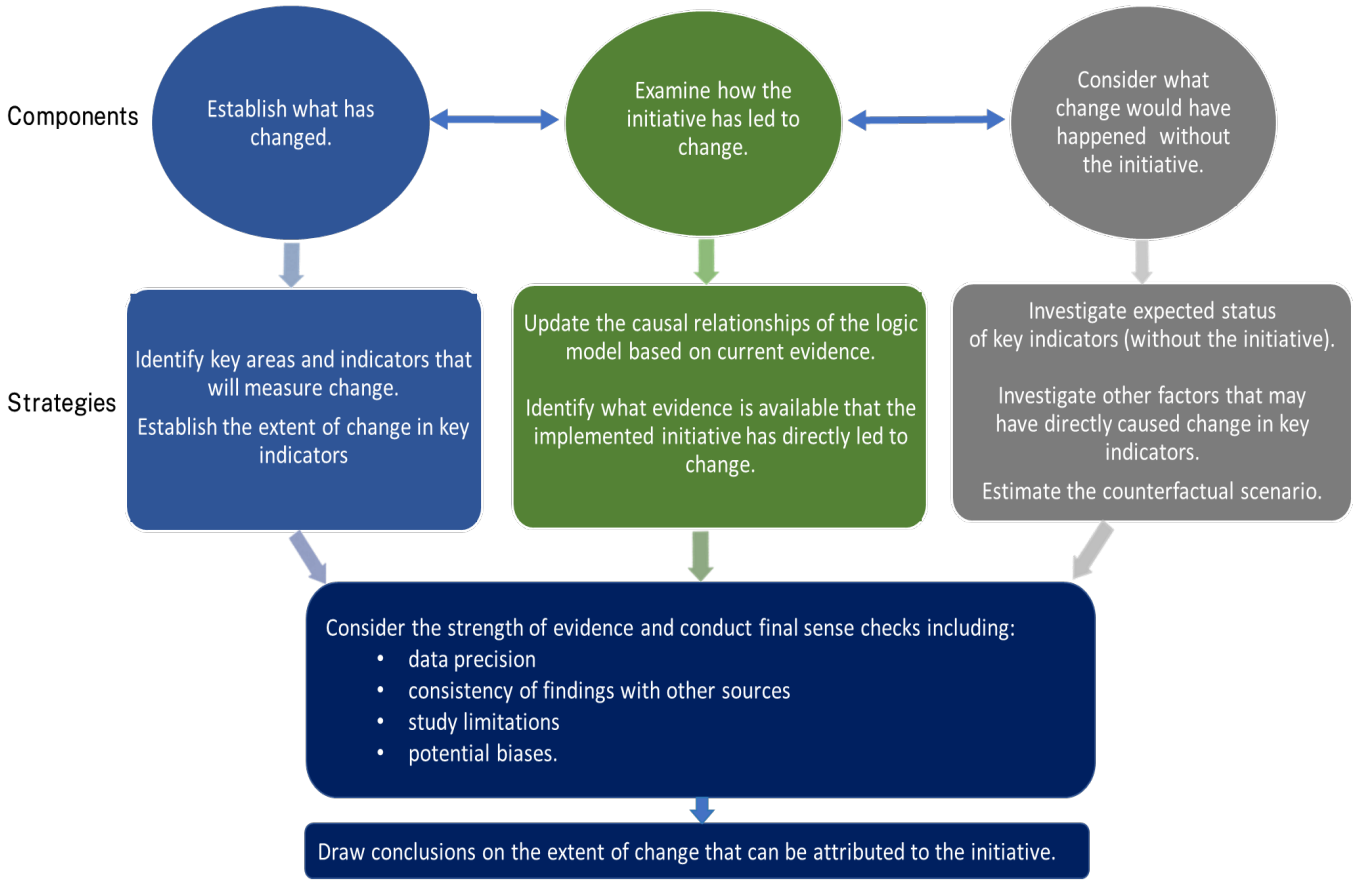
Investigating attribution with non-experimental design

Attribution in a non-experimental design can be examined by:

- assessing the extent of change using relevant indicators
- analysing the causal relationship between the initiative and observed changes
- consider what changes would have occurred without the initiative
- evaluate the strength of evidence sources and do a final sense check.

Figure 1 provides an overview of these interrelated strategies.

Figure 1: Attribution in non-experimental design



Establish what has changed

To establish what has changed, identify key areas and indicators impacted by the initiative and assess the extent of change using relevant indicators.

Table 5: Strategies to establish what has changed

Strategies to establish what has changed	Explanation
Identify key areas and indicators that will identify change. This may be informed by the logic model.	Identify what information can reliably indicate change and can be monitored and collected.
	Identify where baseline data is available or where it will be necessary to estimate the pre-initiative status of key indicators.
Establish the extent of change in key indicators following implementation of the initiative	Where baseline (pre-implementation) data is available, compare pre- and post-implementation data to calculate the extent of change in key indicators
	Where baseline data is not available: use interviews, surveys or focus groups to investigate: • participant's recollections of events before the initiative, and • their perceptions of the extent of change since implementation.

Examine how the initiative has led to change

This component involves identifying key areas and indicators to consider change and understanding the extent of change using the indicators. Logic models are a useful tool when examining how an initiative has led to change. Further information on logic models can be found in *Workbook I: Foundations of Evaluation.*

Table 6: Strategies to examine how the initiative has led to change

Strategies to examine how the initiative has led to change	Explanation
Identify what evidence is available that the implemented initiative has directly led to change	Engage with delivery agents and initiative recipients (using tools such as interviews, surveys or focus groups) to ask what level of change they see as directly attributable to the initiative.
	This should be coordinated with questions related to what change they would have expected without the initiative, and what other factors may have caused changes.
Consider the possibility that the initiative has reproduced or displaced outputs or outcomes that would have been delivered under the base case	Identify what information is available that this type of initiative leads to the assumed change (for example, from evaluations of similar initiatives or from primary research about the initiative).
	Test and analyse the causal relationships in the logic model, based on theory and any new evidence that has become available. Further information on methods to examine these causal relationships is provided in Table 4.
	Involve key stakeholders such as the delivery team, relevant experts (e.g. academic experts) or customers.

Consider what change would have happened without the initiative

What would have happened without the initiative is known as a logical counterfactual. It estimates the change in key indicators in the absence of the intervention. This may require developing a statistical model based on established trends at the baseline and controlling for exogenous variables.

Factors that may influence changes in relevant indicators, such as changes in market conditions or external events or policies, should be identified early (i.e. during business case development).

Monitoring these factors can ensure unexpected changes are captured and lessons learnt can be incorporated into future business cases for similar initiatives.

Quantitative and qualitative evidence can be used to construct a logical counterfactual. Qualitative evidence can be gathered from sources such as interviews and focus groups.

To develop a logical counterfactual:

1. **Investigate established trends of key indicators without the initiative**:

- before the implementation of the initiative
- in similar contexts where the initiative has not been introduced (e.g. similar groups or regions that have not been exposed to the initiative)

Stakeholders and expert opinion can be useful information sources. The case for change and base case analysis from any business case is another valuable source of information.

2. **Investigate other factors that may have caused change in key indicators, e.g.:**

- the impacts of other initiatives that complement, substitute or counteract the initiative (e.g. information campaigns and incentive programs).
- innovation (e.g. technological advancements, improvements to operating processes).
- changes in:
  - economic conditions (e.g. exchange rates, interest rates, stimulus to demand)
  - social conditions (e.g. changing values)
  - environmental conditions (e.g. weather, pollution, natural disasters).

3. **Consider if the initiative has reproduced or displaced outputs or outcomes that would have been delivered under the base case**

As per the NSW Government Guide to Cost-Benefit Analysis TPG23-08 (A3.7):

1. **Additionality:** For example, a business may have already been intending to take on another employee but then use funding from an initiative to do so, or it may replace an established employee with another to benefit from an initiative. This may result in no net increase in employment or output.
2. **Displacement:** The outputs or outcomes delivered by the initiative are fully or partially offset by reductions elsewhere in New South Wales. That is, the outputs or outcomes would have been delivered in the base case but in a different location. For example, businesses or employees may relocate in response to an initiative which causes shifts in employment. Hence, there is no increase in employment within the New South Wales referent group.
3. **Leakages:** Where outcomes or benefits leak out to other jurisdictions. For example, a business operating in New South Wales, but owned predominantly by overseas shareholders, resulting in some profits leaking overseas.

Consider the strength of evidence and conduct final sense checks

Consider the strength of the different sources of evidence, and coordinate findings, to draw conclusions regarding the extent of change that can be attributed to the initiative.

Combining data and findings can help to build and check the accuracy of evidence. Be aware that responses may be influenced by biases such as memory (ability to recall), social desirability (report socially desirable responses to a researcher), or framing (response influenced by the way the question is framed).

Clearly state what data has been used and the strength of each data source. Final sense can include comparing the findings against previous evaluations of the initiative or similar initiatives. Final sense checks should also be conducted to ensure that the conclusions are reasonable. This can include comparing the findings against previous evaluation of the initiative, or evaluations of similar initiatives.

Testing conclusions through expert opinion, peer review or stakeholder review can also help to verify the accuracy of the results.

Example: Hypothetical jobs initiative

This example is a hypothetical initiative providing job grants to businesses in disadvantaged regions.

Table 7: Establish what has changed

Strategy	Explanation
Identify key areas and indicators that will identify change.	Indicators may include: Jobs created in the local area, percentage of target cohorts securing future work, number of employees that targeted firms hire.
	Baseline data may include the unemployment rate and firm size before the initiative.
Where baseline data are not available, consider survey questions	Example 1: How many people do you employ (i) now, and (ii) prior to the initiative?
	Example 2: How many hours do you work (i) now, and (ii) prior to the initiative?

Table 8: Examine how the initiative has led to change

Strategy	Explanation
Identify what evidence is available that the implemented initiative has directly led to change	Survey questions to examine the causal links could include: • To what extent has the initiative supported you to employ more people from the local area compared with other factors? [like scale 1-5]. • To what extent has the initiative increased the number of people you employ who are from the local area? [like scale 1-5].
Where baseline data are not available, consider survey questions	Example 1: How many people do you employ (i) now, and (ii) before the initiative?
	Example 2: How many hours do you work (i) now, and (ii) prior to the initiative [date]?

Table 9: Consider the counterfactual

Strategy	Explanation
Investigate expected status of key indicators (without the initiative)	Identify what change may have happened to key indicators without the initiative using established trends, such as the local unemployment rate or size of targeted businesses.
	Identify similarly disadvantaged groups or regions that have not been exposed to the initiative to consider the possible trajectories of change without the initiative.
Investigate other factors that may have caused change in key indicators	Example: Other jobs grants programs targeting similar businesses. Ask businesses if they have also relied on other programs. If so, ask them which ones.
	Example: Assess whether there have been changes in economic conditions (e.g. economic activity) and social changes (e.g. changing demographics).
Consider survey questions to examine any displacement	To businesses: How many staff hired since the grants were previously unemployed?
	To employees: Were you previously working? Was your role in or outside the area?

Table 10: Consider the strength of the evidence and conduct a sense check

Strategy	Explanation
Consider the strength of difference sources	Compare survey findings with findings from similar initiatives
	Complement survey findings with other sources, e.g. administrative and economic data
	Consider data limitations
Conduct final sense checks on attribution	Analyse regions not exposed to the initiative if not already done so.
	Test the proposed counterfactual with key stakeholders.