Australian Government
Australian Institute of Family Studies

Discovering what
works for families

Evaluation design

FAMILIES AND CHILDREN
EXPERT PANEL PROJECT

Shae Johnson

June 2021

The Australian Institute of Family Studies is committed to the creation and dissemination of research-based information on family functioning and wellbeing. Views expressed in its publications are those of individual authors and may not reflect those of the Australian Institute of Family Studies or the Australian Government.

The Families and Children Expert Panel Project is a team of researchers and evaluators based at the Australian Institute of Family Studies and funded by the Australian Government Department of Social Services. The Expert Panel Project helps service providers to plan, implement and evaluate programs for families and children using evidence-based approaches.

Australian Institute of Family Studies
Level 4, 40 City Road, Southbank VIC 3006 Australia
Phone: (03) 9214 7888 Internet: aifs.gov.au

Cover image: © gettyimages/FreshSplash

Edited by Katharine Day
Typeset by Lisa Carroll

About this resource

Once you have decided on evaluation questions – what you want to know – then you need to decide how you are going to answer those questions. An 'evaluation design' is the overall structure or plan of an evaluation – the approach taken to answering the main evaluation questions. Evaluation design is not the same as the 'research methods' but it does help to clarify which research methods are best suited to gathering the information (data) needed to answer the evaluation questions.

This resource gives a quick overview of some of the main evaluation designs used for outcomes evaluations or impact evaluations. These are evaluations that aim to answer questions about whether a program, service or treatment (often called the 'intervention') is working as intended, or if it is having a positive or negative effect on its intended audience. We also briefly discuss some other types of evaluation design that are sometimes used in outcomes evaluations but are also commonly used to evaluate how programs or services are being delivered.

This resource is intended for use by program managers or practitioners who want a basic understanding of the different types of evaluation design.

Deciding on an evaluation design

Different evaluation designs are suitable for answering different evaluation questions, so the design of an evaluation usually depends on its purpose and the key evaluation questions it is meant to answer (aifs.gov.au/cfca/expert-panel-project/identifying-evaluation-questions). This guide focuses on evaluations that measure a program or intervention's effectiveness or results. An evaluation design with a focus on effectiveness may include questions such as, '*To what extent did the program achieve its expected outcomes?*' or '*What changes occurred as a result of this program?*'. However, evaluations can also have a different purpose, such as determining if a program or service was implemented as intended, if it was appropriate for its intended client group or what the cost versus benefit was. These different types of evaluation can require different kinds of evaluation design.

There are also other factors to consider when deciding on an evaluation design, and these are listed below. Working though these factors will help to inform the design and methods that will be most suitable for your evaluation. The last two factors in this list will also help establish the scope of the evaluation. Additional support to work through these factors is provided under Further reading.

Important points to consider when deciding on an evaluation design are:

- the questions you want to answer
- the audience for the evaluation
- the maturity of your program (i.e. is it ready to evaluate outcomes or has it only just started?)
- the type of program or intervention you are seeking to evaluate
- your client or target group (e.g. who the program is for, how many people are in the program or receive a service and what their characteristics are)
- what data are already available
- your resources (e.g. funding, staff, skills) and time frame
- whether you will conduct an evaluation internally or contract an external evaluator.

The following designs are most appropriate for conducting an outcomes evaluation. However, an outcomes evaluation is most useful if it is accompanied by a detailed understanding of how the program was delivered and to whom. For example, did the program reach the intended participants? Were all components of the program delivered? These types of questions are explored in a different type of evaluation called a 'process evaluation' (aifs.gov.au/cfca/expert-panel-project/what-evaluation) and it can be useful to combine process evaluations with those that look at outcomes. Knowing a program was delivered as planned will then allow you to link the program activities to the outcomes.

Evaluation designs

Researchers and evaluators sometimes refer to a 'hierarchy of evidence' for assessing the effectiveness of a program or intervention. The evaluation designs that are thought to produce the most powerful evidence that a program or intervention works are usually situated at or near the top of this hierarchy.

The hierarchies usually have randomised controlled trials (RCTs) at or near the top. These are usually followed by 'quasi-experimental' designs using comparison groups. These types of evaluation designs aim to measure changes for participants before and after the program or intervention and may compare these changes to other groups of participants that did not attend the program or intervention. There are also a range of other non-experimental designs such as pre- and post-test studies or case studies; these may not be able to produce such strong evidence for program effectiveness but can be more appropriate depending on the situation.

Experimental

- Randomised control trial

Quasi-experimental

- Case comparison groups

Non-experimental

- Pre- and post-test studies
- Case studies

If you are planning an evaluation, you can use these hierarchies to guide your decisions about which evaluation design to use but the choice of design should also be guided by key questions outlined in the section above. RCTs may be considered the most powerful evidence but they are not always possible or appropriate. So, what do some of the main designs look like?

Experimental designs

Randomised controlled trials (RCT) are the main experimental evaluation design. RCTs are a method of systematically testing for differences between two or more groups of participants. This usually means one group receives the intervention, treatment or service that is being evaluated or tested (the 'intervention group') and the other does not (the 'control group'). Differences in results between the groups can indicate whether an intervention is effective or not.

Besides comparing the results between the groups, the main distinctive feature of an RCT is the random allocation of participants to the control and intervention groups. Randomisation provides each participant with an equal chance of being allocated to receive or not receive the intervention.[1] This is important because it means there is a greater chance that the people in the intervention and control groups will have a similar mix of attributes such as gender, health, attitudes, past history or life circumstances. Without randomisation there is more chance of systematic bias; that is, where one group is different to the other and this difference can affect the results. An example of systematic bias would be if the people in the treatment group for an anger management intervention already had lower-conflict relationships than the people in the control group. If this were so, it would not be possible to tell if any positive results were due to the intervention or to the pre-existing differences between the groups.

In RCTs, data are collected from participants before and after (and sometimes during) the program. If there is no bias in the way individuals are allocated to the groups, you can probably conclude that any differences between the groups after completing the program are due to the intervention rather than to pre-existing differences among participants. Since RCTs are typically conducted under conditions that provide a high degree of control over factors that might provide alternative explanations for findings, RCTs can provide a relatively high degree of certainty that the outcomes for participants are a direct result of the program.

[1] Participants allocated to the non-intervention group may receive an alternative intervention or receive the intervention at a later time.

Although RCTs are good at answering questions about intervention effectiveness (i.e. 'does it work?') they are less useful for answering questions about how or why an intervention works. From a child and family services perspective, RCTs cannot always accommodate the complex and challenging nature of service delivery (Tomison, 2000). In order to link participant outcomes to a program, RCTs need to be conducted under tightly controlled conditions. This can be difficult to do in real life situations and the evidence that RCTs produce is sometimes difficult to apply to everyday practice.

There are some RCT designs, such as cluster RCTs, that can be more useful for generating practice-based evidence than traditional RCTs (Ammerman, Smith, & Calancie, 2014). In cluster RCTs, groups – or clusters – of individuals such as those within schools, medical practices or entire communities are randomised to treatment or control conditions. For example, six schools may be selected to take part in a RCT and three are allocated to be treatment groups and three are allocated to be control groups.

There are also other experimental study designs that offer alternatives to traditional RCTs, such as time series analyses (Bernal, Cummins, & Gasparrini, 2017) and natural experiments (Dunning, 2012). However, these experimental designs, like most RCTs, require sophisticated statistical and methodological expertise.

As RCTs are not always practicable or appropriate, evaluators and researchers often employ the next best thing – comparison groups as part of quasi-experimental designs.

Quasi-experimental designs

A quasi-experimental design differs from an RCT in that it does not randomly assign participants to an intervention or control group. Quasi-experimental designs identify a comparison group that is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics. There are statistical techniques for creating a valid comparison group; for example, regression discontinuity design and propensity score matching, which reduces the risk of bias (White & Sabarwal, 2014).

Comparison groups are often used when the random allocation of program participants to control and intervention groups is not possible for practical or ethical reasons. Comparison groups can include waiting lists for an intervention and participants attending other programs where the participants are not able to be randomly allocated into groups. Participants on a waiting list are a good source of comparison data, because (a) they are available to you, and (b) you can collect the same data from them as you do from those participating in the program. The two groups are likely to be reasonably well-matched in terms of demographic characteristics as long as participants in the program group have not been given prioritised entry over the waiting list group.

Comparison groups may also be found in population data that have already been collected; for example, from health datasets. In this instance, it is important that they can be statistically matched to your control group to take into account any differences in the two groups. The outcome measures used would also need to be comparable.

Evidence of greater benefits to those who participated in an intervention compared to a comparison group can suggest the program is effective, but it is more difficult to say with certainty that the program caused the change. Because there has not been a random assignment of participants, it is not always possible to say with certainty that any differences or benefits observed in the evaluation are the result of the intervention rather than pre-program differences between the groups of participants. For example, the clients in one comparison group might experience less severe problems, be from a particular cultural group, be older or have a different family type from those who participate in your program. Therefore, they might have better or worse outcomes than the other group that are not explained by the intervention. Nonetheless, if consistent results are found in repeated studies of a given type of program using a variety of quasi-experimental (and other, non-experimental) methods, then it is possible to have greater confidence in the effectiveness of the program.

Non-experimental designs

Most other evaluation designs fall under the broad heading of 'non-experimental' designs. When the use of control or comparison groups is not feasible, non-experimental designs can be appropriate.

Some common non-experimental designs (and approaches) are:

- pre- and post-test studies
- case studies
- most significant change (MSC)
- developmental
- realist
- empowerment

Some of these approaches, such as pre- and post-test studies, usually focus on an intervention's effectiveness or outcomes. However, others may more often be used for other forms of evaluation, such as understanding how a program has been implemented or whether it is appropriate for its intended audience. We list a few here that are sometimes used for measuring outcomes. More detail on these and other designs can be found in the Further reading section.

Pre- and post-test studies

Pre- and post-test studies examine the effect of a program without the use of either a control or comparison group. In this evaluation design, data are ideally collected (e.g. via survey or outcomes measure) from participants immediately before the program starts and again at its completion. Any change is then measured. If a program is ongoing, data might be collected from a client when they start the program. When the client leaves the program is the time for the post-test collection.

Outcomes can be measured at additional timepoints during or after the program, as well as pre and post. For example, if a client is expected to attend a program for an extended time, taking measurements mid-program can provide an opportunity to measure if it is having the expected outcomes for that client. If positive changes are found, this can also be an opportunity to provide feedback directly to the client. Outcomes measured at follow-up timepoints, such as three or six months after the program, can provide additional evidence about the long-term effectiveness of the intervention.

If the program works, the program logic would lead you to expect any changes recorded will be in the direction that supports the program goals. For example, participants completing a program may show increased self-esteem or a reduction in behaviour problems.

Pre- and post-test designs are often relatively easy to run and can require less specialised expertise than experimental or quasi-experimental designs. However, there are some important limitations to pre- and post-test designs. In these studies, even if there are differences between the pre- and post-test measures, it is difficult to say for certain whether the effects are due to participation in the program. This is because we cannot know if similar changes might have occurred anyway, even if the program had not been run. All that can be said is that some aspect of this group's behaviour (or attitudes, knowledge, skills, etc.) changed in the period between the start of the program and its conclusion.

Thinking about other explanations that may impact client outcomes is useful when looking at the findings of any evaluation. For example, improvements in child development may happen naturally as children age over the period of a program. In complex settings or when there may be other possible causes for the observed outcomes, further investigation can be useful.

Case studies

Case studies are another common evaluation design. These are often used to get an in-depth understanding of a single activity or instance within a program setting. This is useful when an evaluation aims to capture information on more explanatory 'how', 'what' and 'why' questions (Crowe et al., 2011). Case studies can be used to show personal experiences or unique program processes with both qualitative and quantitative data. For example, a case study evaluation for a parenting program may evaluate a small number of clients who provide detailed stories of their experiences. In this way, case studies do allow for a richness of information, but they are not able to provide a generalisation about the program as a whole. A case study is often combined with other evaluation designs.

Most significant change

When there is a focus on identifying what the outcomes of an intervention are (i.e. what changes result from an intervention) the **most significant change** design may be suitable. This story-based technique involves a form of continuous inquiry whereby designated groups of stakeholders search for significant program outcomes and then deliberate on the value of these outcomes in a systematic and transparent manner (Dart & Davies, 2003).

Developmental evaluation

Developmental evaluation (aifs.gov.au/cfca/publications/developmental-evaluation) is a structured way to monitor, assess and provide feedback on the development of a program while it is being designed or modified (Child Family Community Australia [CFCA], 2018). The focus here is not on fully developed interventions but on programs or services where inputs, activities and outputs are not yet entirely decided on or are changing. Developmental evaluations attempt to address the challenges of evaluating developing or changeable programs and services by adopting a more responsive and adaptive approach. This is done by asking evaluative questions, applying evaluation logic, and gathering and reporting on evaluative data to support project, program, product and/or organisational development with timely feedback (Patton, 2012). Although this approach can measure outcomes it is less useful for undertaking a rigorous assessment of whether an intervention 'works'.

Realist evaluation

A **realist evaluation** is an approach to evaluation that uses qualitative methods (such as interviews or focus groups) to understand in detail the underlying mechanisms of a program or intervention. This may be the case when an experimental or quasi-experimental design cannot provide the level of understanding needed of the mechanisms of a program. Realist evaluation is less often used to understand whether a program is effective (i.e. did it achieve the desired outcomes) and more often used for evaluating new initiatives or programs that seem to work but where 'how and for whom' they work is not yet understood. This can include programs that have previously demonstrated inconsistent outcomes as well those that will be scaled up or implemented in new contexts (Westhorp, 2014).

Empowerment evaluation

Empowerment evaluation (aifs.gov.au/cfca/publications/empowerment-evaluation) is more a set of principles that guide the evaluation at every stage than an evaluation design (CFCA, 2015). This approach is drawn from the participatory or collaborative field of evaluation and seeks to involve all stakeholders (i.e. evaluators, management, practitioners, participants and the community) in the evaluation process. This approach can potentially be combined with other evaluation designs.

In conclusion

This resource has provided a basic overview of the different types of evaluation design used for outcomes evaluations. There are a range of evaluation designs that allow for different types of evaluation questions to be answered. However, evaluations that focus on effectiveness – how well a program works – differ in their strength of evidence. These include experimental, quasi-experimental and pre- and post-test evaluations. Evaluation designs that are non-experimental may focus more on the 'why' and 'how' of a program. Identifying the right evaluation design for the right situation is the first step to a successful evaluation.

Further reading

Identifying an evaluation design

CDC Centres for Disease Control and Prevention, program Evaluation Framework Checklist
www.cdc.gov/eval/steps/step3/index.htm

Evaluation designs and approaches

Better Evaluation
www.betterevaluation.org/en/approaches

WK Kellogg Foundation (2017). The Step-by-Step Guide to Evaluation
www.wkkf.org/resource-directory/resources/2017/11/the-step-by-step-guide-to-evaluation--how-to-become-savvy-evaluation-consumers

Haynes, L., Goldacre, B., & Torgerson, D. (2012). Test, learn, adapt: developing public policy with randomised controlled trials. Cabinet Office-Behavioural Insights Team.
www.researchgate.net/publication/256031307_Test_Learn_Adapt_Developing_Public_Policy_with_Randomised_Controlled_Trials/link/5b582bb9a6fdccf0b2f35012/download

The Indigenous Evaluation Strategy

Productivity Commission. (2020). A Guide to Evaluation under the Indigenous Evaluation Strategy.
www.pc.gov.au/inquiries/completed/indigenous-evaluation/strategy/indigenous-evaluation-guide.pdf

Developmental, realist and participatory evaluations are particularly suited to allowing Aboriginal and Torres Strait Islander knowledges, perspectives and world views to be incorporated into the design and delivery of evaluations (Productivity Commission, 2020). Culturally valid methods, such as yarning (storytelling), ganma (knowledge sharing) and dadirri (listening) can also be used to engage Aboriginal and Torres Strait Islander people throughout the evaluation process.

Evaluation Methods

Once you know what you want to collect for your evaluation the next steps are to decide how you will collect the data. Further information on research or data collection methods and more detail on conducting an evaluation can be found here: aifs.gov.au/cfca/publications/planning-evaluation-ii-getting-detail

References

Ammerman, A., Smith, T. W., & Calancie, L. (2014). Practice-based evidence in public health: Improving reach, relevance, and results. *Annual Review of Public Health*, *35*, 47–63.

Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, *46*(1), 348–355.

Child Family Community Australia (CFCA). (2015). *Empowerment evaluation* (CFCA Practitioner Resource). Melbourne: Child Family Community Australia, Australian Institute of Family Studies.

Child Family Community Australia (CFCA). (2018). *Developmental evaluation* (CFCA Resource Sheet). Melbourne: Child Family Community Australia, Australian Institute of Family Studies.

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC Medical Research Methodology*, *11*(1), 1–9. doi.org/10.1186/1471-2288-11-100

Dart, J., & Davies, R. (2003). A dialogical, story-based evaluation tool: The Most Significant Change technique. *American Journal of Evaluation*, *24*(2), 137–155. doi.org/10.1177/109821400302400202

Dunning, T. (2012). Natural experiments in the social sciences: A design-based approach. Cambridge, U.K.: Cambridge University Press.

Patton , M. Q. (2012). Planning and evaluating for social change: An evening at SFU with Michael Quinn Patton. [Web Video]. Retrieved from www.youtube.com/watch?v=b7n64JEjUUk&list=UUUi_6IJ8IgUAzI6JczJUVPA...

Tomison, A. (2000). Evaluating child abuse protection programs (Issues in Child Abuse Prevention No. 12). Melbourne: National Child Protection Clearinghouse. Retrieved from www.aifs.gov.au/nch/pubs/issues/issues12/issues12.html

Westhorp, G. (2014). Realist impact evaluation: An introduction. London: Overseas Development Institute.

White, H., & Sabarwal, S. (2014). Quasi-experimental design and methods. *Methodological briefs: Impact Evaluation*, *8*, 1–16.