Fact sheet: An overview of research designs for impact evaluations

This fact sheet outlines a variety of experimental and quasi-experimental research designs for your impact evaluation, with a brief description and information on when they are best used.

Top TSIP evaluation tips

- The pre-post design (without comparison group) is the simplest, but also the weakest, design to estimate the impact of your programme: It doesn't allow you to distinguish between the effect of your programme versus the effect of other factors on your outcome indicator(s)
- Pre-post designs with (non-randomised) comparison groups allow you to distinguish to some extent between the effect of your programme versus the effect of other factors on your outcome indicator(s), but you cannot be very confident about it
- The pre-post design with a randomised comparison group (RCT) is the most rigorous design and can confidently distinguish between the effect of your programme versus the effect of other factors on your outcome indicator(s)
- All designs can be improved by increasing the sample size and ensuring that participants are not lost during the evaluation

Introduction

Generally speaking, there are two different categories of research designs for impact evaluations:

Quasi-experimental designs

- The researcher controls the independent variable i.e. chooses which participants receive treatment and when. There is no randomised comparison group, but other (less rigorous) methods are used to simulate the counterfactual.
- Such designs include:
  - Pre-post with no comparison group: the treatment group's pre-test scores are used as the counterfactual
  - Pre-post with non-randomised comparison group: the treatment group's scores are compared with those of a constructed comparison group

Randomised experiment designs: the RCT

- The researcher controls the independent variable i.e. chooses which participants receive treatment and when. There is a randomly assigned comparison group.

In order to show the impact of an intervention you have to be able to compare the change in outcome indicators of the intervention with the counterfactual, which is defined as what would have happened without the intervention. All of the following designs aim to address the counterfactual with varying degrees of rigour.

Pre-post with no comparison group[1] (quasi-experimental)

Standard of Evidence: Level 2

Description: This is the weakest form of quantitative evaluation. There is just one group of participants – a treatment group – consisting of individuals receiving the intervention. Outcome indicators are generally measured i) shortly before participants start receiving the intervention (pre-test, or baseline), and ii) at one or more points after they have started receiving the intervention (post-test and follow-ups, or endlines). The period of time between pre- and post-tests depends on the intervention, but the post-test outcome indicators are usually measured immediately after the beneficiaries have completed the intervention, and follow-ups should on average not be collected before 6 months after the post-test. If there is a positive change between these the pre-test and the post-test (and follow-ups), it may have been caused by the intervention.

Counterfactual[2]: The measure shortly before participants start receiving the intervention (pre-test) is used as the counterfactual. This assumes that if the intervention hadn't taken place, this measure would have remained the same. Whilst this may be true in some cases, this is generally a flawed assumption – there are potentially a large number of other factors that could have had an effect on the outcome indicators being measured, and so it is difficult to be sure that the change was due to the intervention itself.[3]

What data is essential: Pre- & post-test scores for the treatment group.

Improving the design: You can try to identify and take into consideration any influences other than your programme that may have had an effect on the measured outcome indicators in order to produce a better estimate of the true impact of your programme itself. This is possible by i) providing a theoretical analysis as to what other factors may or may not have had an impact on them (this may include a literature review or asking participants (by survey or interview) what factors had an impact on them (in terms of the outcome indicator(s) being measured)), and ii) adjust for these factors in your analyses in a transparent way. However, these approaches are considered much weaker than having a comparison group.

[1] Learn more about pre-post with no comparison group on pp.108-111 in Shadish 2002
[2] Learn more about the counterfactual in Morgan 2007
[3] Learn more about contextual factors that could have an impact on your outcome (other than your programme) on pp. 10-15 in Togerson 2008

Pre-post with non-randomised comparison group[4] (quasi-experimental)

Standard of Evidence: Level 2 or Level 3 – dependent on how comparable the comparison group is.

Description: The same guidelines apply as above with regards to measuring outcome indicators at pre-, post-test and follow-up, but this design also includes a non-randomised comparison group – a group of similar individuals that will not be receiving the intervention for the duration of the evaluation. These will often be:

- Individuals from a site / region that the programme is not currently operating in (e.g. classroom, school, prison, GP practice, job centre etc).
- Individuals that have applied to the programme but have been unsuccessful in their application

Counterfactual: The outcome indicator(s) in the comparison group. This is preferable to the pre/post-test with no comparison group because with the comparison group you can take into account some contextual factors that have an impact on the outcome indicator(s) of your evaluation (e.g. indirect effects of an economic crisis); however, because the comparison group is not randomised, it may still be different to the treatment group in terms of other factors, some of which may affect the outcome indicators being measured. For example, a comparison group consisting of unsuccessful applicants to a tutoring programme may not be very comparable as they are probably less motivated and/or academically capable on average than the successful applicants in the treatment group.

Improving the design: This depends on how comparable the chosen comparison group is to the treatment group. There are (a) different things you can do to make your comparison group as comparable as possible, and (b) different types of statistical approaches to adjust for the fact that the comparison group is not (statistically) identical to the treatment group.[5] These statistical approaches ideally should be specified during the planning of the evaluation design.

(a) The different types of things you can do to make your comparison group as comparable as possible include:

- Selection of participants: Try to ensure that the treatment group participants do not differ systematically from the comparison group (e.g. the treatment group should not on average have higher motivation levels)
- Matching: If possible, participants in both groups should be matched in terms of variables that might affect the outcome indicators being measured (e.g. gender, age, prior attainment / health / offending etc).
- Comparing pre-test scores: Comparing the pre-test scores of the two groups is a useful way of gauging how appropriate the comparison group is – if the two groups'

[4] Learn more about non-randomised comparison groups on pp. 136-153 in Shadish 2002
[5] Learn more about these statistical approaches in the Boslaugh 2012, Morgan 2007 and Shadish 2002 books

pre-test scores are far apart, you know that the two groups are too different to make a good treatment and comparison group.

(b) There are a number of different statistical approaches to analysing the data and/or adjusting for differences between the comparison and treatment groups. Some of these are quite advanced, and it is recommended that you read about them in more detail to get a better understanding of what they entail (please see the books recommended in footnote 5 below).

- Simple difference (very basic)
  - Description: the difference between outcome indicators of participants and non-participants at post-test.
  - What data is essential: post-test scores of treatment & comparison groups on main outcome indicators
- Difference-in-difference (very basic)
  - Description: difference in change in outcome indicators of participants and non-participants at post-test (statistics difference)
  - What data is essential: pre- & post-test scores of treatment & comparison groups on main outcome indicators
- Interrupted time series (basic)
  - Description: Graphical approach to a natural experiment - you are looking for sudden differences in your main outcome indicator that you observe over a longer time period (before, during and after you intervene with your programme), which enables you to observe the difference between the overall trend versus the impact of your programme.
  - What data is essential: several pre- & post-test scores for treatment & comparison groups on main outcome indicators
- Instrumental variables (if done well, quite good)
  - Description: uncorrelated variable (to outcome) predicts participation
  - What data is essential: pre- & post-test scores for treatment & comparison groups on main outcome indicators and other indicators that you think predicts participation
- Multivariate regression (if done in advance, good)
  - Description: compare participants with non-participants and control for potential influences
  - What data is essential: pre- & post-test scores for treatment & comparison groups on main outcome indicators and other indicators you want to control for (confounding variables)
- Regression discontinuity (if done well, very good)
  - Description: methods for generating an unbiased comparator group in a natural experiment by looking at cases either side of an important threshold (potential participants ranked, define cut-off which defines treatment group) the design selects people into their intervention groups on some pre-test

variable with a pre-defined cut-off; if properly implemented this approach can produce unbiased estimates of effect sizes – albeit less efficiently than an RCT.

- What data is essential: pre- & post-test scores for treatment & comparison groups on main outcome indicators and the outcome indicator used for ranking & defining the threshold (if different from main outcome indicators)

- **Statistical matching: (if done in advance, very good, otherwise still good)**
  - Exact matching
    - Description: find someone (statistically) identical for each participant
    - What data is essential: pre- & post-test scores for treatment & comparison groups on main outcome indicators and other indicators that are needed to match participants
  - Propensity score matching
    - Description: used for matching control cases in a natural experiment, based on identifying characteristics most likely to be linked to the participation in the intervention arm (according to mix of what are thought to be predictive characteristics)
    - What data is essential: pre- & post-test scores for treatment & comparison groups on main outcome indicators and other indicators that are needed to match participants

However, each type of quasi-experimental design has its particular weaknesses.[6]

Randomised experiment design: pre-post-test with randomised comparison group[7] (experimental)

Description: Randomised controlled trials (RCTs) are the gold standard among experimental designs in social intervention research. An RCT is defined as an evaluation of an intervention which is manipulated so that at least one randomly allocated sub-group receives the treatment and at least one does not. Through random assignment you can eliminate the risk of allocation bias, allowing for two statistically identical groups to be created. If you then administer your programme to one group but not the other, any differences in outcome indicators between these two groups after the programme can be confidently attributed to the programme (i.e. not to external factors) since the two groups only differ in whether or not they receive the treatment. This is because randomisation evenly distributes characteristics across groups (e.g. if you have 50 men and 50 women, and you randomise all of them to treatment and control groups, you will get close to 25 men and women in both groups, and all other observable and unobservable characteristics will also be evenly distributed). You can either randomise individuals or clusters (i.e. groups of individuals, e.g. schools). Generally, it is better to randomise individuals because you need much smaller

[6] Learn more about the weaknesses of these quasi-experimental designs in the Gorard 2013 book
[7] Learn more about randomised designs in the fact sheet on RCTs or refer to the Togerson 2008 book on RCTs

sample sizes in order to demonstrate a significant effect, but in some cases that is not possible.

Counterfactual: The outcome indicator(s) in the comparison group. A statistically identical comparison group is created through random allocation. It is absolutely essential to comply with the random allocation as even just the slightest change will compromise the validity of the RCT's results.

What data is essential: Pre- & post-test scores for the treatment & control groups

Improving the design: Since RCTs are the gold standard, it does not need to be improved if it is conducted properly. However, it is important to ensure that a large enough sample size is available in order to be able to detect the expected effect. Please refer to the RCT fact sheet for further information on how to optimise your RCT design.

Books referenced in this fact sheet

Boslaugh, S. (2012), *Statistics in a nutshell*, O'Reilly Sebastopol

Gorard, S. (2013), *Research design: Creating Robust Approaches for the Social Science*, SAGE Publications Ltd. London

Morgan, S. L., Winship, C. (2007), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge University Press, Cambridge

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Wadsworth Cengage Learning, Belmont

Togerson, D. J., Togerson, C. J. (2008), *Designing Randomised Trials in Health, Education and the Social Sciences*, Palgrave Macmillan Basingstoke

Glossary

Allocation bias—when there is a systematic difference between the individuals that are assigned to the treatment group and the individuals that are assigned to the comparison group (this leads to selection bias).

Baseline—the measurement of the outcome indicator taken shortly before the intervention is started (also called pre-test).

Bias—A term denoting that a known or unknown variable is or may be responsible for an observed effect other than the intervention.

Confounders— A variable associated with cause and outcome; can mask a true relationship between another variable and outcome. E.g. if you are trying to measure the impact of an afterschool maths club on children's numeracy in London, and your comparison group consists of schools in a less affluent area of London, the general quality of the schools' teaching might be a confounding variable – i.e. if numeracy scores improve less among the comparison group, that might be due to a poorer quality of teaching rather than the lack of the afterschool maths club.

Counterfactual—what change in outcome indicators would have been seen in the absence of the intervention. It is not possible to observe the counterfactual (since it does not actually exist), but it can be simulated approximately with a comparison group.

Experimental designs[8]—a design that involves deliberatively introducing an intervention to observe its effects against a (randomly assigned) comparison group.

Impact—the long-term, cumulative effect of programmes/interventions over time on what they ultimately aim to change.

Independent variable—the things whose impact is being measured (in our case this is generally the intervention).

Indicator—a quantitative or qualitative variable that provides a valid and reliable way to measure achievement, assess performance, or reflect changes connected to an intervention.

Natural experiment—evaluation of an intervention not manipulated by the researcher, meaning that cases cannot be randomly allocated to control and treatment groups.

Outcome—short-term and medium-term effect of an intervention's outputs, such as change in knowledge, attitudes, beliefs, behaviours.

Post-test—the measurement of the outcome indicator taken shortly after the intervention has finished (also called endline).

Randomised controlled trial (RCT)[9]—a highly rigorous experimental design which is able to attribute program effects exclusively to the programme (i.e. not external factors or 'noise') by comparing a treatment group with a (statistically) 'identical' comparison group, which is achieved by randomly allocating participants in either treatment or comparison group.

Selection bias—a bias that happens when choosing to put participants into a treatment group and comparison group for a specific reason (e.g. because they failed in their application to be on the programme) rather than through randomisation. This means that the treatment and control groups differ already *before* they receive the intervention.

[8] Learn more about randomised experiments in Shadish chapter 8, pp. 246-278
[9] Learn more about the randomised controlled trial in Torgerson