

Fact Sheet 3: Randomised and non-randomised designs

This fact sheet summarises features of randomised and non-randomised designs, and the advantages and disadvantages of each. The fact sheet also explains confounding factors and how to manage them.

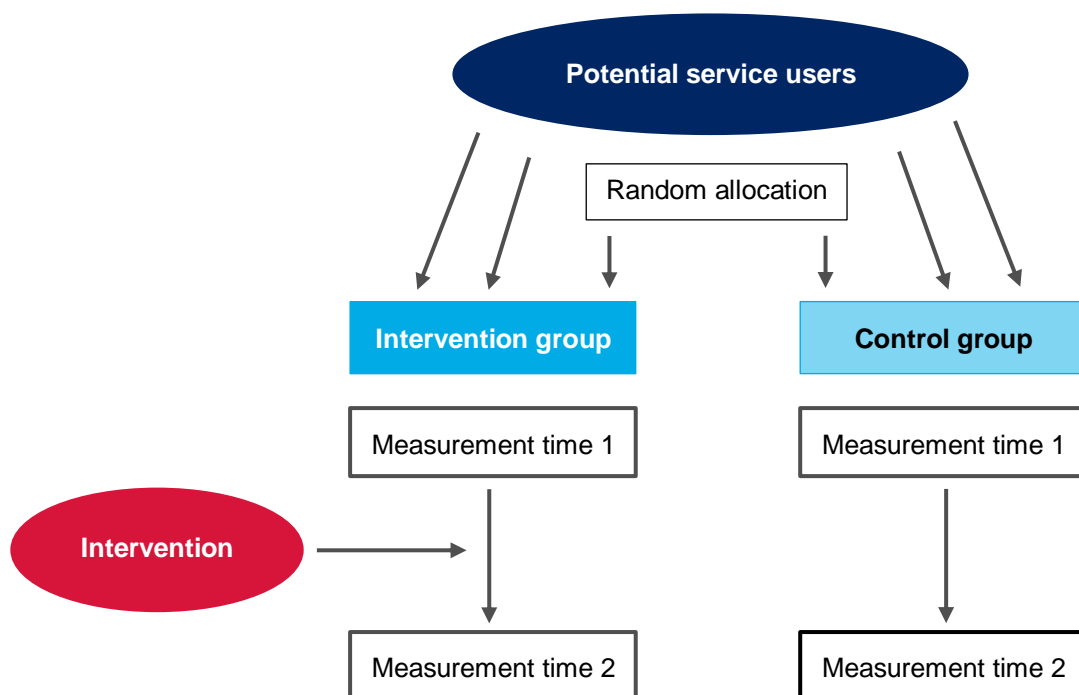
📖 Refer to [HM Treasury's Magenta Book](#) for more information on these designs.

1. Randomised designs

A randomised design is one where eligible individuals or groups of individuals are randomly allocated to either the intervention group (to receive the intervention) or the control group (do not receive an intervention).

With a sufficiently large number of individuals, randomisation ensures that the characteristics of the individuals in each group are the same. This means that any characteristics that could influence the outcome of interest (e.g. the number of cigarettes smoked in a smoking cessation program) are equally balanced between the two groups and that any difference between the two groups in terms of the outcome can be attributed to the intervention.

Figure 1: The logic of random allocation



1.1 Individual randomised design

An individual randomised design randomly allocates eligible individuals to either the intervention group or the control group. Typically, in a randomised design with two groups (one intervention and one control), there is a 50:50 chance of eligible individuals being allocated to either group.

Box 1: Example of individual randomised design

Let us consider a new rehabilitation program aimed at reducing recidivism following release from custody. At the time of release, eligible individuals could be randomly allocated to the rehabilitation program (intervention group) or standard practice (control group).

1.2 Cluster randomised design

When randomisation of individuals is not practical or desirable, you could consider cluster randomised designs where groups (or “clusters”) of individuals are randomly allocated to either the intervention or control group. This is often the case when an intervention is better delivered to groups of individuals (e.g. group training sessions or geographical rollout). Cluster trials also help prevent “contamination bias” where individuals receiving the intervention influence those not receiving the intervention when in close proximity.

Box 2: Example of cluster randomised design

Let us consider an intervention that aims to prevent harassment and bullying in primary school children via face-to-face training sessions. In this instance, it would be difficult to randomise individuals from the same school to either the intervention or control group for two reasons:

1. Children from the same school are likely to talk to each other and share the information learned during the training sessions with those randomly allocated to the control group (“contamination” bias).
2. It seems more natural and effective for such an intervention to be delivered at the school level.

Using a cluster randomised design, you could select all schools within a geographical area (e.g. Sydney) and randomly allocate schools to either receive or not receive the intervention.

1.3 Stepped-wedge design

One disadvantage of traditional randomised designs, both with individual and cluster randomisations, is that (typically) half of the eligible population is allocated to a control group and does not receive the intervention. While this is acceptable when there is clear uncertainty about the potential benefits of an intervention (e.g. a new drug in a drug trial), it can pose ethical and logistical problems when the intervention is strongly believed to have benefits which are difficult to withhold from eligible individuals. In this instance, alternative randomised designs involving a delayed intervention can be considered. Those designs are generally called “stepped-wedge” designs and involve randomly allocating individuals or groups of individuals to different starting times.¹

For example, with three randomly allocated groups, one group would start the intervention straight away, the second group a bit later, and the third group even later. With this design, everyone who is eligible receives the intervention while maintaining the advantages conferred by randomisation (i.e. making sure that all three groups are comparable).

¹ Hemming et al., 2015

Box 3: Example of stepped-wedge design

Let us consider again the school bullying example. There may be some who consider the intervention should be offered to all schools or perhaps some schools are reluctant to participate. In these cases, you could consider a stepped-wedge design where schools are randomly allocated to different starting dates. A possible stepped-wedge design could involve three groups of schools starting the intervention at three different times. All end up receiving the training.

The design is represented in the table below where dark blue cells represent the period/s when schools are receiving the intervention. In the first period, none of the schools receive the intervention, providing baseline information on the rate of bullying and harassment in the absence of an intervention. At each subsequent period, the intervention is introduced to a new group of schools. By the fourth period, all schools are receiving the intervention.

	Period 1	Period 2	Period 3	Period 4
1 st group of schools				
2 nd group of schools				
3 rd group of schools				

2. Non-randomised designs

Randomised designs may not always be appropriate or feasible. This may be the case when implementing macro policies or system-wide changes where it is difficult to influence intervention allocation or when it may be seen as unethical to withhold an intervention already proven to be effective. Under those circumstances, it is still essential to have an appropriate control group which will be observed and measured as similarly as possible as the intervention group.

The main drawback of non-randomised designs is the limited ability to guarantee the comparability of the intervention and control groups. Proposals to use a non-randomised control group should, as much as possible, adhere to the following rules (note that points 3 and 4 also apply to randomised designs):

1. Collect baseline data on all known confounds (i.e. characteristics with the potential to influence the outcomes) for both the intervention and control groups.
2. Collect data on the primary outcome at multiple times in both groups, with at least one data point before the start of the intervention and at least one after. This is so you can try to separate the effect of the intervention from other factors in the environment.
3. Ensure that both the intervention and control groups are observed at the same time.
4. Ensure that both the intervention and control groups follow the same data collection procedure.

However, sophisticated non-randomised designs are available to help strengthen basic controlled before-and-after approaches, especially where confounds are well understood and measured, causal pathways to outcomes are known, and the impact is anticipated to be large.

Assuming the above rules can be followed, a range of methods is then available for establishing the counterfactual. The main ones involve:

- multivariable analyses
- matching (e.g. using propensity scoring)
- comparing changes over time between the intervention and control groups (sometimes called “difference in differences”).

In most cases, analysis involves a mixture of the above methods. For example, by first identifying a well-matched subset of individuals using a propensity score and then by comparing changes over time between the intervention and control groups. More details about these methods are available in Sections 9.33 to 9.48 of the [Magenta Book](#).

2.1 Non-randomised designs with a control group

Propensity score matching

Propensity scoring calculates the probability of being a participant in an intervention given a range of individual characteristics measured before enrolling in the intervention, both for intervention participants and individuals not participating. This is typically done using a logistic regression where the outcome is an indicator of enrolment (yes or no) and predictors include a range of factors with the potential to influence enrolment and/or outcomes. Once the probability has been calculated for everyone, the idea is to match each intervention participant to one or more control participants with the same probability. Those with no appropriate match are not included in the analysis and individual characteristics across the two groups are balanced.

The main limitation of propensity matching is that, as opposed to randomisation, it is only able to balance observed characteristics. For more details, refer to Peter Austin’s introduction to propensity score methods.²

Box 4: Example of propensity score matching – UK’s Peterborough social impact bond (SIB)³

The Peterborough SIB aimed to reduce reconviction rates for short sentence male prisoners leaving HMP Peterborough. It was launched in September 2010 and provided interventions for adult males (aged 18 years or over) receiving custodial sentences of less than 12 months (‘short sentence prisoners’) and discharged from HMP Peterborough.

The UK Ministry of Justice and Social Finance proposed a matched control group to remove the influence of external events on reconviction levels (e.g. changes in sentencing policy, the economic environment, etc.). Consequently, the approach outlined in the SIB contract was to develop a control group of prisoners discharged from other prisons during the same time period as the Peterborough cohort. This control group was developed using propensity score matching.

The 936 men released from HMP Peterborough and 9,360 released from other prisons were successfully matched on the propensity score in terms of demographics and criminal history. The analysis showed an 8.39% reduction in reoffending rates within the Peterborough Cohort 1, which was greater than the control group but insufficient to trigger payment under the terms of the bond.

² Austin P. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res.*; 46(3): 399–424

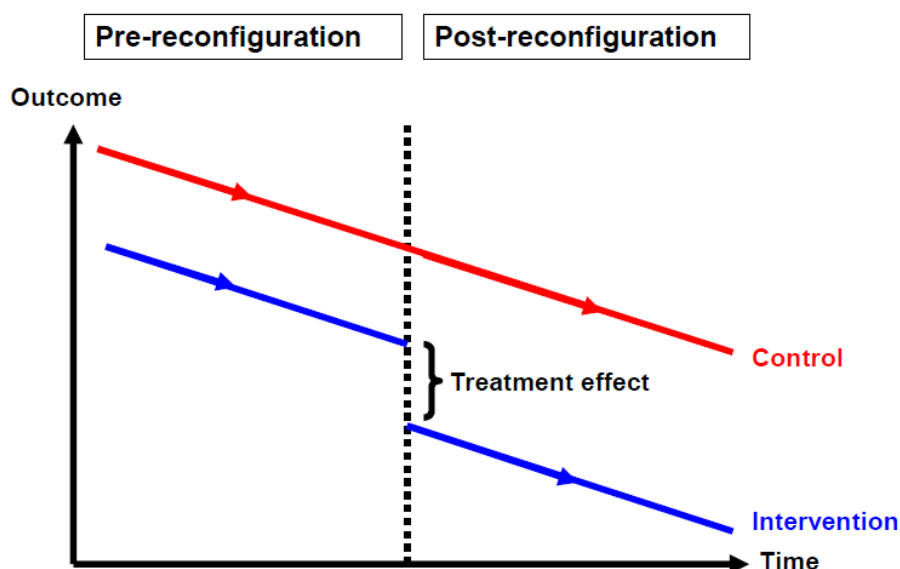
³ Cave, Williams, Jolliffe and Hedderman (2012) and Jolliffe and Hedderman (2014)

Difference-in-difference design

Difference-in-difference designs use time trends in an outcome. In this design, trends in associated outcomes between intervention and control groups are compared over a time period relevant to the intervention. While unobserved factors might affect the outcome, if they do not affect trends in the outcome, then the trends for both groups in the absence of the intervention will be the same. Any significant difference in trends can be interpreted as an intervention effect. This is the so-called parallelism or “common trends” assumption.

The parallelism assumption should always be verified where possible, either by examining the pre-intervention trends in historical time series data or from previous studies. Where the assumption does hold, this design is a useful method that can address selection bias in the absence of rich information about the participants. But the parallelism assumption should not be automatically assumed true, and this approach would not be recommended if, for example, data are only available at two time points (before and after the implementation of an intervention). The basic approach of this design to consider an intervention reconfiguration is illustrated below.⁴

Figure 2: Difference-in-difference design



Box 5: An example of a difference-in-difference evaluation⁵

WHC ran a pilot from February 2006 to February 2008. It was a free, no-obligation service that aimed to provide small and medium-sized enterprises with advice on workplace health issues to increase the level of healthy workplaces across England and Wales.

The primary outcome of interest was a net beneficial impact on the incidence and duration of occupationally related ill-health and injury. Employers operating in regions where the WHC workplace visit service was not provided were the control group, on the basis that they were similar (in terms of their size and sector) to those participating in the WHC pilot. Their outcomes, therefore, constituted the best available estimate of the counterfactual.

⁴ Morris, S. et al. (2014). Impact of centralising acute stroke services in English metropolitan areas on mortality and length of hospital stay: difference-in-differences analysis. *BMJ*, 349, g4757

⁵ For more guidance, see HM Treasury (2011), *The Magenta Book. Guidance for evaluation*

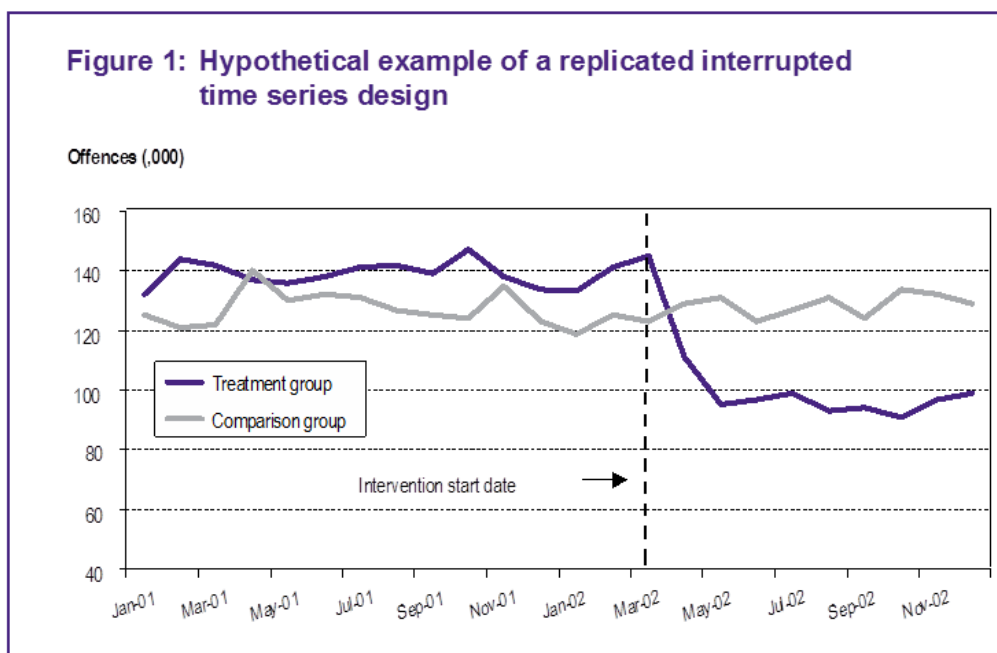
One way of evaluating the impact of the WHC pilot would have been to look directly at the relationship between involvement in the pilot and final outcomes. However, this approach was considered unlikely to produce robust results because, in addition to improving safety, using the pilot can change the way final outcomes are recorded.

Instead the relationship was analysed in two stages, looking first at the effect of the WHC pilot on intermediate outcomes and then looking at the effect of the intermediate outcomes on the final outcomes. These relationships were examined using difference-in-difference analysis. This looked at the changes in outcomes between the two survey waves, and tested whether these changes were different for the WHC pilot intervention and control groups. There was no evidence that taking part in WHC had a direct measurable effect on absenteeism due to illness. There was, however, evidence that involvement in WHC improved a range of health and safety practices. These were linked to reduced accident rates.

Replicated interrupted time series

An extension of the simple difference-in-difference design consists of taking one or more observations from both the intervention and control groups prior to introducing an intervention. This is sometimes called an 'interrupted time series'. If the intervention is working, we would expect the second set of observations to move in a more favourable direction for the intervention group than for the control group. Notice that by comparing outcomes for both intervention and control groups after the intervention, we are able to control for any known factors that might be influencing trends in both groups. An example of this design is provided in the figure below.

Figure 2: Interrupted time series design



Historical baseline

Historical data can be used to compare outcomes for the intervention group against past outcomes for a similar cohort. To establish a similar cohort, characteristics such as age, needs, and referral pathways are controlled for in the analysis.

This approach requires a stable cohort with a stable level of outcomes over a number of years.

Analysis of historical baseline information is used to develop a rate card approach to outcomes contracting. In a rate card approach, a historical baseline is analysed to establish and control for dead weight i.e. the outcomes that would have been achieved without an intervention. This analysis is used to inform the pricing that Government is willing to pay per outcomes achieved.

Box 6: Example of a rate card underpinned by a historical baseline⁶

In the London Homelessness Social Impact Bond (SIB), one of the outcomes targeted was a reduction in rough sleeping. This would be measured by a reduction in the number of individuals recorded in CHAIN.

CHAIN is a comprehensive database that records individuals' demographic information, support needs, and movement in and out of rough sleeping and hostel accommodations. The database is unique to London. This also meant it was not possible to create a matched comparison group.

By modelling cohorts over time using data in CHAIN and other available data, a baseline of what is expected to be achieved without a specific (targeted) intervention was established i.e. the historical baseline.

The baseline then provided the basis for modelling what incremental outcomes the proposed intervention might achieve and what savings and benefits are generated. By understanding and accounting for the 'deadweight' (i.e. what would be achieved in the absence of the intervention) in the calculation of savings and benefits, Government can determine a price per outcome that reflects value to Government.

2.2 Non-randomised designs without a control group

Simpler comparisons can be illuminating, but can be difficult to interpret because they do not include a way to estimate the counterfactual. These methods⁷ include for example:

- Comparing before and after the intervention
- Comparing 'dose' of change or degree of exposure to change across settings
- Regression discontinuity design
- Instrumental variable estimation.

In general, designs with no control group are discouraged as they do not provide a reliable opportunity to separate the effect of an intervention from possible simultaneous changes occurring over the course of the program. However, it is recognised that in exceptional circumstances such designs may be justifiably proposed.

3. Summary

In summary, a randomised trial provides the most reliable framework for assessing the impact of an intervention because, when sufficiently large it provides a control group that only differs from the intervention group in terms of the intervention being received or not.

⁶ Department for Communities and Local Government (2014). *Qualitative Evaluation of the London Homelessness Social Impact Bond: First Interim Report*. Department for Communities and Local Government: London

⁷ For more guidance, see HM Treasury (2011), *The Magenta Book. Guidance for evaluation*

Design	Advantages	Disadvantages
Randomised individual trial	Balances both measured and unmeasured confounds. Provides the most robust evidence of impact.	Can be resource intensive and/or not acceptable for certain interventions. Those randomised out of the intervention will not be able to receive the intervention.
Randomised cluster trial	Ability to randomise clusters/groups of individuals when individual randomisation is not appropriate.	Increases the sample size compared to an individual randomised trial.
Stepped-wedge trial	Allows everyone to receive the intervention while retaining the benefits of randomisation.	Takes longer than an individual or cluster trial and is more complex to analyse.
Propensity matched control	Provides a more flexible framework than a randomised design with a plausible counterfactual.	Substantially increases the risk of bias in the presence of unmeasured confounds.
Difference-in-difference	Provides a more flexible framework than a randomised design with a plausible counterfactual.	Substantially increases the risk of bias in the presence of unmeasured confounds.
Interrupted time series	More flexible than randomisation, making use of time trends and a plausible counterfactual.	Difficult to attribute effect to the intervention due to the absence of a counterfactual; can be strengthened by addition of an observed control group.
Historical baseline	Provides a more flexible framework than a randomised design with a plausible counterfactual. Has some control for deadweight.	Substantially increases the risk of bias in the presence of unmeasured confounds.

Table 1 below briefly summarises the advantages and disadvantages of each design.

Table 1: Advantages and disadvantages of main designs

Design	Advantages	Disadvantages
Randomised individual trial	Balances both measured and unmeasured confounds. Provides the most robust evidence of impact.	Can be resource intensive and/or not acceptable for certain interventions. Those randomised out of the intervention will not be able to receive the intervention.
Randomised cluster trial	Ability to randomise clusters/groups of individuals when individual randomisation is not appropriate.	Increases the sample size compared to an individual randomised trial.
Stepped-wedge trial	Allows everyone to receive the intervention while retaining the benefits of randomisation.	Takes longer than an individual or cluster trial and is more complex to analyse.
Propensity matched control	Provides a more flexible framework than a randomised design with a plausible counterfactual.	Substantially increases the risk of bias in the presence of unmeasured confounds.

Design	Advantages	Disadvantages
Difference-in-difference	Provides a more flexible framework than a randomised design with a plausible counterfactual.	Substantially increases the risk of bias in the presence of unmeasured confounds.
Interrupted time series	More flexible than randomisation, making use of time trends and a plausible counterfactual.	Difficult to attribute effect to the intervention due to the absence of a counterfactual; can be strengthened by addition of an observed control group.
Historical baseline	Provides a more flexible framework than a randomised design with a plausible counterfactual. Has some control for deadweight.	Substantially increases the risk of bias in the presence of unmeasured confounds.