Cabinet Office

# Experimental Designs

Evaluation Task Force Academy 2.0

# Hello

# ETF Evaluation Academy

**Module 1:** Introduction to Evaluation

**Module 2:** Developing a Theory of Change

**Module 3:** Scoping an Evaluation

**Module 4:** Process Evaluation

***Module 5:* Impact Evaluation - Experimental Designs**

**Module 6:** Impact Evaluation - Quasi-Experimental Designs

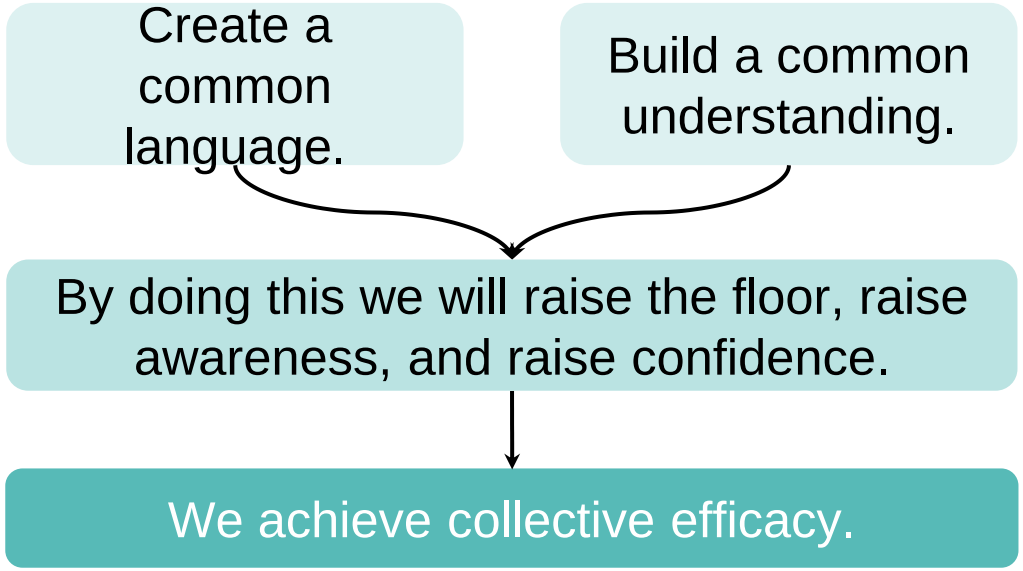
**Module 7:** Impact Evaluation - Theory-Based Designs

**Module 8:** Value for Money Evaluation

**Module 9:** Planning and Managing an Evaluation

**Module 10:** Communicating Evidence and Decision Making

*The Evaluation Academy will upskill analysts across HMG departments in key evaluation methodologies and evaluation management techniques and will result in better and more evaluation across HMG.*



# Module 5: Overview of contents

- Learning outcomes
- Hierarchy of evidence
- Randomised control trials
- Designing randomised controlled trials
- Types of RCT
- Comparing RCTs to other methods
- Critically evaluating RCTs
- Advocacy and application of learning

**Key**

**Know/Comprehend**: lowest level of complexity

**Apply/Analyse:** medium level of complexity

**Evaluate/Create:** highest level of complexity

# Learning outcomes

- I can explain the role and value of key experimental evaluation designs (e.g. experimental vs quasi-experimental, different types of RCTs such as staged, waitlist, parallel)
- I can identify the types of evaluation questions that experimental evaluations can or cannot answer
- I can explain how experimental evaluations can work alongside other evaluation methods
- I can explain the key things that need to be considered when preparing and running each experimental design, such as randomisation strategies, sample size, and analysis plan
- I can explain the benefits and risks of different experimental evaluation designs
- I can contrast the most appropriate experimental evaluation method(s) based on the policy context
- I can critically assess the findings of an experimental evaluation
- I can advocate for including experimental designs across the policy cycle

# ETF Evaluation Academy: Hierarchy of evidence

The Nesta Standards of Evidence

The objective of developing Standards of Evidence is to help us know how confident we can be in the evidence provided to show that an intervention is having a positive impact.

**Level 5**
You have manuals, systems and procedures to ensure consistent replication and positive impact

**Level 4**
You have one + independent replication evaluations that confirms these conclusions

**Level 3**
You can demonstrate causality using a control or comparison group

**Level 2**
You capture data that shows positive change, but you cannot confirm you caused this

**Level 1**
You can describe what you do and why it matters, logically, coherently and convincingly

## Ladder of evidence

**How can we be confident our activity makes a difference?**

Statements about 'what works'

- **Systematic reviews** (Based on level 3-5 studies)

Statements about 'what's promising'

- **5 Randomised controlled trials**
- **4 Before/after measures** Multiple site comparisons
- **3 Before/after measures** Two site comparisons

Statements about possible impact

- **2 Before/after measures** No comparison site
- **1 One-off measures** No comparison site

Study designs increasingly rule out potential alternative causes

Study designs cannot rule out potential alternative causes

# Randomised controlled trials

# Impact questions

- Focused on **outcomes**
- Tell you **the effect of something**
- Make a **comparison**

## Example questions

- What effect is my programme or policy having?
- Are the people that the programme or policy is supposed to serve better off because of it?
- Do more people sign up for my programme if I change the recruitment materials?

**Pros**

- ✓ Tell us whether a programme is effective or not
- ✓ More generalisable results

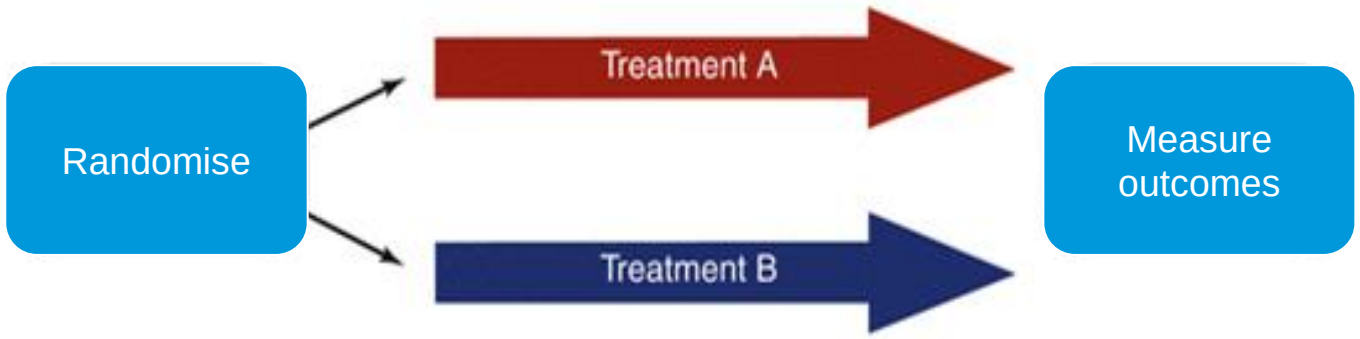
**Cons**

- × Can't explain *why* a programme does or doesn't work

# Why do we use experimental evaluations?

Experimental evaluations (also known as Randomised Controlled Trials - RCTs) can establish if programmes **causally** impact an outcome of interest.

Example: Does an anti-bullying campaign lead to better pupil outcomes in schools?



# RCTs have a control group

The control group serves as the **counterfactual**.

**Business as usual**

**People who aren't offered a programme**

**People who are on a waiting list**

**People who get a placebo drug**

# Why can’t we just compare any two groups?

**Observable differences**

**Unobservable differences**

**Imagine an educational programme evaluated across schools:**

School A has teachers with more experience than School B

School C has teachers with higher motivation than School D

# Activity

You work in DLUHC and are asked to support a business case for a programme which looks to simplify planning applications to increase the number of houses built in England.

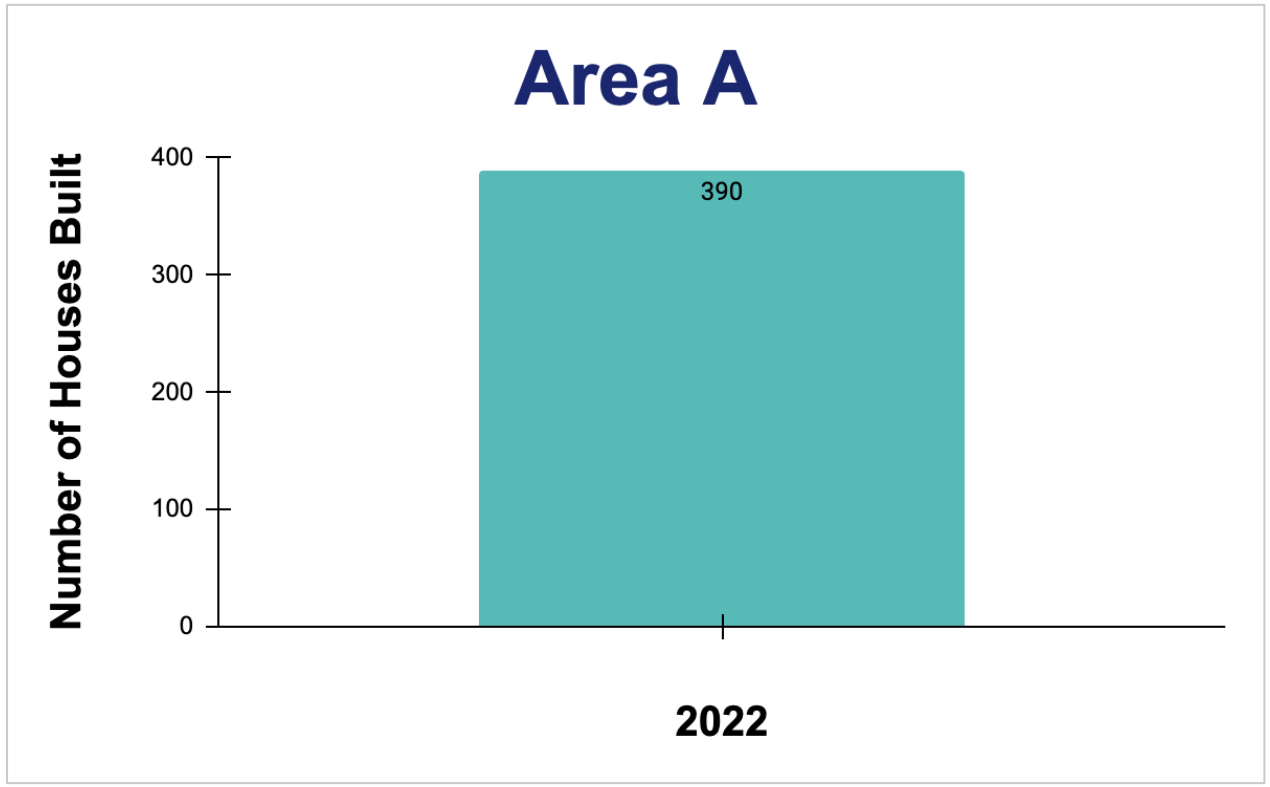
A group of Areas called ‘A’ implemented the programme in 2022, those in ‘B’ did not.

The programme can be scaled up at a cost of £10M. You will fund it if you believe it will increase the number of houses built by 5%.

We’re going to decide as a group if we are confident to fund the programme.

# Decision 1

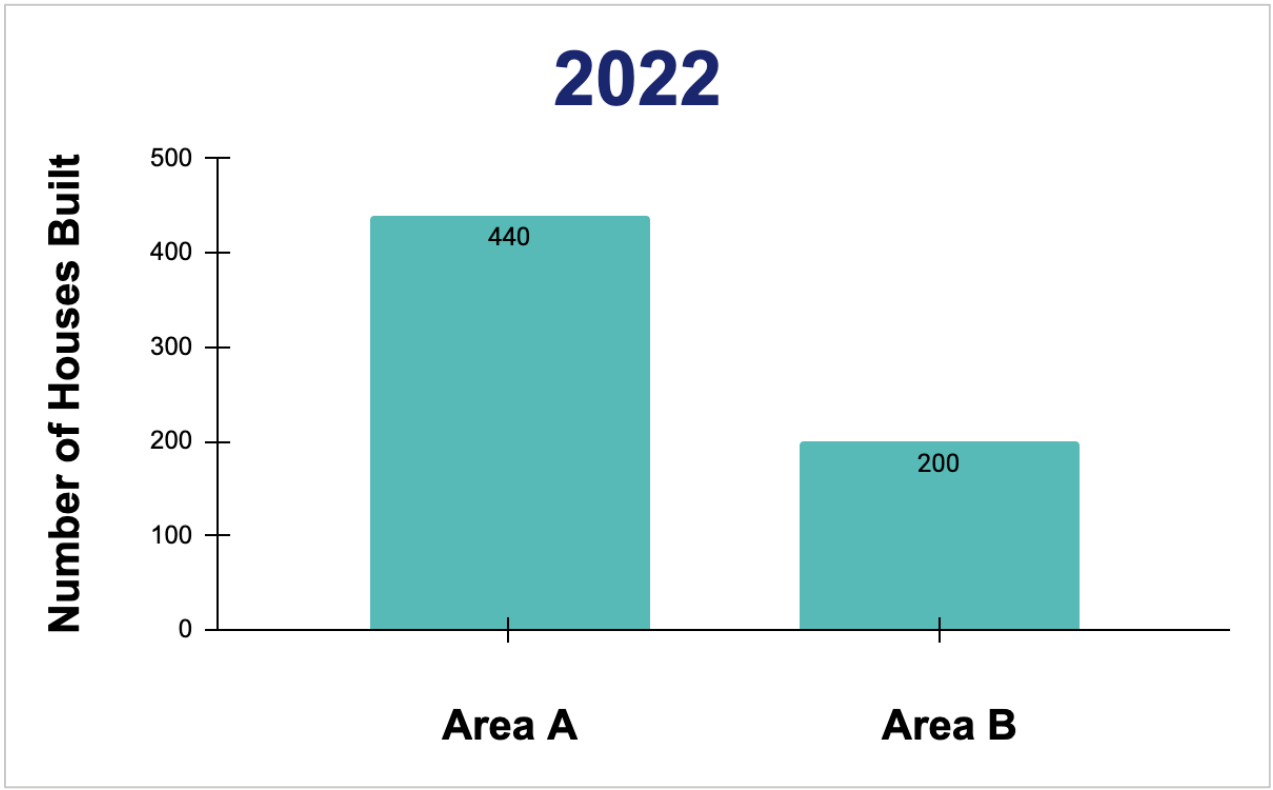
90% of planning officers liked the programme.




Fund

Do not fund

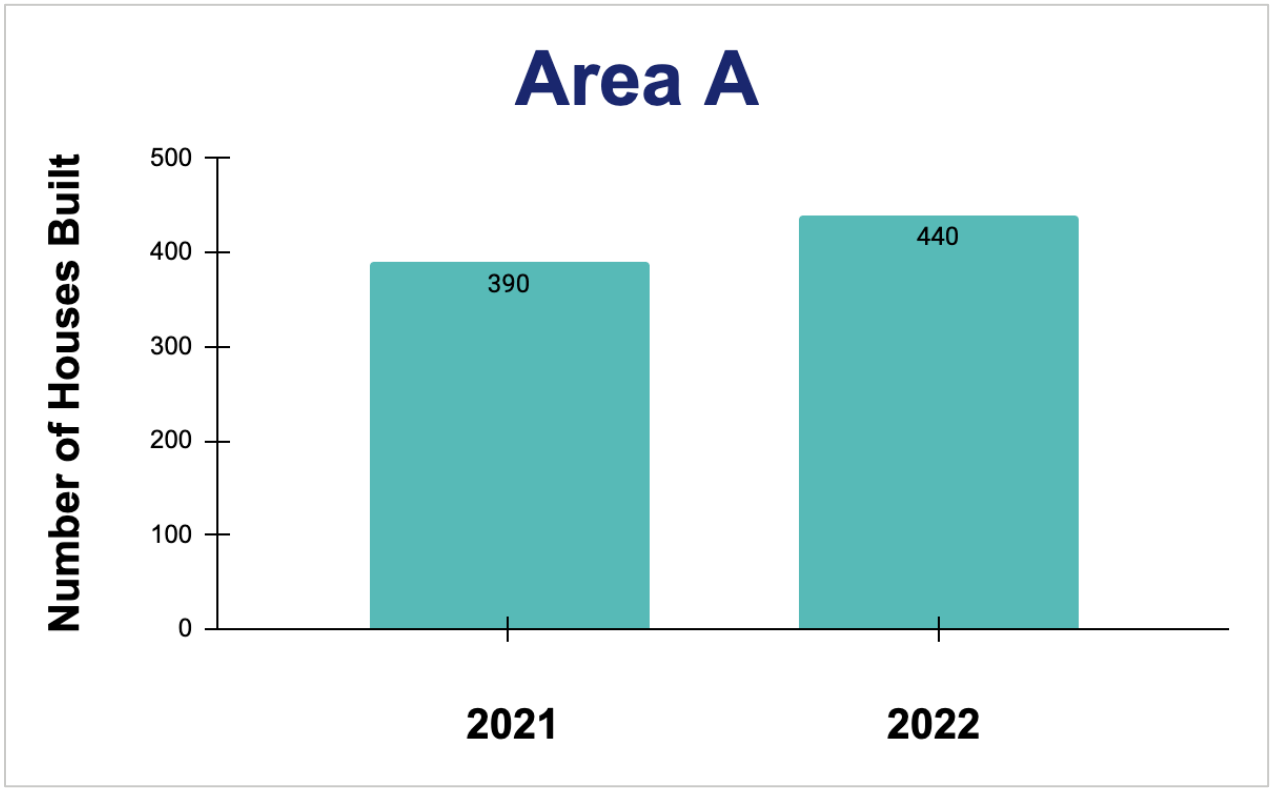
# Decision 2




Fund

Do not fund

# Decision 3

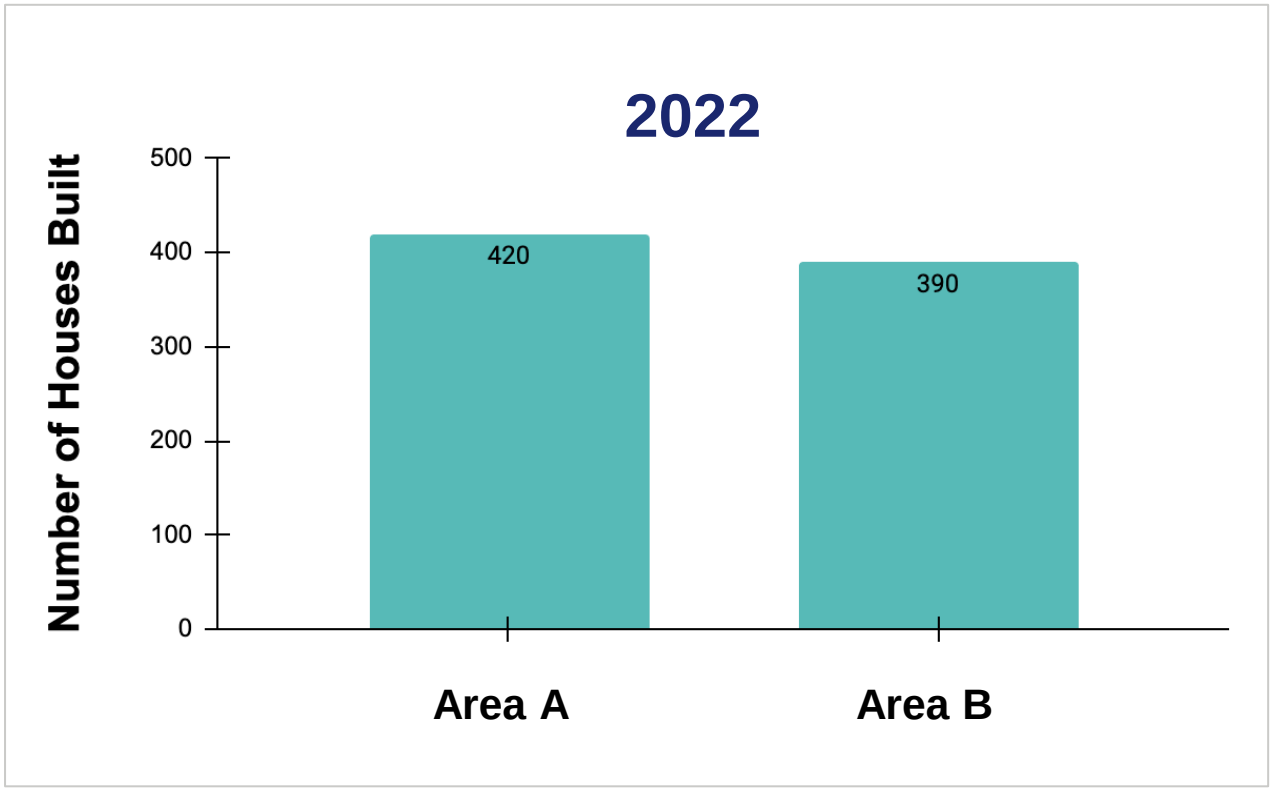


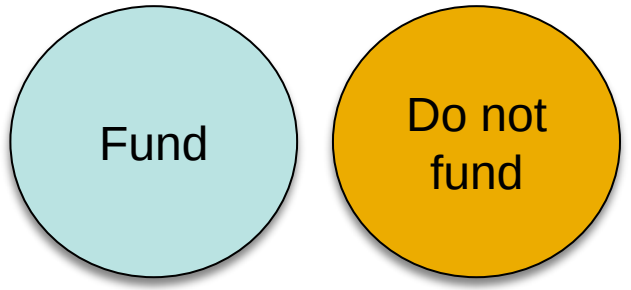

Fund

Do not fund

# Decision 4

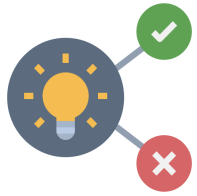
**Allocation to Area A and B was determined by a random lottery.**



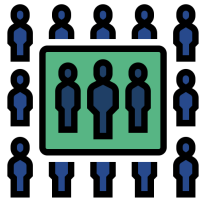



# Designing Randomised Controlled Trials

# Key things to consider when designing an RCT



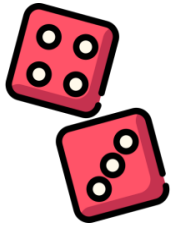
**Evaluation Question / Hypothesis**



**Sample size**

**Statistical power**

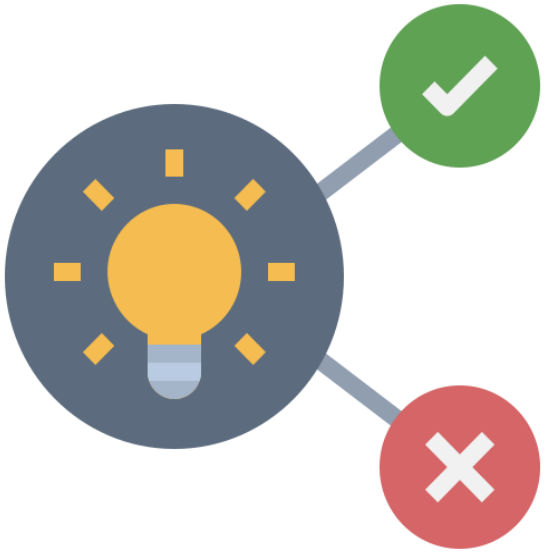
**Ethics**



**Randomisation**

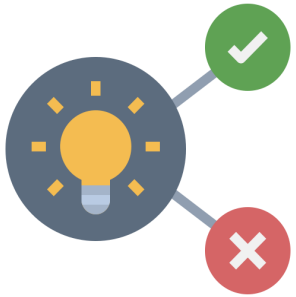
# Start with an evaluation question / hypothesis

- People who go to university earn more than those who don’t
- People spend too much on housing
- A school programme will increase student grades
- Not enough people are installing heat pumps in their home



# How to create an evaluation question / hypothesis

**Null hypothesis:** assumes a programme does not make a difference

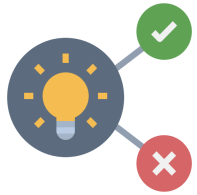


An educational programme has **no impact or a negative impact** on a student's grades.

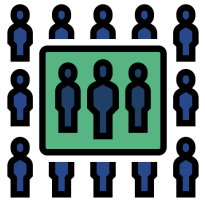
**Alternative hypothesis:** what we consider to be true if the null hypothesis is likely to be *false*

An educational programme **has a positive impact** on students' grades.

# Key things to consider when designing an RCT



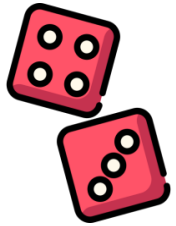
Evaluation Question / Hypothesis



Sample size

Statistical power

Ethics

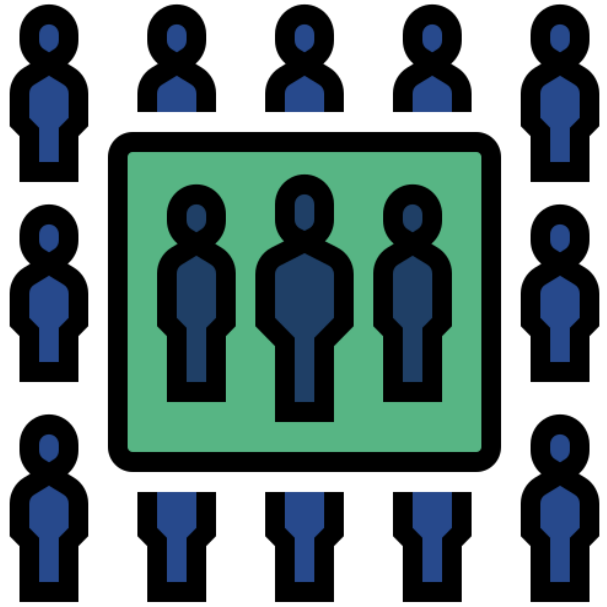


Randomisation

# Sample size

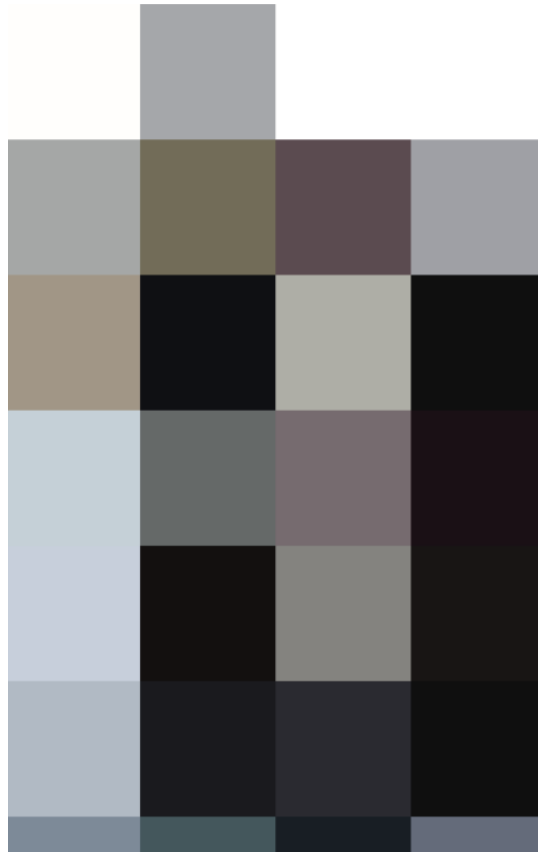
This is the number of observations in an experiment.

An educational programme that was delivered to 2,000 pupils, with 2,000 pupils in the control group, has a sample size of 4,000



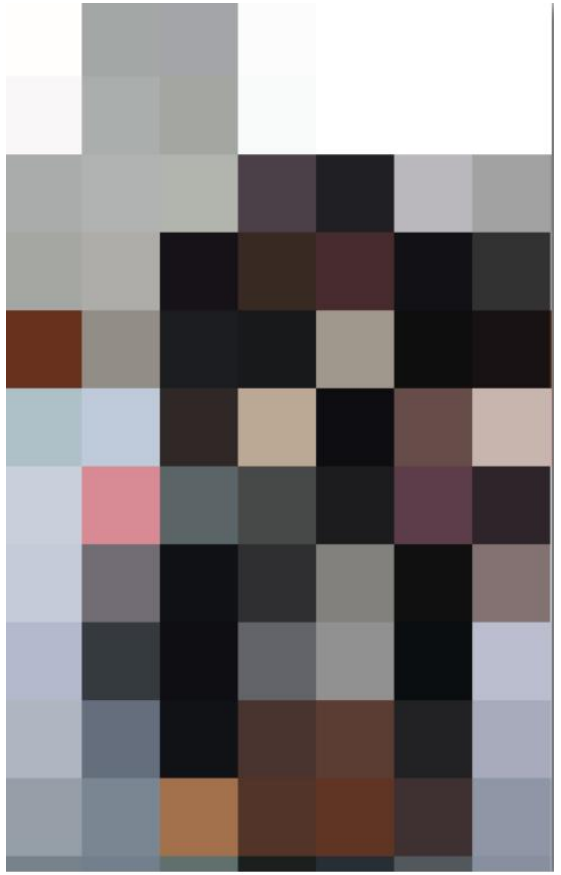
# How do I know if my sample is big enough?

28 data points?



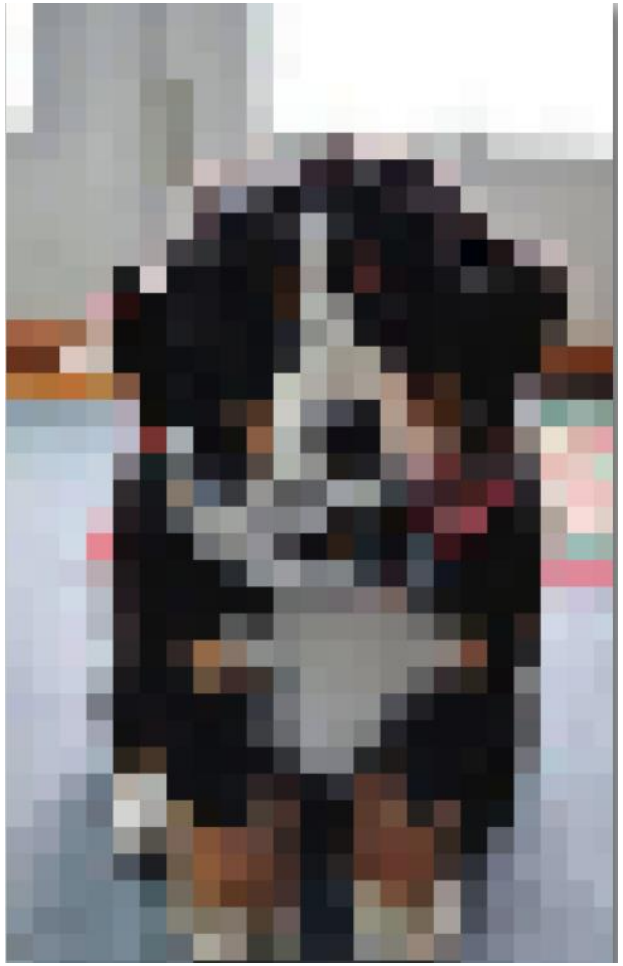
# How do I know if my sample is big enough?

84 data points?



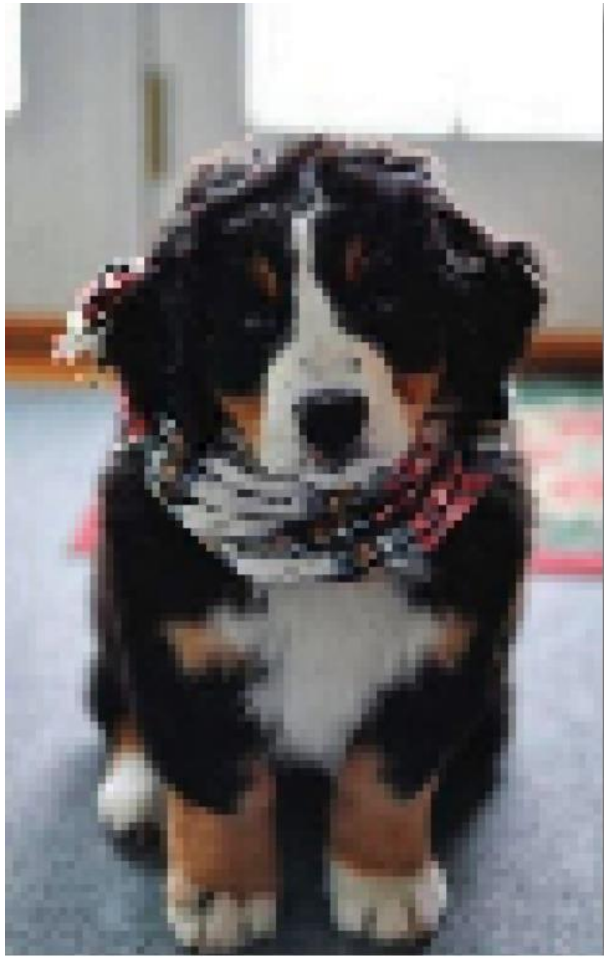
# How do I know if my sample is big enough?

828 data points?



# How do I know if my sample is big enough?

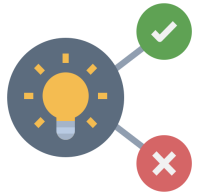
8,000 data points?…Hello, puppy!



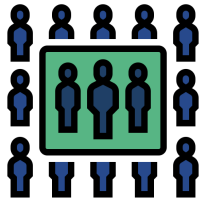
# How do I know if my sample is big enough?

237,777 data points?…too much data… but you're welcome!

# Key things to consider when designing an RCT



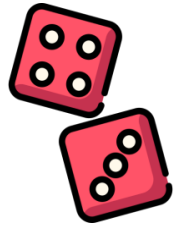
**Evaluation Question / Hypothesis**



**Sample size**

**Statistical power**

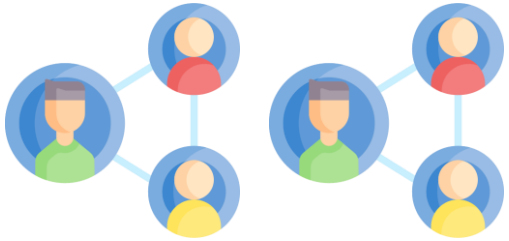
**Ethics**



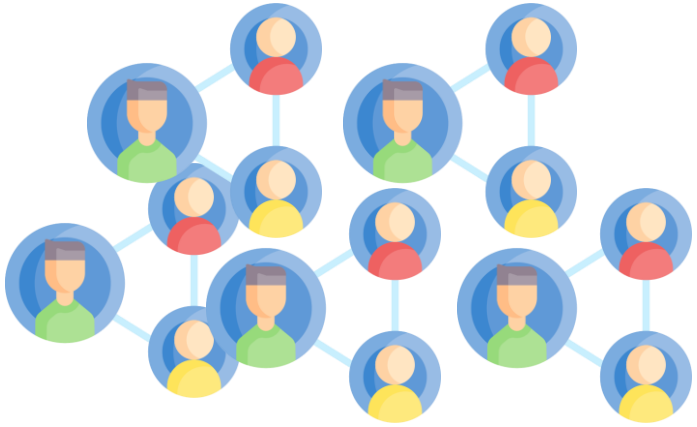
**Randomisation**

# How do I select my sample size?

Too little



The right amount



Too much

Power is the probability of detecting an effect *where one exists.*

# Power calculations have two adjustable parameters



**Statistical power**

- Usually 80%
- A 4 in 5 chance of finding a real effect, if it exists

**Significance threshold**

- Usually 5%
- A 5% likelihood that results we observe are due to chance

# Using power analysis to determine sample size

“What sample size do I need to determine my expected effect?”

You are sure of your expected effect size

You have a policy rationale for the effect size

Your output is the number of observations required per trial arm

**Activity: What would the effect size of interest be for our housing intervention?**

# Using power analysis to determine the minimum effect size

“What is the minimum effect I can detect?”

Your trial has practical constraints

Your output is the **minimum** effect size you can expect to detect

Power calculations indicate that the educational programme **can detect at least a 3%** impact on students’ grades.

# Using power analysis to determine the minimum effect size

Let's revisit our housing RCT where a programme to simplify planning applications is considered successful and *feasible to roll out in England* if it leads to 5% more houses built.

Imagine we can implement the programme in 50 local authorities.

We want to conduct a power analysis to understand:

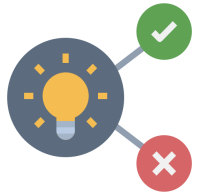
- **Will we be able to detect a 5% difference under these constraints?**
- **How many local authorities would we need to detect a 5% difference?**

Local authorities *on average* build 390 houses a year, so a 5% difference is 20 houses.

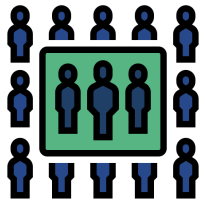
The variation of building rates across authorities is high, so we will put in a standard deviation of 100.

We need 80% power and use a 5% significance threshold.

# Key things to consider when designing an RCT



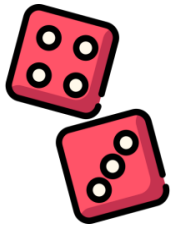
**Evaluation Question / Hypothesis**



**Sample size**

**Statistical power**

**Ethics**



**Randomisation**

# Is it ethical to randomise?

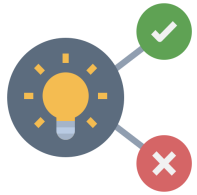
We want to evaluate because:

- We can be surprised by the results of well-intentioned programmes
- The cost of evaluation can be much lower than the cost of persisting with policies of uncertain / poor effectiveness

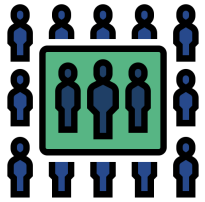
RCTs are the gold standard for evaluating and tend to be the most robust way to understand impact. In rare cases they may not be suitable.

**Case study: For the Coronavirus Job Protection programme, the team decided that it would not have been ethical to withhold the programme from some individuals to determine the impact during a time of high uncertainty.**

# Key things to consider when designing an RCT



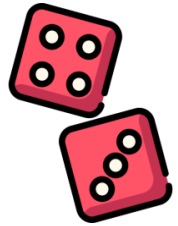
**Evaluation Question / Hypothesis**



**Sample size**

**Statistical power**

**Ethics**



**Randomisation**

# Individual randomisation

Allocate individual observations (e.g., a pupil, a taxpayer, a business).

✓ Benefits: Larger effective sample size

✗ Risks: Spill-over or contamination

**Example: Imagine we are delivering a parent engagement programme. We allocate children in the same family to both treatment and control groups (one sibling to treatment, the other to control). Are there any challenges to this?**

# Clustered randomisation

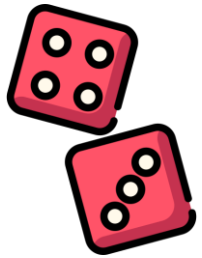
Allocate **clusters** of observations (e.g., a school or a local area).

✓ Benefits: Reduced risk of spill-over

✗ Risks: Smaller effective sample size

**Example: Imagine we decide to randomise our parent engagement programme across schools. All parents in a treatment school now receive the programme. Are there any challenges to this?**

# Group allocation



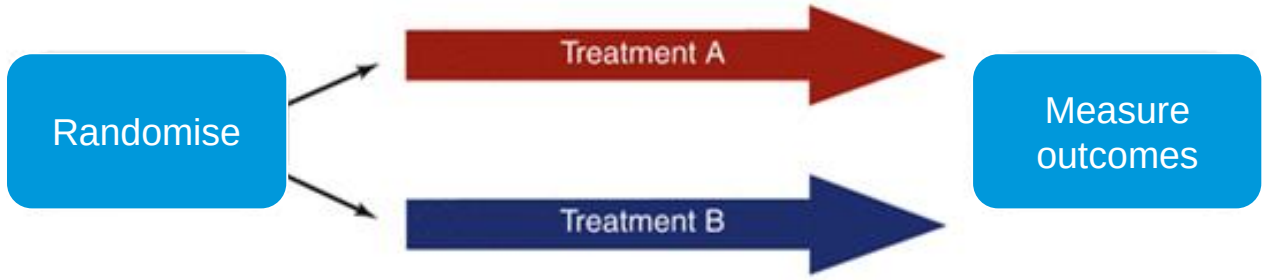
**Simple randomisation**: randomly assign units to groups

**Stratified randomisation:** set constraints on how many units you allocate to groups

# Types of RCT

# Parallel design

**Parallel design:** Some people experience a programme, others do not.

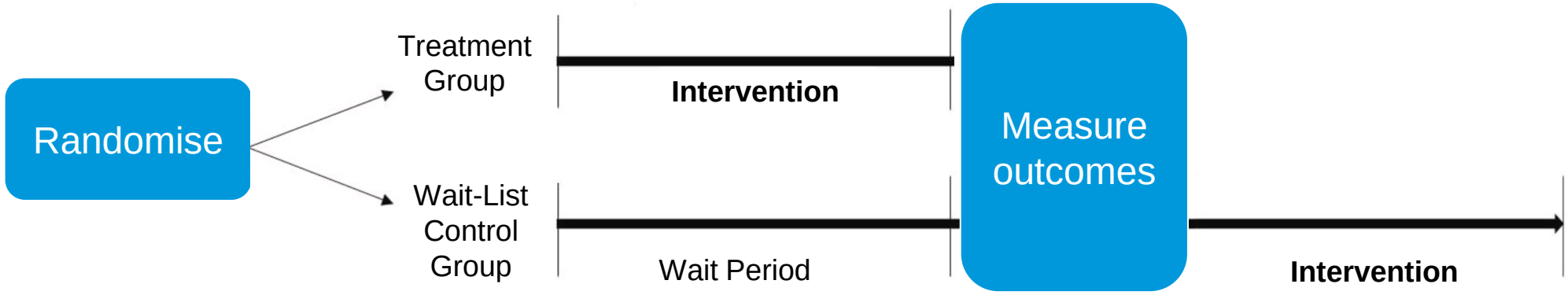



**Case study: A RCT run in the prison service was run using a parallel design where, as each individual entered the trial, they would join a block who would all be assigned to the same condition.**

**Those in the control condition were assigned a diary writing task for the last 4 weeks of their sentence; those in the treatment got interventions that target identity and self-perception, and a peer support offer.**

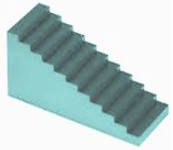
# Waitlist design

**Waitlist design:** Some participants receive the programme, others are on a waiting list. After the trial is over, the waitlist group are offered the programme.

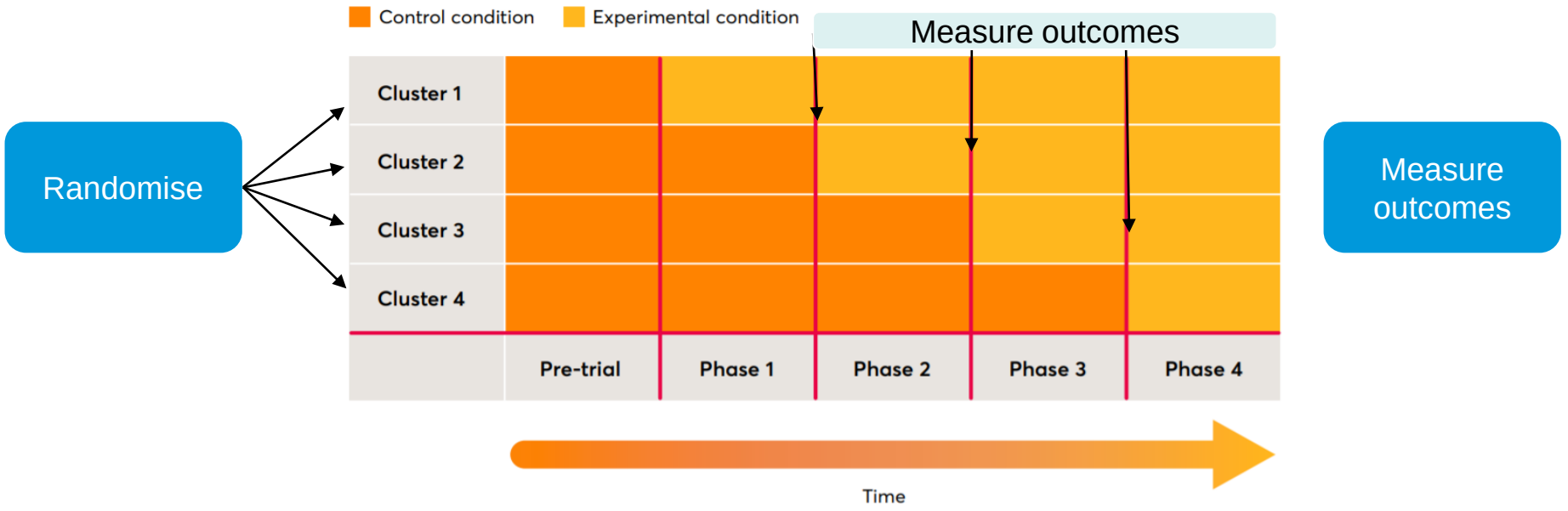



**Example: The Community-Based English Language (CBEL) programme, an RCT with over 500 participants, was run using a waitlist design where the control group were offered the intervention after the treatment group had completed it.**

# Stepped wedge design

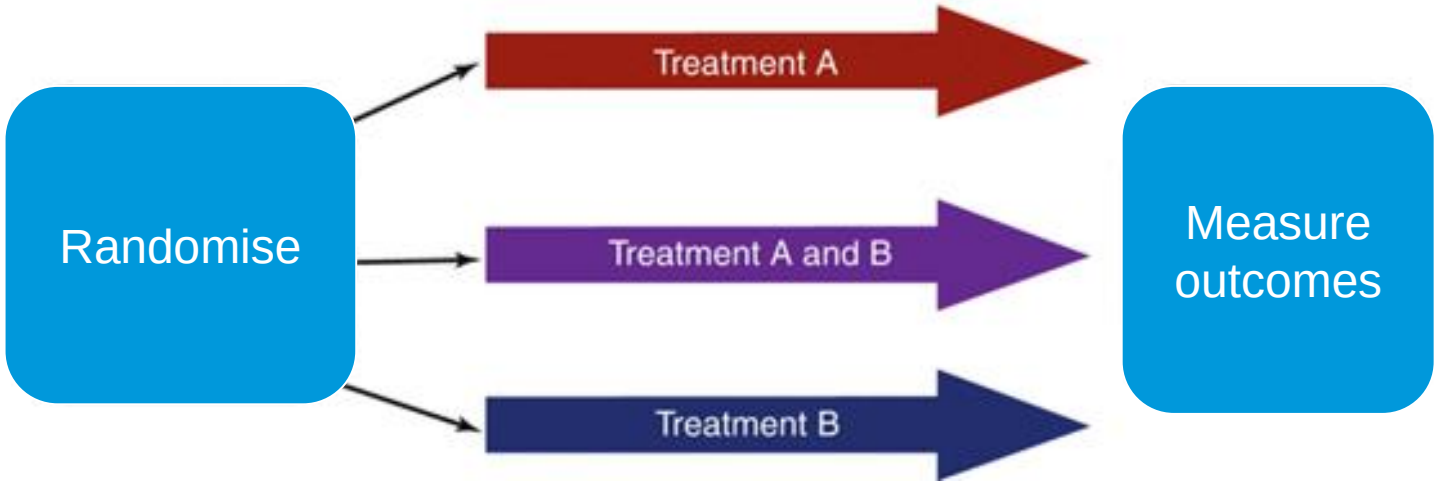


**Stepped wedge design:** The programme is gradually rolled out across clusters / participants. Everyone starts in the control; one by one participants receive treatment.



# Factorial designs

**Factorial designs:** Randomly allocate participants to receive a combination of different treatments.



# What experimental evaluation design would you use?

**Activity: What type of randomisation and RCT would be most appropriate for our housing example?**

- Cluster at the LA level
- Stratify to ensure that an equal number of deprived/non-deprived areas are in the treatment and control
- Possible designs
  - Parallel RCT at the LA level
  - Waitlist or Stepped wedge RCT rolling out the changes sequentially across LAs

# Comparing RCTs to other methods

# How RCTs fit alongside other methods

As with all impact evaluations, **RCTs benefit hugely** from a **Theory of Change (ToC)** detailing how the programme is intended to work

A **PE** or **Theory Based Evaluation** may be more appropriate if:

- You want to understand ***what*** is happening and ***how*** or ***why***

**The Education Endowment Foundation (a What Works Centre for Education) typically commissions Impact Evaluations *alongside* Process Evaluations.**

A **QED** or **Theory Based Evaluation** may be more appropriate if:

- You have **no time** to set up an RCT before programme delivery starts
- You are changing a policy that **affects the entire population** of interest
- It is **impossible to randomise**

**Beware of ruling out an RCT too quickly for cost or time reasons.** Consider the cost of an ineffective programme and the benefits of learning.

# Critically evaluating RCTs

# Critically evaluating the findings of RCTs

We trust RCT results when:

- The evaluation was **implemented as intended**
- We **randomised correctly** and **used good data**
- We had **enough participants** to make a valid claim

**Activity: What are some questions you could ask when reviewing the results of an RCT?**

# Questions to ask yourself

**Was randomisation successful?**

- Are groups comparable on descriptive statistics?
- Were balance checks carried out? Did they find differences between groups?

**How effective was the roll-out of the programme?**

- Did everyone who was meant to receive the programme receive it?
- Were spillover risks adequately addressed?
- How confident are we that the programme was administered as intended?

**Was there a pre-specified analysis plan?**

- Did the analysis adhere to the plan?  Were deviations justified?
- Does the analysis account for multiple comparisons?
- Was the sample size large enough to be powered?

# Activity

**Imagine you are reading the final report of the housing RCT. Discuss in groups how you would critically assess the following findings reported:**

- **Richer Local Authorities were more likely to be in the treatment group**
- **Many Local Authorities in the country have the same IT provider, so when they rolled out the new changes, they accidentally implemented the simpler planning rules in Local Authorities which were supposed to be in the control**
- **The main outcome measure changed from number of houses built to number of planning applications submitted with no explanation**
- **There were 6 main outcomes (whether apartments, detached houses, semi detached houses, terraced houses, duplexes, houses with garages were built) and the analysis didn't acknowledge this**

Imbalance

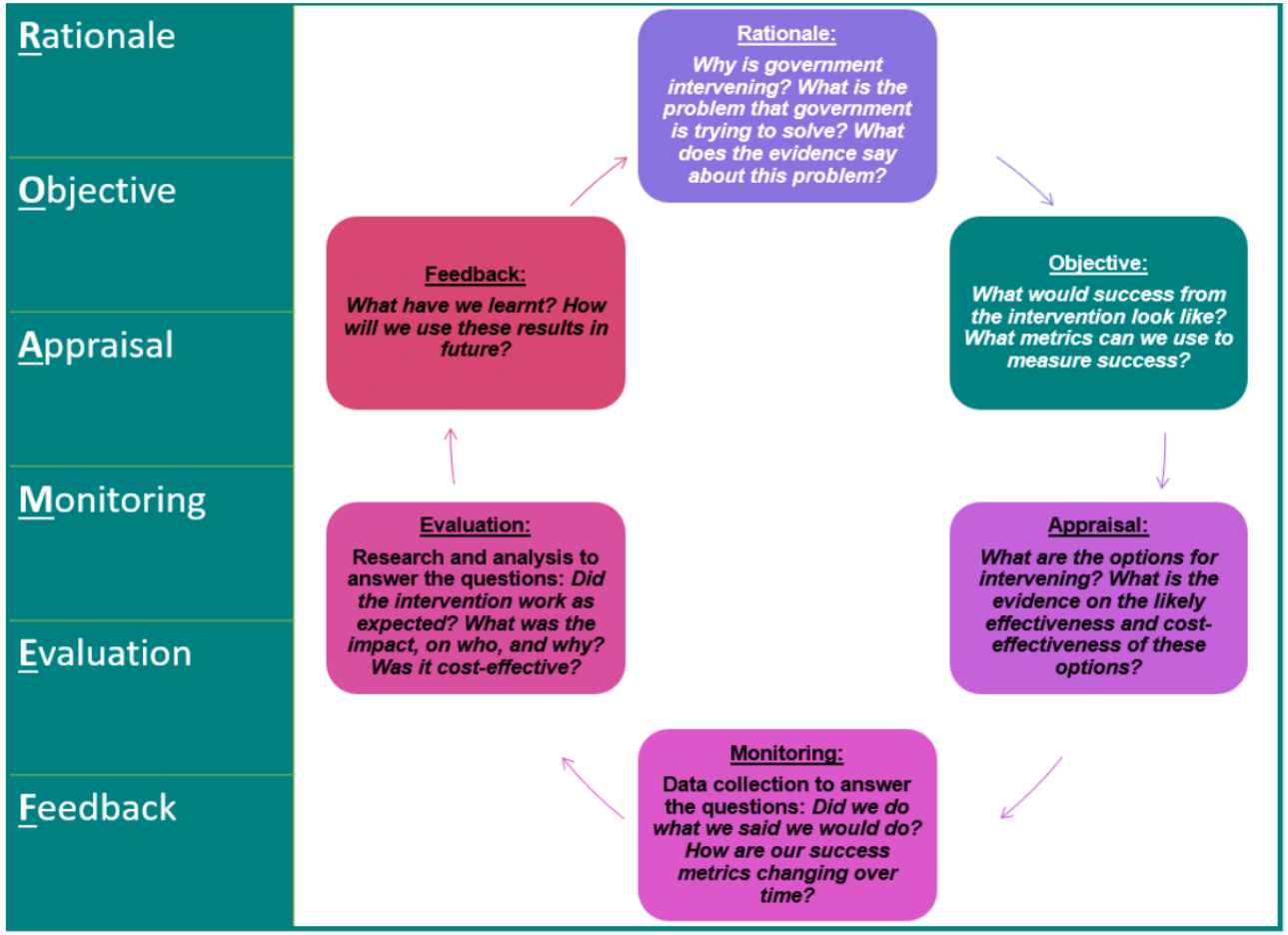
Spillover

Deviation from protocol

Didn’t correct for multiple comparisons

# Advocacy and application of learning

# Including experimental designs across the policy lifecycle




**Activity: How does what you have learned today fit into the ROAMEF cycle?**

- Think about an upcoming or current evaluation or policy you are involved in. How can you apply your learning from this module to influence that work?
- What barriers exist? How do you push through? What people or resources can support you?
- Write an intention for how you will use this in your work in the next 1-2 months.

# Summary

In this module, we have learnt:

- The role and value of key experimental evaluation designs and why we use them
- When we can use an experimental design vs a quasi-experimental design and the different types of RCTs such as staged, waitlist, parallel designs.
- The types of evaluation questions that experimental evaluations can or cannot answer.
- How experimental evaluations can work alongside other evaluation methods.
- The key things that need to be considered when preparing and running each experimental design, such as randomisation strategies, sample size, and analysis plan.
- The benefits and risks of different experimental evaluation designs.
- The most appropriate experimental evaluation method(s) based on the policy context.
- How to critically assess the findings of an experimental evaluation and be a stronger advocate for including experimental designs across the policy cycle.

# Further resources

| Resource |
| --- |
| Evaluation and Trial Advice Panel: Evaluation experts providing free advice to civil servants |
| The Magenta Book: Central Government guidance on evaluation |
| ETF: Further resources for evaluating policy in government (incl. Advanced RCT training resources) |
| BIT: TESTS |
| The Green Book |
| The Experimenter's Inventory |