Cabinet Office

Quasi-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

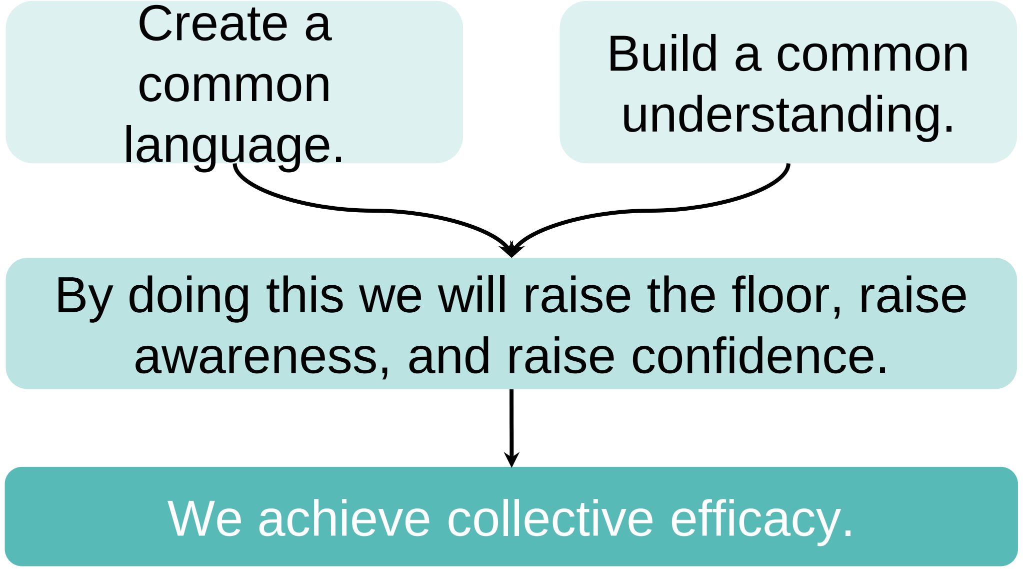
Module 7: Impact Evaluation - Theory-Based Designs

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.



Create a common language.

Build a common understanding.

By doing this we will raise the floor, raise awareness, and raise confidence.

We achieve collective efficacy.

Module 6: Overview of contents

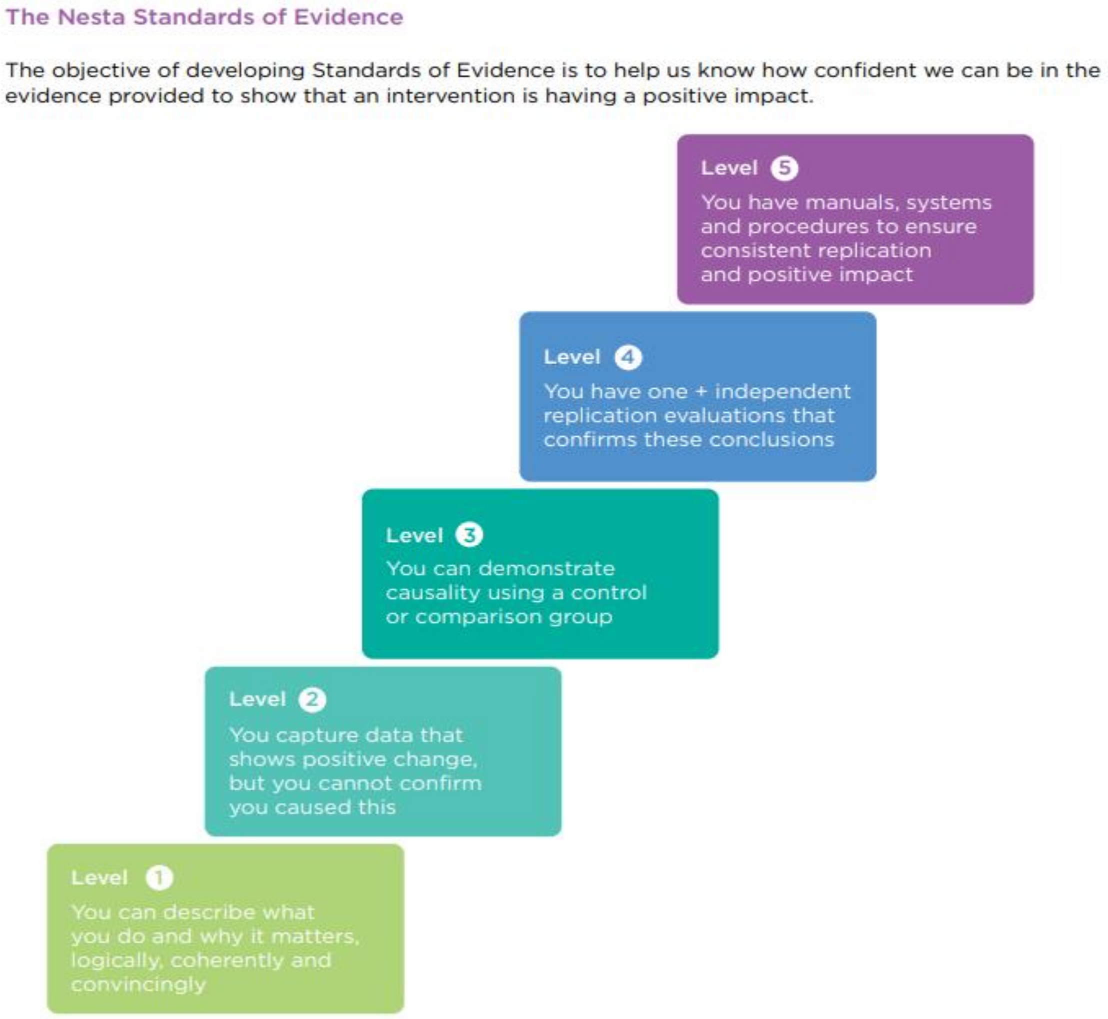
- Learning outcomes
- Hierarchy of evidence
- Introduction to QEDs
- Exploration of individual QEDs
- Choosing the right QED
- Advocacy and application of learning

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 quasi-experimental designs (in particular RDD, SC, DiD)
- I can identify the types of research questions that QEDs can or cannot answer
- I can explain the trade-offs and practical considerations when deciding between experimental and quasi-experimental methods
- I can explain the key things that need to be considered when preparing and running each QED design
- I can explain the benefits and risks of different QED designs
- I can contrast the most appropriate QED method(s) to use based on the policy context
- I can critically assess the findings of a QED evaluation
- I can advocate for including QEDs 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.

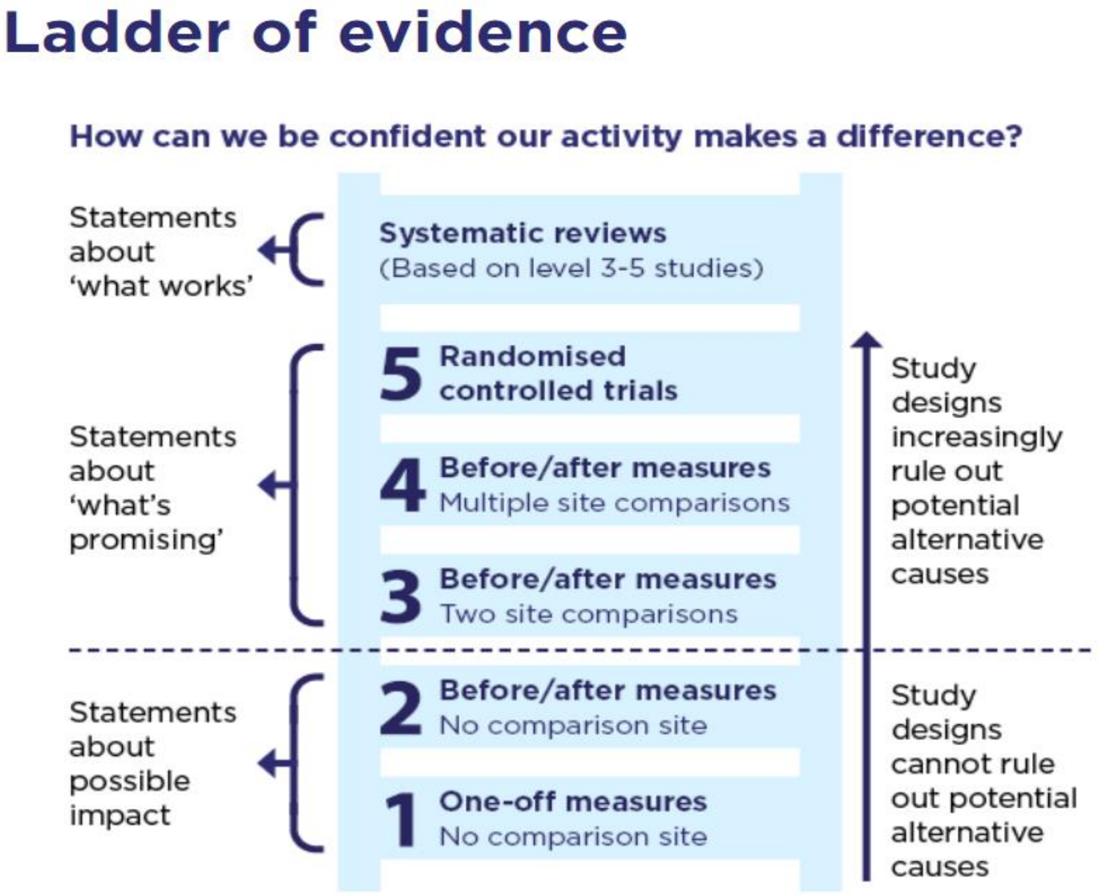
Level 5
You have manuals, systems and procedures to ensure consistent replication and positive impact

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

Study designs increasingly rule out potential alternative causes

Statements about possible impact
2 Before/after measures
No comparison site
1 One-off measures
No comparison site

Study designs cannot rule out potential alternative causes

Introduction to QEDs

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

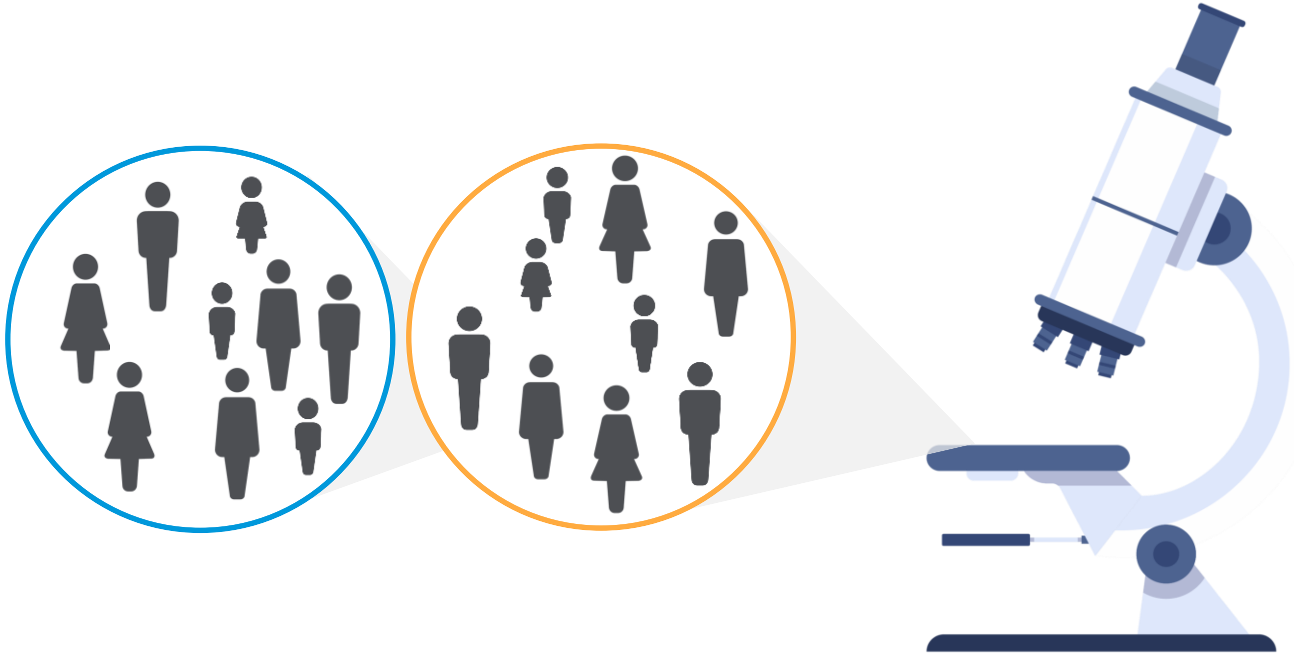
- × Can’t explain *why* a programme does or doesn’t work

What are quasi-experimental designs (QEDs)?

There are a variety of QEDs, can you name any?

We will cover these 5 methods (*note: there are more QEDs beyond these*)

- Regression discontinuity (RDD)
- Difference-in-differences (DiD)
- Synthetic control
- Matching
- Pre-post



Why would you run a QED instead of an RCT?

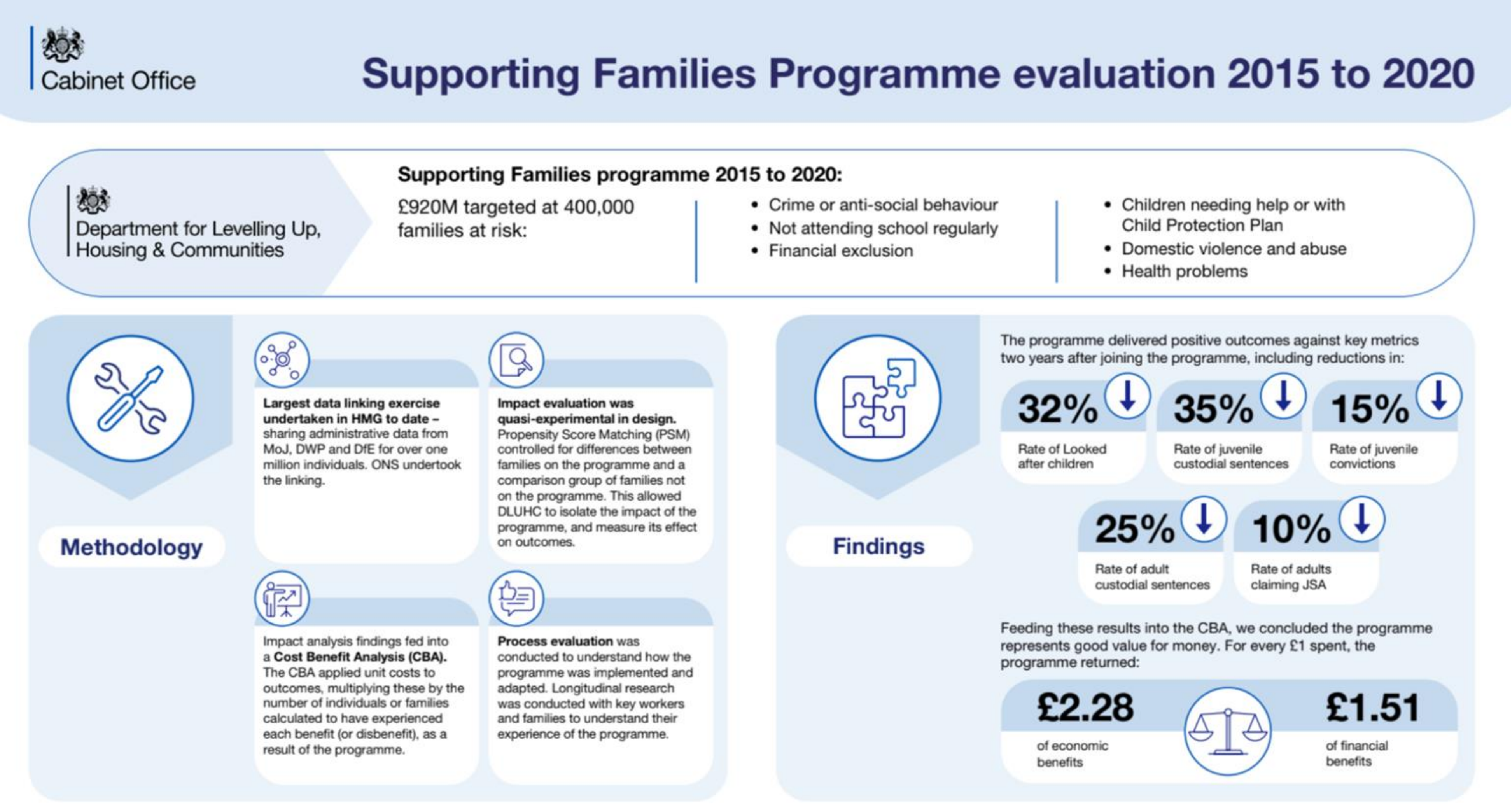
The programme has already been implemented

Ethical reasons

Practical reasons

Political reasons

Example of a QED (using Propensity Score Matching)



Cabinet Office

Supporting Families Programme evaluation 2015 to 2020

Department for Levelling Up, Housing & Communities

Supporting Families programme 2015 to 2020:

£920M targeted at 400,000 families at risk:

- Crime or anti-social behaviour
- Not attending school regularly
- Financial exclusion
- Children needing help or with Child Protection Plan
- Domestic violence and abuse
- Health problems

Methodology

Largest data linking exercise undertaken in HMG to date – sharing administrative data from MoJ, DWP and DfE for over one million individuals. ONS undertook the linking.

Impact evaluation was quasi-experimental in design. Propensity Score Matching (PSM) controlled for differences between families on the programme and a comparison group of families not on the programme. This allowed DLUHC to isolate the impact of the programme, and measure its effect on outcomes.

Impact analysis findings fed into a **Cost Benefit Analysis (CBA).** The CBA applied unit costs to outcomes, multiplying these by the number of individuals or families calculated to have experienced each benefit (or disbenefit), as a result of the programme.

Process evaluation was conducted to understand how the programme was implemented and adapted. Longitudinal research was conducted with key workers and families to understand their experience of the programme.

Findings

The programme delivered positive outcomes against key metrics two years after joining the programme, including reductions in:

- 32% ↓ Rate of Looked after children
- 35% ↓ Rate of juvenile custodial sentences
- 15% ↓ Rate of juvenile convictions
- 25% ↓ Rate of adult custodial sentences
- 10% ↓ Rate of adults claiming JSA

Feeding these results into the CBA, we concluded the programme represents good value for money. For every £1 spent, the programme returned:

£2.28 of economic benefits

£1.51 of financial benefits

What are the drawbacks of QEDs compared to RCTs?

RIGOUR

Require extra assumptions to make causal claims

ANALYSIS

Require specialised skills

DATA COLLECTION

Require additional data to validate assumptions

When is an RCT not possible?

Activity: List some reasons why you might not be able to run an RCT in your context.

Possible answers

You are changing a policy that affects the entire population of interest. E.g,:

- You change the age students can leave school from 16 to 18, this will affect all pupils
- You are implementing a new railway

There are ethical questions about treating people differently (and a wait-list RCT, where you give the comparator group the programme at the end of the data collection period, is not feasible).

The policy has already been implemented!

Exploration of individual QEDs

Exploration of individual QEDs

5 Types of QED

1. Regression discontinuity
2. Difference-in-differences
3. Synthetic control
4. Matching
5. Pre-post

5 questions for each QED

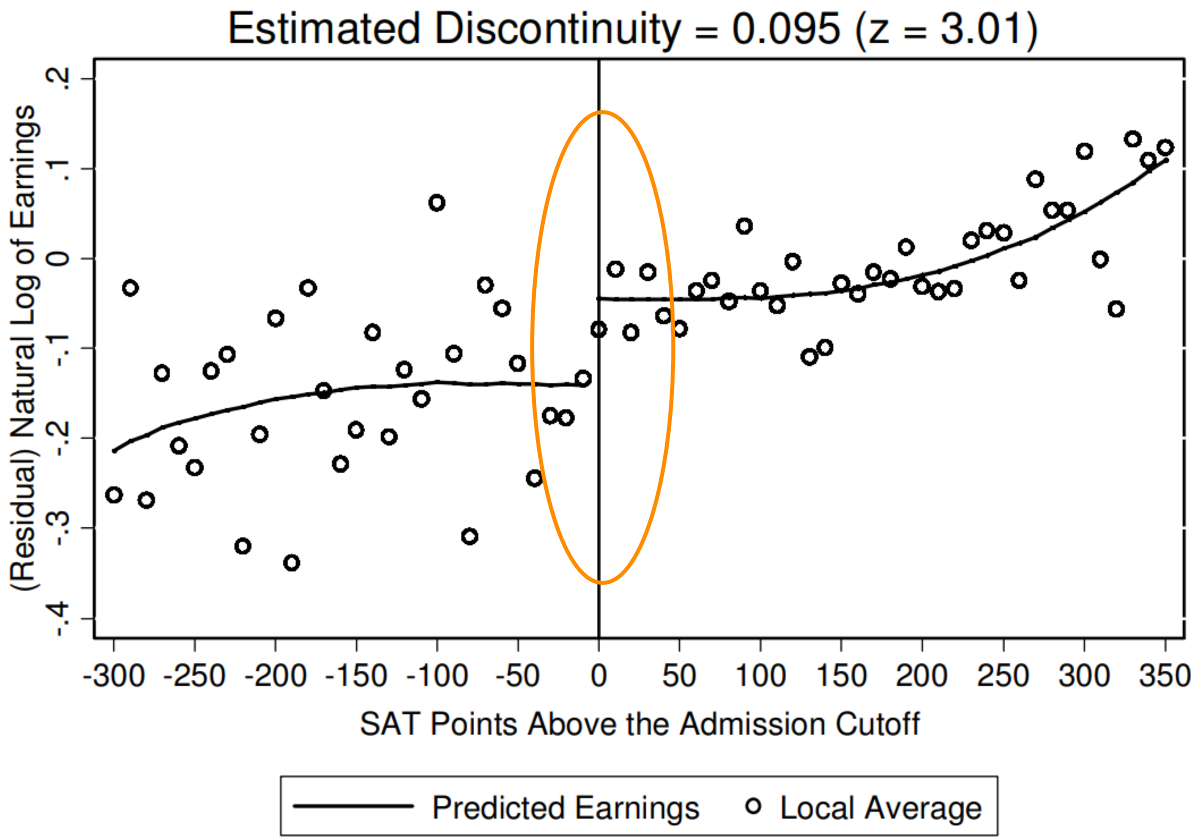
1. What is this method?
2. When should you use this method?
3. When should you not use this method?
4. What critical choices do evaluators make with this method?
5. What are the key limitations of this method?

Regression Discontinuity Design

What is a regression discontinuity design (RDD)?

RDDs are used when there is a **threshold** that can be exploited to study differences between groups.

Whether an individual is just **above** or **below** this threshold is close to **random**.

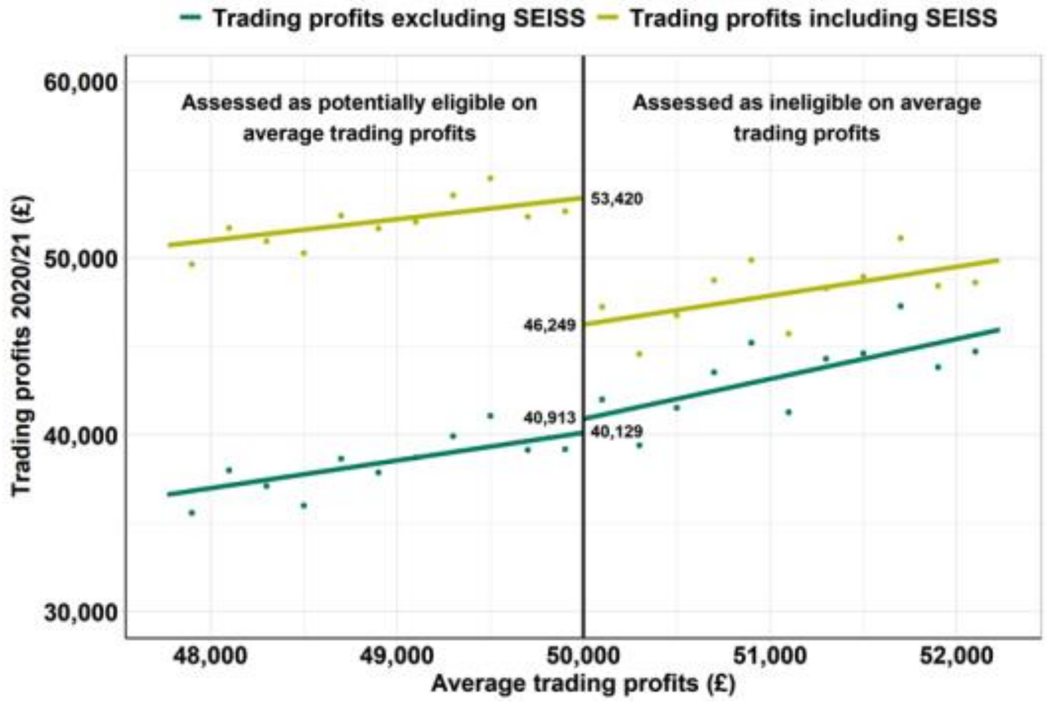


Case study: RDD

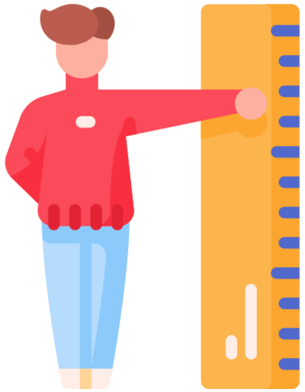
In the case of the COVID SEISS (Self Employment Income Support Scheme), **a RDD approach was used:**

> *“An RDD estimates the impact of an intervention by using a discontinuity in the probability of treatment (the probability of having claimed the SEISS).*
>
> *A discontinuity occurs because individuals with less than £50,000 average trading profits often chose to claim the SEISS, whereas those above this threshold rarely claimed it.”*

Figure 5.5a: Regression discontinuity design of average trading profits against trading profits in 2020 to 2021

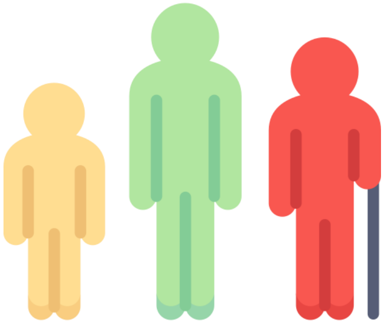


When should you use an RDD design?

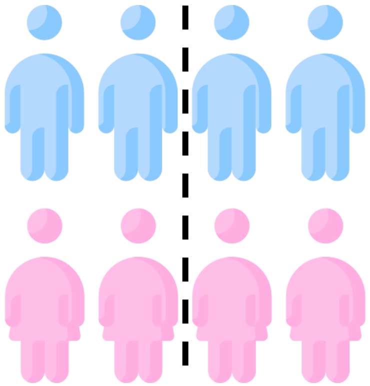


There is a natural cutoff

You know and can access the running variable



You have access to other important variables



You can identify lots of people around the cutoff

When should you use an RDD design?

Example: We want to estimate the impact of a training programme on future earnings.

- Admission to the programme is conditional on getting a test score of 50/100.
- Unsuccessful applicants with scores just below the threshold (like 48 or 49) are likely to be very similar to successful applicants who just made it (by scoring 50 or 51).

Activity: What else needs to be in place, in addition to the above information, to make this a good candidate for RDD?

1. We have a ‘sizeable’ group of applicants who scored around 50 (instead of a very polarised scenario of many 10s and 90s)
2. We can access information on test taker characteristics like their earnings when taking the test, gender, age

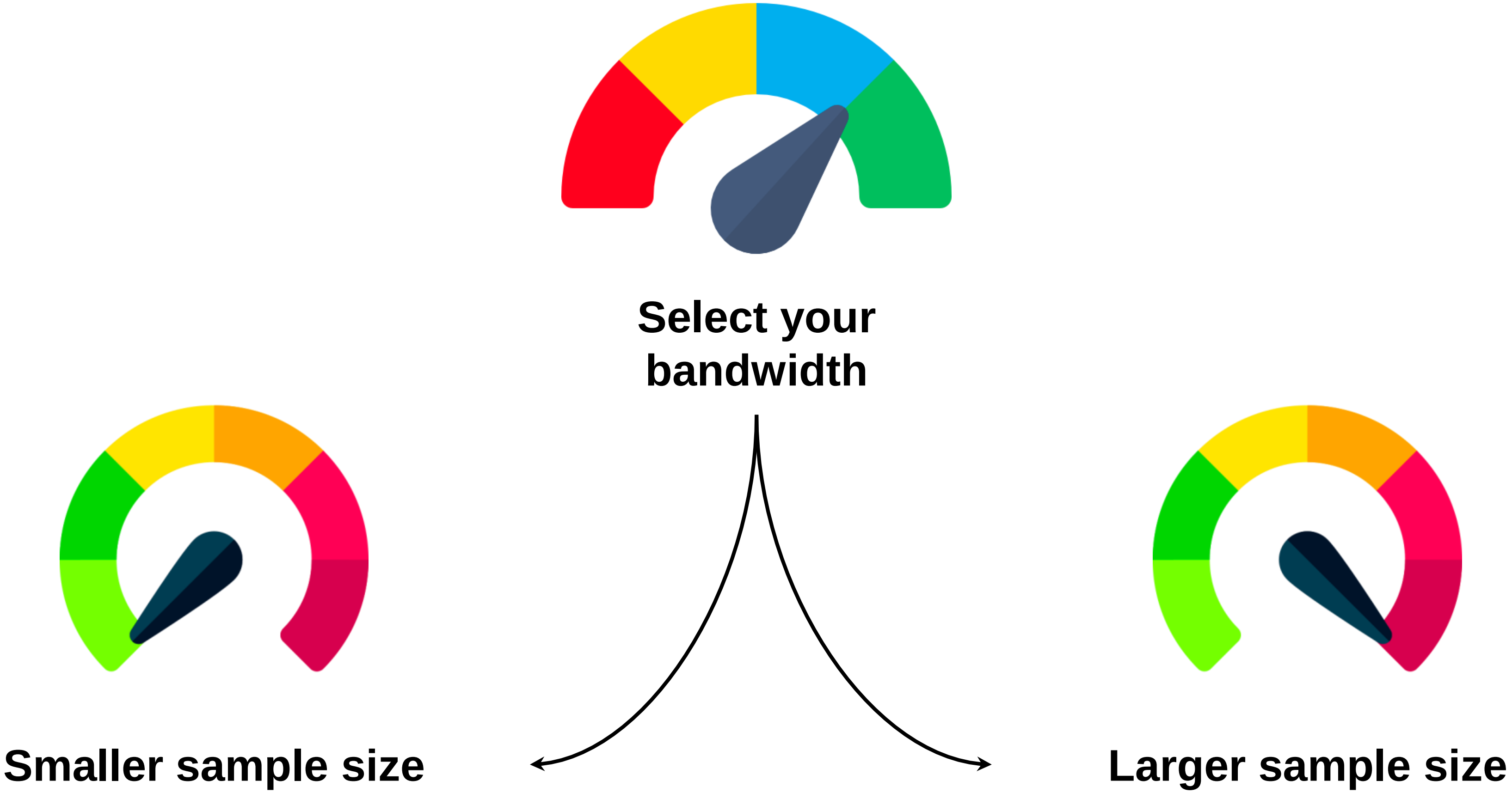
When should you not use RDDs?

The running variable can be manipulated

The cutoff triggers multiple interventions

You care about the effect for all treated individuals

What critical choices do evaluators make with RDD?



Select your bandwidth

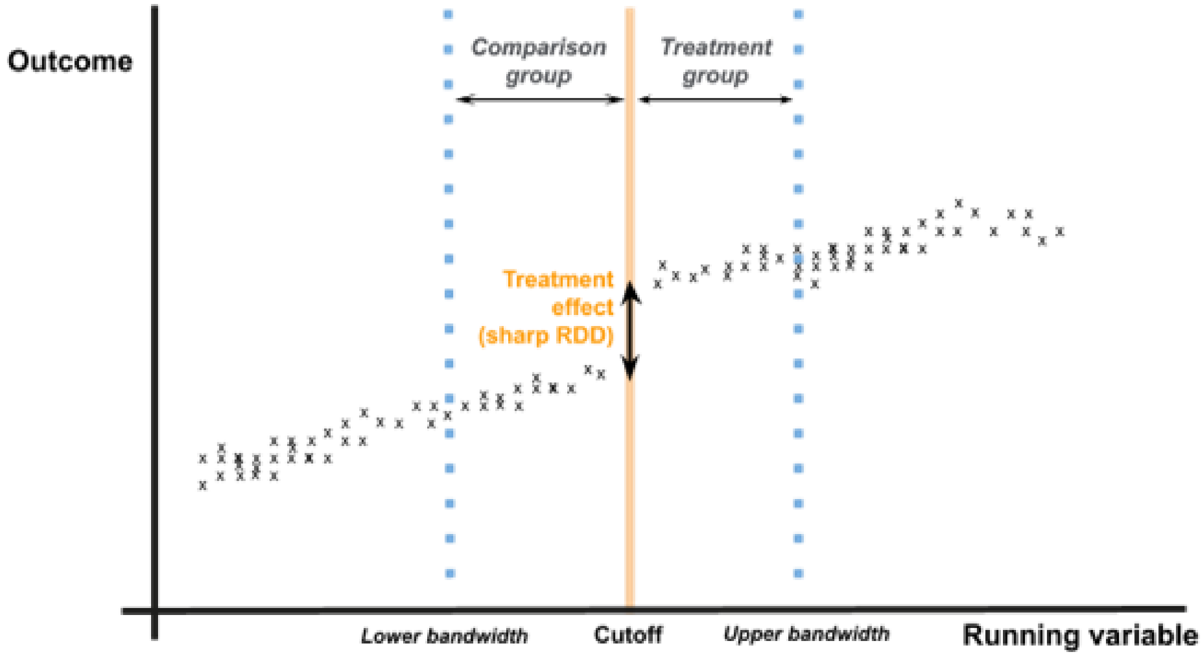
Smaller sample size

Larger sample size

What are the key limitations of an RDD?

Only estimate the effect of the programme for individuals close to the cutoff…

…therefore you can't understand the effect on those further away from it.

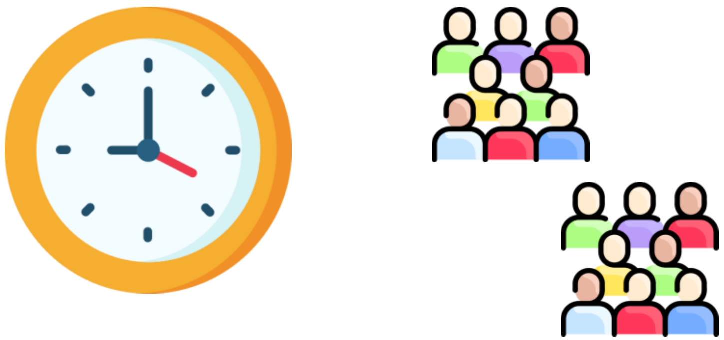


Difference-in-differences

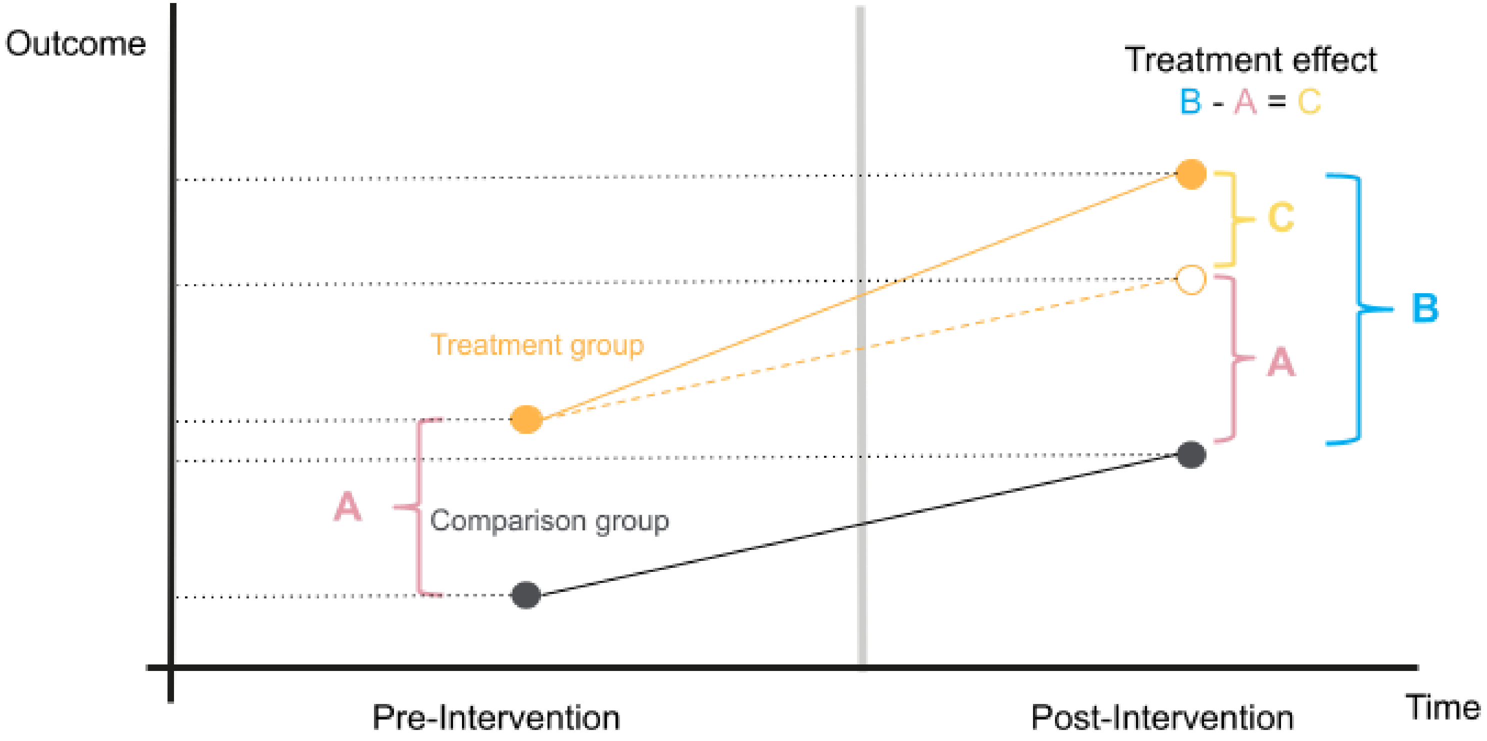
What is a difference-in-differences (DiD) design?

Time

Group



It takes the difference in the outcome before and after it was implemented for the treatment group and subtracting this same time difference for a comparator group



When should you use a DiD design?

The two groups have **parallel trends**, and could credibly have followed the same trend **without the intervention.**

Pre-intervention, comparator closely tracks treatment

The programme assignment is based on **eligibility**

No change in behaviour in anticipation of the intervention

When should you use a DiD design?

Example: We want to understand the effect of compulsory schooling laws on the years of schooling children obtain.

England has changed its schooling law (compulsory until 18). Wales has not (compulsory until 16). Here, England is the treatment group, Wales is the control group. 2021 is a pre period and 2022 is the post period.

What else needs to be in place, in addition to the above information, to make this a good candidate for DiD?

1. **Parallel trends**.
2. **No anticipation effects** Parents *could* anticipate the change and move to Wales, so you would want to test for this.
3. **Spillover caution:** There is nothing preventing people living in one country and attending school in another. Whether this matters will depend on exactly how our data are collected.

When should you not use a DiD design?

The two groups don’t follow the same trend before the intervention

Small sample size

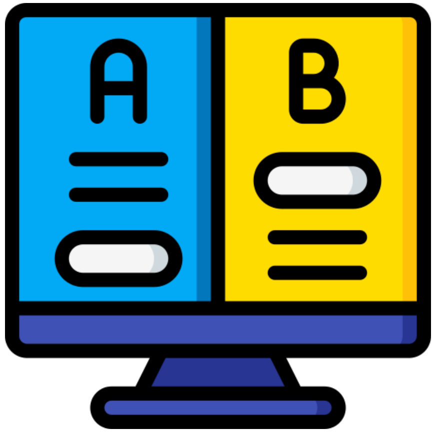
Participants can anticipate the change before it happens

Case study: DfE conducted a QED evaluation of funding for T-level placements. The DiD had two main challenges:

1. Small sample size: Only ten providers were funded meaning that the sample size was very small.
2. The parallel trends assumption was violated.

As a result, the team had to exercise caution with the results of the DiD.

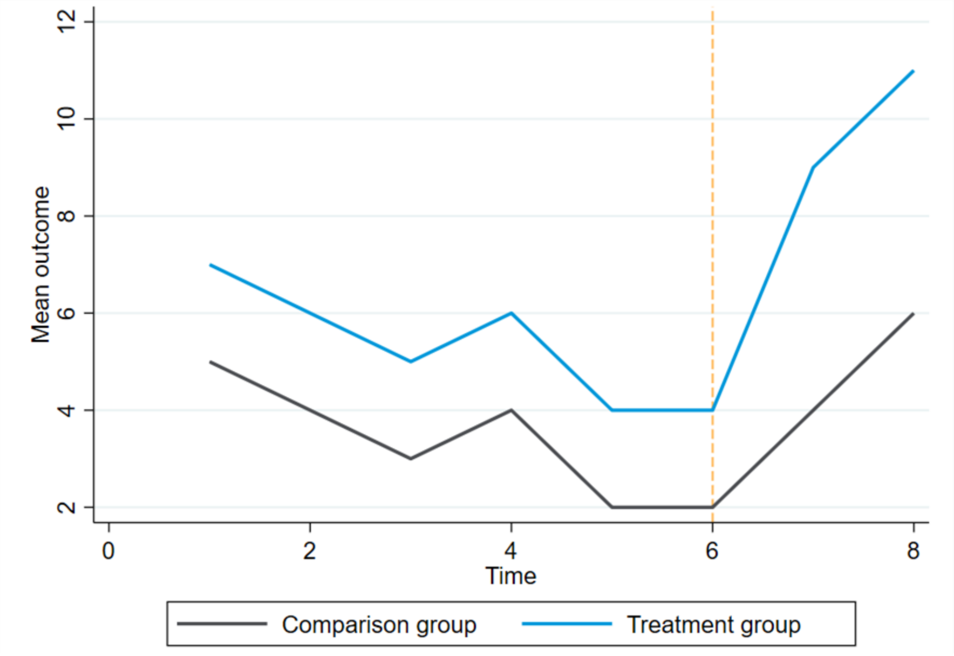
What critical choices do evaluators make with a DiD?



Selecting an appropriate comparison group

What are the key limitations of using DiD?

Estimates are based on the parallel trends assumption

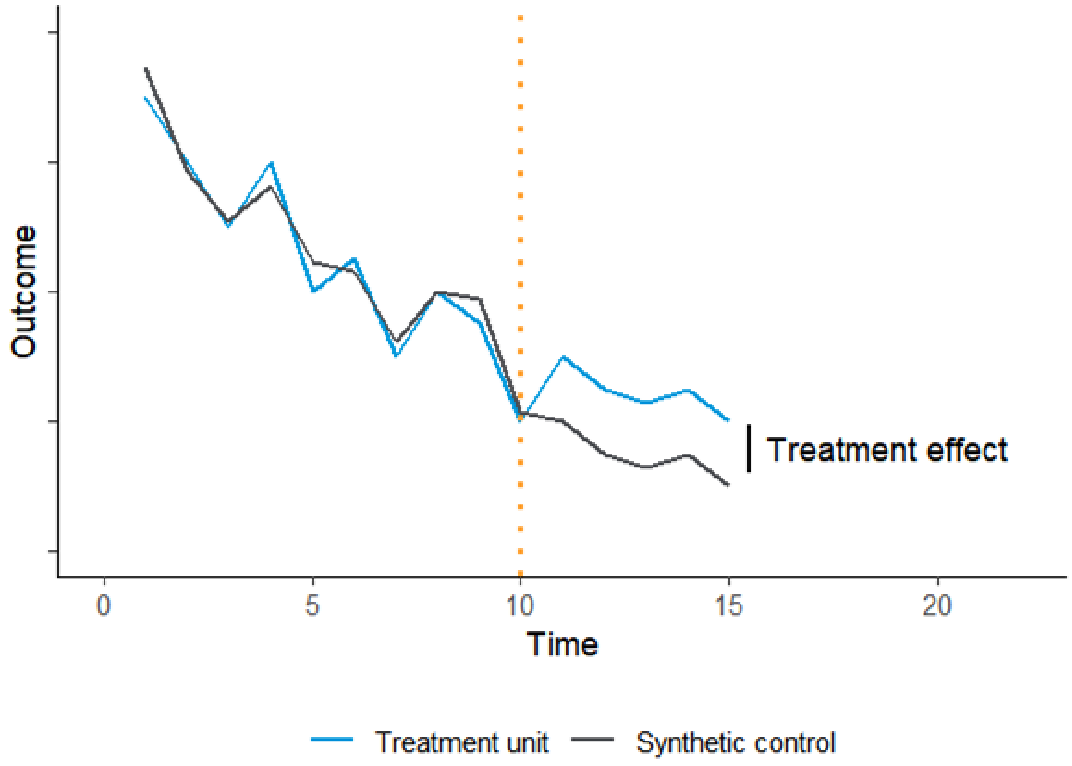


Synthetic control method

What is a synthetic control method (SCM)?

A synthetic control evaluates interventions implemented at an **aggregate level** in a **small number of units**.

It compares post-intervention outcomes of the treated unit to those of a ‘fictional’ (synthetic) control unit, created by an algorithm.



Intervention started in period 11

When should you use a SCM design?

One (or few) treated unit(s) and multiple untreated units

Many pre-intervention outcome measurements

Past outcomes strongly predict future outcomes

Pre-intervention, control closely tracks treatment

No change in behaviour in anticipation of the intervention

When should you use a SCM design?

Example: We want to understand the effect of a new public transport discount scheme, implemented in only one English county, on transport ridership.

Fictional control = data from all *other* counties in England.
Units = English counties (donor pool)

What else needs to be in place, in addition to the above information, to make this a good candidate for SCM?

1. Past public transport ridership in English counties is strongly predicted by previous levels of ridership
2. Past public transport ridership moves in a similar way in time in the intervention county and in the synthetic control county

When should you not use a SCM design?

You expect your programme to have smaller effects

Your evaluation team does not have any prior experience of running SCMs

What critical choices do evaluators make when using SCM?

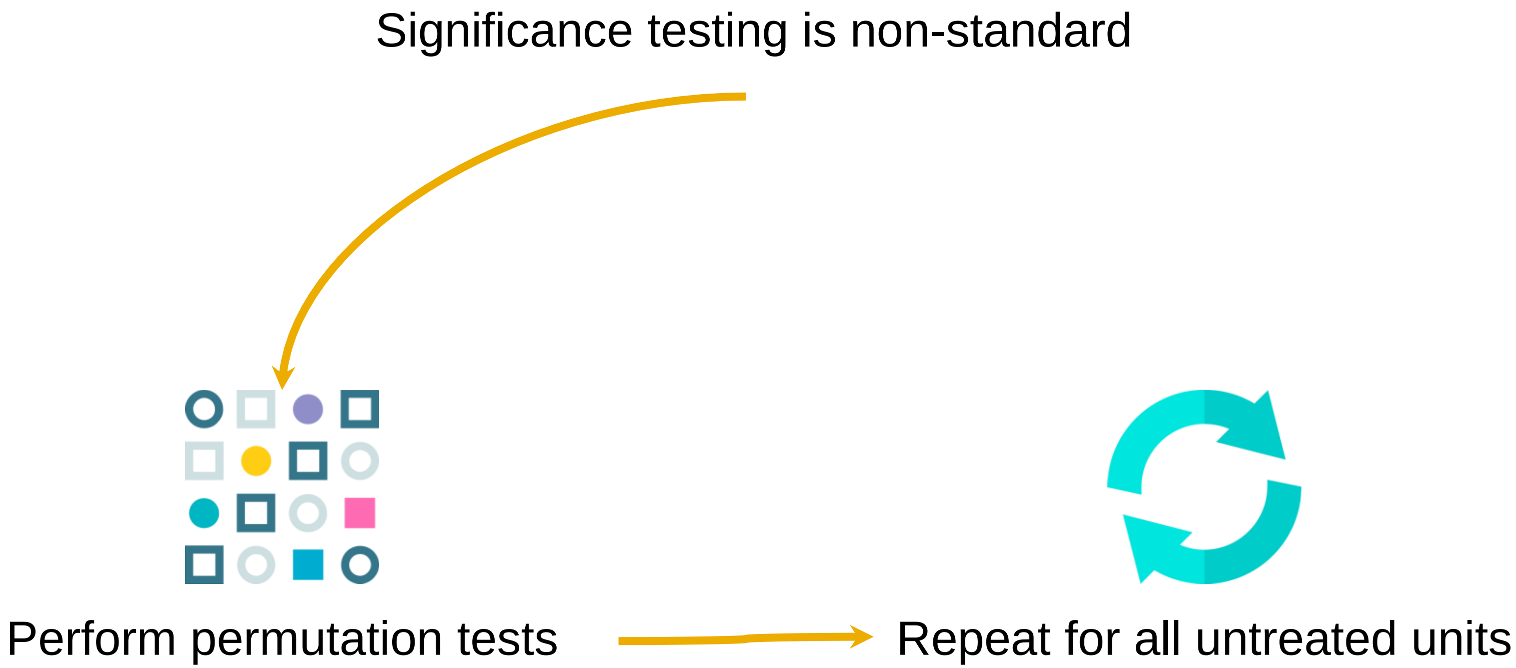
Choose your donor pool carefully

10 → 50

Use 10-50 units in the donor pool **per treated unit**

Test whether findings are robust to different predictors

What are the key limitations of using SCM?



Significance testing is non-standard

Perform permutation tests

Repeat for all untreated units

Cross sectional design with matching estimator

What is a cross sectional design with matching estimator?

Matching methods compare two groups after a programme has occurred. They try to **maximise** the **comparability** of individuals in the treatment and comparator group.

They apply **weights** to individuals in the data based on **observable characteristics**

When should you use a matching design?

Only have post-intervention outcome data

Can observe characteristics which are important to both outcomes and treatment status

Treatment status is not dictated by personal choice

When should you use a matching design?

Example: we want to estimate the impact of a training programme on employment.

Matching analysis compares:

The individuals enrolled with other similar individuals

Based on data on their employment characteristics and other demographics

In the same local area who didn’t enrol

Imagine you are explaining this evaluation to a Minister or Deputy Director. What is or is not convincing about it?

When should you not use a matching design?

Whenever you can use a more robust method!

Have pre-treatment data for the treatment and comparator group? → Use a DiD design.

What critical choices do evaluators make with a matching design?

Choosing a matching method.

Entropy balancing (recommended)

Propensity score matching (PSM)

Coarsened exact matching (CEM)

What are the key limitations of a matching design?

Less convincing: Important for your advocacy of evaluation in policy design.

Scenarios for perfect matching are **rare**

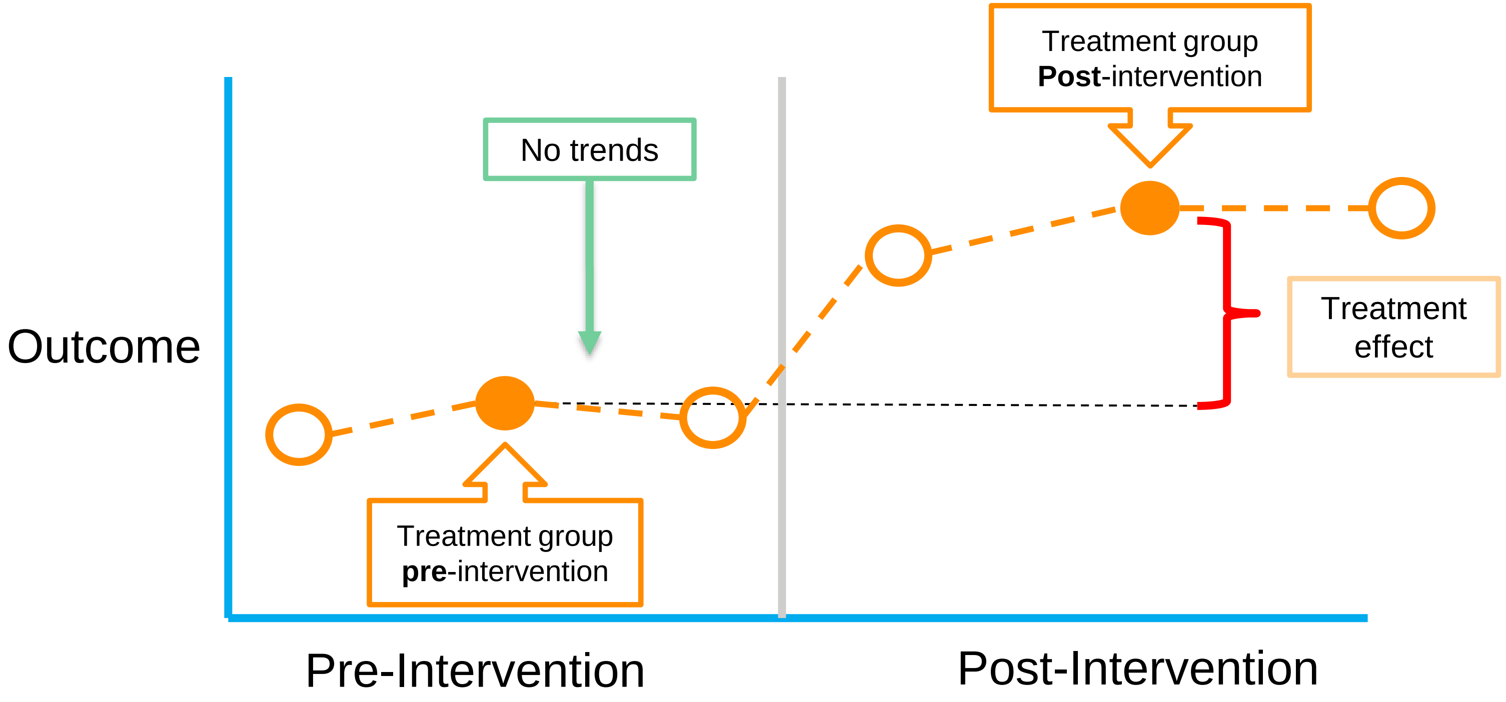
Can overstate impact

Pre-post

What is a pre-post design?

Pre-post analyses estimate the effect of an intervention by comparing outcomes within a treatment group **before** and **after** it is implemented.

The impact of an intervention is the **change in the outcome** between the pre and the post measure.



When should you consider using a pre-post design?

An intervention is applied to an entire population

Data for a comparator group is impossible to get

There is a short measurement window and a sharp change in outcomes

When should you consider using a pre-post design?

Example: Let’s imagine that you have been asked to evaluate the impact of reducing the length of work shifts in a factory from 10 hours to 8 hours on productivity.

What needs to be in place to make this a good candidate for pre-post?

1. You can measure productivity daily
2. The change is sharp - it affects the outcome straight away from day 1

When should you not use a pre-post design?

There are underlying time trends in the outcome

There are concurrent events that might affect the outcome

There are changes in unobservable characteristics over time

Participants can anticipate the change before it happens

What are the key limitations of a pre-post design?

Less convincing

Scenarios for a **robust** pre-post are **rare**

Choosing the right QED

What specific QED should I use?

Aim for the QED with the highest robustness, **unless**:

- The design is not possible
- It is not practical to get data
- You have very few observations

Design	Robustness rank (1-3, 1 = highest)	Minimum data requirements	Power (i.e., sample size required to detect an effect)
Regression discontinuity	1	● Post-outcomes for T + C groups ● ‘Running variable’ based on which treatment is assigned	Much worse than individual-level RCT
Difference-in-differences	2	● Pre- and post- outcomes for T + C groups	Lower than RCT
Synthetic control	2.5	● Pre- and post- outcomes for treated unit and donor pool	Much worse than individual-level analysis
Matching	3	● Post-outcomes for T + C groups	Comparable to RCT
Pre-post	3	● Pre- and post- outcomes for T group only	Similar to RCT if pre and post are two separate arms Higher than RCT if same individuals are in pre and post

Further details on each method and the advantages/disadvantages of each can be found in the Magenta Book here.

Activity: What QED would work best for…

Example: Imagine you are working with a local authority (LA) who rolled out a new type of Covid testing procedure in 2021. They now want to understand its impact on the number of new Covid cases in the LA - to see if this approach should be used in future pandemics.

What QED would work best for…

Goal: Local authority (LA) want to understand impact of a new Covid testing procedure on the number of new Covid cases in the LA. Should this approach should be used in future?

Scenario 1: You have access to aggregated weekly local authority level case rates from DHSC from 2020 to 2022. You have this data for both the treated LA and every other LA in England. You know that case rates within areas have remained stable over time.

	Week 1	**Week 2**	**Week 50**	**Week 100**
Treated LA	200	204	202	210
Untreated LA 1	300	302	295	298
Untreated LA 2	650	645	655	647

Synthetic control. Why?

- Systemic intervention affects everyone in the area
- Only one treatment unit
- High-frequency, long-term data plus comparator
- Pre-programme measurements

What QED would work best for…

Goal: Local authority (LA) want to understand impact of a new Covid testing procedure on the number of new Covid cases in the LA. Should this approach should be used in future?

Scenario 2: You have access to a database of all Covid tests taken and their results across England. You can match this to a HMRC dataset of all employed people in England. This means that for every person in employment in England from 2020-2022, you know where they live and their testing history.

	Individual	Tests May 2020	Outcome May 2020	Tests July 2021	Outcome July 2021	Tests July 2022	Outcome July 2021
Treated LA	1	Yes	Positive	No	-	No	-
Treated LA	2	No	-	Yes	Negative	Yes	Positive
Untreated LA	3	No	-	Yes	Positive	No	-
Untreated LA.	4	Yes	Negative	Yes	Positive	No	-

Difference in Differences.

- Intervention is systemic, and affects everyone in the area, but we can identify control areas and have pre/post data
- We have a lot of treatment units
- Data allows us to check for parallel trends

What QED would work best for…

Goal: Local authority (LA) want to understand impact of a new Covid testing procedure on the number of new Covid cases in the LA. Should this approach should be used in future?

Scenario 3: You have access to a database of all Covid tests taken and their results across your local area only. You can match this to a HMRC dataset of all employed people in your local area. This means that for every person in employment in your local area 2020-2022, you know their testing history.

	Individual	Tests May 2020	Outcome May 2020	Tests July 2021	Outcome July 2021	Tests July 2022	Outcome July 2022
Treated LA	1	Yes	Positive	No	-	No	-
Treated LA	2	No	-	Yes	Negative	Yes	Positive

Pre-post

- Intervention is systemic and affects everyone in the area
- Can access individual level data
- Only have data for our treatment group. No comparison area data available

What QED would work best for…

Example B: Imagine you want to understand the impact of a new regulation across the 6,904 electoral wards in England.

The regulation allows wards that had 25% or more green areas (like a park or AONB) in 2021 to add an extra storey to existing units without requiring planning permission.

You want to know if this regulation has increased the number of housing units in eligible wards.

This is a one-year only regulation, and it was announced on December 31st 2021.

What QED would work best for…

Goal: Understand the impact of a new regulation (allows wards that had 25% or more green areas in 2021 to add an extra storey to existing units without requiring planning permission). Has this regulation increased the number of housing units in wards? One-year only regulation, announced on Dec 31st 2021.

Scenario 1: You know the percent of green space in each ward in 2021, the eligibility criteria for the new regulation. You have data on the number of units built across each ward in 2021 (year before the introduction of the policy) and 2022 (the year after the introduction of the policy).

	2021 Green Space	2021 Built	2022 Built
Ward A	24%	374	400
Ward B	27%	377	450
Ward C	22%	355	360
Ward D	30%	380	405
Ward E	24%	390	395

RDD:

- People couldn't anticipate the change
- There is a discrete threshold
- Many wards with a % of green space close to the threshold
- We know the % of green space in each ward in 2021

What QED would work best for…

Goal: Understand the impact of a new regulation (allows wards that had 25% or more green areas in 2021 to add an extra storey to existing units without requiring planning permission). Has this regulation increased the number of housing units in wards? One-year only regulation, announced on Dec 31st 2021.

Scenario 2: You have data on the number of units built across each ward in 2020 and 2021 (year before the introduction of the policy) and 2022 (the year after the introduction of the policy). You do not know the % of green space in each ward in 2021, but you do know which wards had more or less than 25% of green space in 2021.

	2021 Green Space Threshold	2021 Built	2022 Built
Ward A	Below	374	400
Ward B	Above	377	450
Ward C	Below	355	360
Ward D	Above	380	405
Ward E	Below	390	395

Difference in Difference

- A change that people couldn’t anticipate
- Based on clear eligibility conditions
- Many treatment units
- Enough pre-treatment data on the outcome to test for the parallel trend assumption

What QED would work best for…

Goal: Understand the impact of a new regulation (allows wards that had 25% or more green areas in 2021 to add an extra storey to existing units without requiring planning permission). Has this regulation increased the number of housing units in wards? One-year only regulation, announced on Dec 31st 2021.

Scenario 3: You have data on the number of units built across each ward in 2022 (the year after the introduction of the policy), but not for the years before. You do not know the % of green space in each ward in 2021, but you do know which wards had more or less than 25% of green space in 2021. You can access a set of ward characteristics from 2021 census (e.g. population number, avg housing composition, existing housing stock).

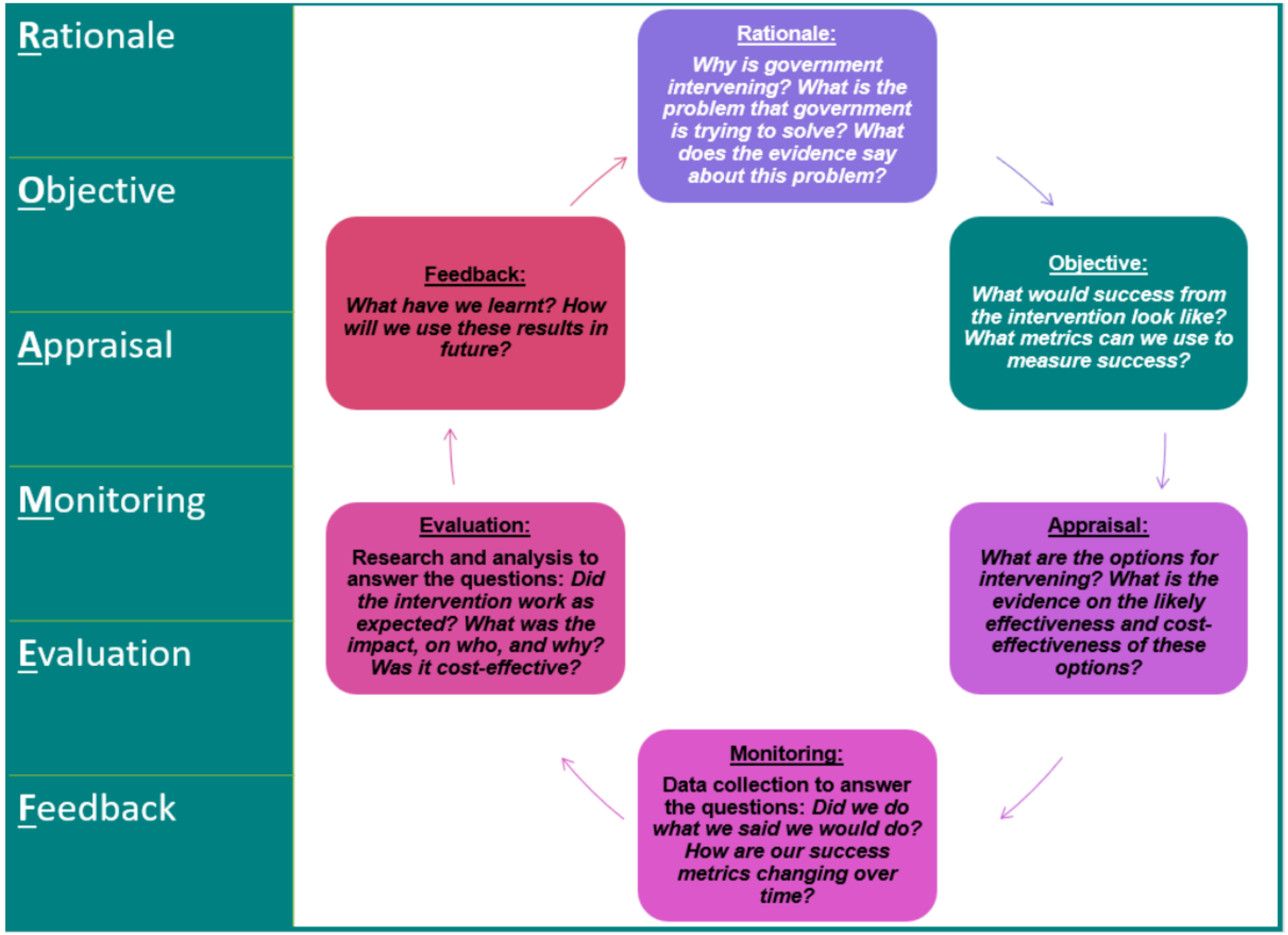
	2021 Green Space Threshold	2022 Built	Ward Characteristics
Ward A	Below	400	Yes
Ward B	Above	450	Yes
Ward C	Below	360	Yes
Ward D	Above	405	Yes
Ward E	Above	395	Yes

Matching

- Do not have data for the periods before the change was introduced
- Have access to a rich set of characteristics to match treated wards to similar untreated wards

Advocacy and application of learning

Including QED across the policy lifecycle



Rationale

Objective

Appraisal

Monitoring

Evaluation

Feedback

Rationale:
Why is government intervening? What is the problem that government is trying to solve? What does the evidence say about this problem?

Objective:
What would success from the intervention look like? What metrics can we use to measure success?

Appraisal:
What are the options for intervening? What is the evidence on the likely effectiveness and cost-effectiveness of these options?

Monitoring:
Data collection to answer the questions: *Did we do what we said we would do? How are our success metrics changing over time?*

Evaluation:
Research and analysis to answer the questions: *Did the intervention work as expected? What was the impact, on who, and why? Was it cost-effective?*

Feedback:
What have we learnt? How will we use these results in future?

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 quasi-experimental designs (in particular RDD, SC, DD).
- The types of research questions that QEDs can or cannot answer.
- The trade-offs and practical considerations when deciding between experimental and quasi-experimental methods.
- The key QED and when to use them: i) regression discontinuity; ii) difference-in-differences; iii) synthetic control; iv) matching; v) pre-post.
- The key things that need to be considered when preparing and running each QED design.
- The benefits and limitations of different QED designs.
- How to contrast the most appropriate QED method(s) to use based on the policy context
- How to critically assess the findings of a QED evaluation and advocate for including QEDs across the policy cycle.

Further resources

Resource
Evaluation and Trial Advice Panel
The Magenta Book: Central Government guidance on evaluation
ETF: Resources for evaluating policy in government
BIT: TESTS
The Green Book
Robust Nonparametric Confidence Intervals For Regression-Discontinuity Designs
The Experimenter’s Inventory