



OECD Working Papers on Public Governance No. 64

Seven routes  
to experimentation  
in policymaking: A guide to  
applied behavioural science  
methods

**Chiara Varazzani,  
Henrietta Tuomaila,  
Torben Emmerling,  
Stefano Brusoni,  
Laura Fontanesi**

<https://dx.doi.org/10.1787/918b6a04-en>

OECD Working Papers on Public Governance

# **Seven routes to experimentation:**

A guide to applied behavioural science methods

By Chiara Varazzani, Torben Emmerling, Stefano Brusoni, Laura Fontanesi, and Henrietta Tuomaila



OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors.

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to OECD Directorate for Public Governance, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: [gov.contact@oecd.org](mailto:gov.contact@oecd.org).

This publication was funded by the European Union. Its contents are the sole responsibility of the authors and do not necessarily reflect the views of the European Union.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at <https://www.oecd.org/termsandconditions/>.

# Table of contents

Acknowledgements	5
About this paper	6
1 Introduction	8
2 The case for experimentation in policymaking	10
3 How to choose a fit-for-purpose method	12
Why use a map?	12
How to use the map	13
5 key questions for choosing a fit-for-purpose method	14
4 Seven routes to experimentation	22
Experimental Methods	22
Observational Methods	28
Combining Methods	33
Pros and cons of the 7 methods	34
5 Conclusions	36
6 References	37
<b>FIGURES</b>	
Figure 0.1. Map of 7 Routes to Applied Behavioural Science	7
Figure 3.1. Five key questions guiding the choice of a method in the map.	12
Figure 4.1. Graph of a Randomised Controlled Trial (RCT) study.	23
Figure 4.2. Graph of an A/B Testing study.	24
Figure 4.3. Graph of a Difference-in-Difference (Diff-in-Diff) study.	26
Figure 4.4. Graph of a Before-After study.	28
Figure 4.5. Graph of a Longitudinal study.	29
Figure 4.6. Graph of a Correlational study.	31
Figure 4.7. Graph of a Qualitative study.	32
Figure 4.8. A table summarising the main goals, priorities and limitations of the seven methods.	35
<b>BOXES</b>	
Box 3.1. Observational and experimental methods	13
Box 3.2. Internal and External Validity	14
Box 3.3. Experiments, quasi-experiments, and natural experiments	15
Box 3.4. Designing for Generalisability	16
Box 3.5. Sample size, power, and effect size	18

Box 3.6. Labelling the methods	20
Box 4.1. A Randomised controlled case study: Behavioural prompts to increase early filing of tax returns: a population-level randomised controlled trial of 11.2 million taxpayers in Indonesia	23
Box 4.2. An A/B Testing case study: Enhancing the design of a webpage to increase the number of consumer product safety reports in Canada	25
Box 4.3. A diff-in-diff case study: Energy efficiency in Switzerland	27
Box 4.4. A before-after case study: Digital health in the UK	28
Box 4.5. A longitudinal case study: Physical and mental health in the US	30
Box 4.6. A correlational case study: The role of political ideology in energy conservation in the US	31
Box 4.7. A qualitative case study: Drought management in Australia	33
Box 4.8. A mixed methods case study: Transparency on online platforms in four European countries	34

# Acknowledgements

This working paper has been developed in collaboration between the OECD's Public Governance (GOV) Directorate, Affective Advisory GmbH, and the Chair of Technology and Innovation Management (ETH Zürich), under the guidance of Marco Daglio, the Head of Unit of the Observatory for Public Sector Innovation (OPSI), Carlos Santiso, Head of Division, and the leadership of Gillian Dorner, Deputy Director, and János Bertók, Deputy Director of the Public Governance Directorate.

The authoring team comprised Chiara Varazzani (OECD), Torben Emmerling (Affective Advisory), Stefano Brusoni (ETH Zurich), Laura Fontanesi (Affective Advisory), and Henrietta Tuomaila (OECD). This paper builds on a previous map ideated, researched, and designed by Laura Castro Soto, Judith Wagner, and Torben Emmerling (affiliated with Affective Advisory).

The authors would also like to express their gratitude to Ralph Hertwig, Lucia Reisch, and Kai Ruggeri for their helpful comments and suggestions on the paper, the members of the OECD Network of Behavioural Insights Experts in Government who contributed to this paper, and the collaborative efforts of the OECD Secretariat officials from across directorates, who provided their valuable insights and comments, including Katherine Hassett, Cale Hubble, David Jonason, Claire Karle, Nicolina Lamhauge, Bruno Monteiro, Jack Orlik, Silvia Picalarga, Claire Salama, Julia Staudt, and Piret Tõnurist. The authors would like to thank Laura Castro Soto and Judith Wagner for assistance with the design of the map and figures of this working paper.

# About this paper

In today's world, policymakers need to make timely and evidence-informed decisions to address complex societal challenges. Behavioural science methods are increasingly considered valuable tools for designing and evaluating policy solutions that better account for human behaviour. These approaches increase the effectiveness of the solutions and improve citizens' responsiveness to and trust in them.

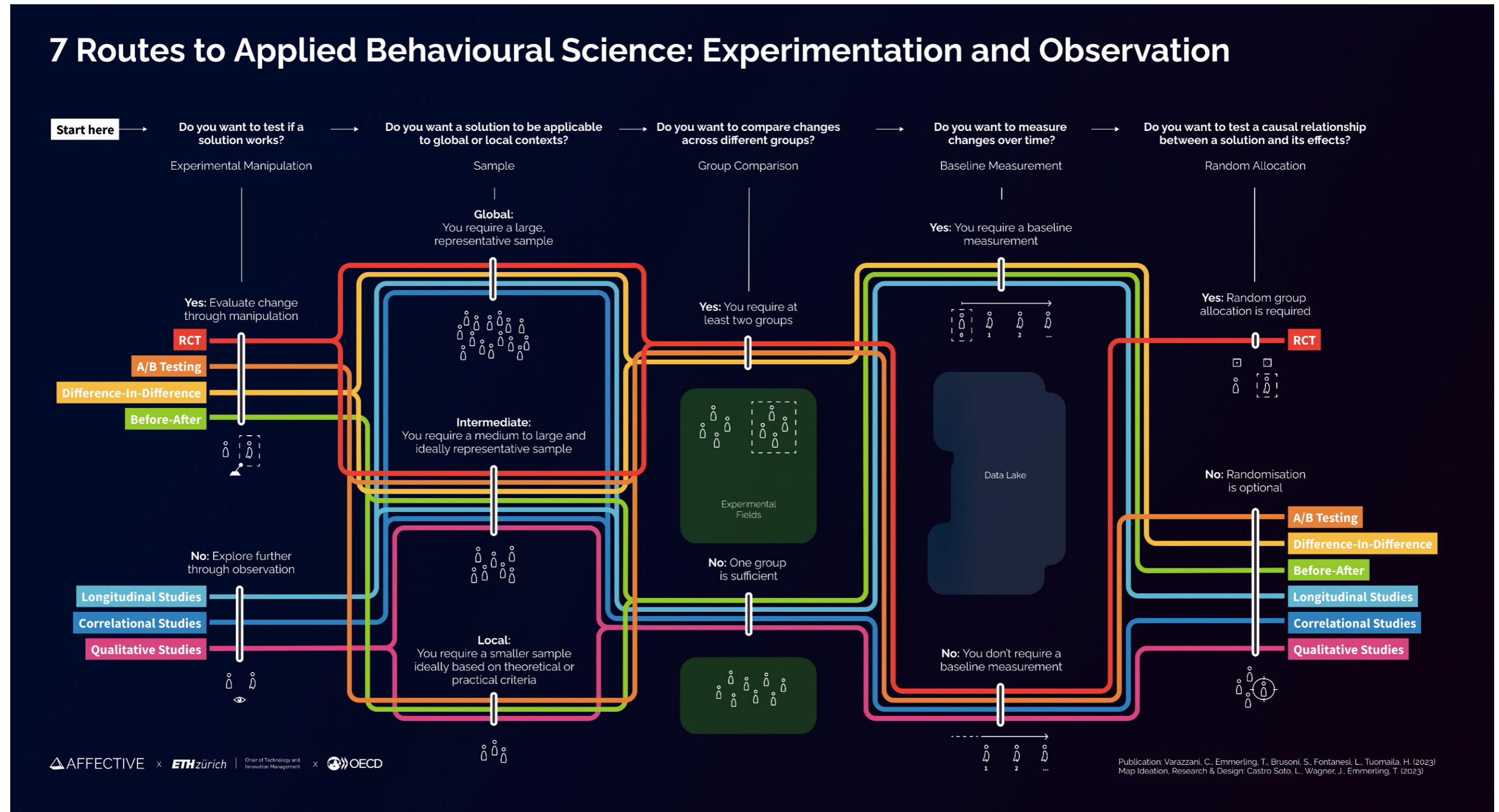
However, many practitioners found an obstacle in the lack of a standardised and accessible approach to guide them in choosing different experimental or observational methods to evaluate solutions in specific policy contexts. As a first step in creating such an approach, this paper provides a clear and intuitive guideline and an attractive map. Moreover, common guidelines and method labels help practitioners communicate and share their findings across jurisdictions.

The paper's five key questions (the five stations in the map) are designed to help navigate the choice among seven different methods (the seven routes in the map) considering the problem at hand, as well as time and resource constraints. Moreover, the paper offers a comprehensive overview of each method's distinguishing features and criteria for use. Each method is presented in a simple and easy-to-understand way.

By using the experimental and observational methods presented in this paper, policymakers can effectively measure policy solutions' impact and compare their effects across diverse cultural and geographical contexts. The visual aid and guidelines introduced in this paper will help policymakers find suitable experimental and observational methods to test and explore the effect of innovations and policies across borders.

By helping practitioners and policymakers select behavioural science methods to support evidence-informed policymaking, this guide will contribute to more effective, robust, and human-centred policies.

Figure 0.1. Map of 7 Routes to Applied Behavioural Science



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris. Note: The authors elaborated the map based on a previous map ideated, researched, and designed by Laura Castro Soto, Judith Wagner, and Torben Emmerling ([sevenroutes.com](http://sevenroutes.com)). For instructions on how to read the map, please see Chapter 3.



# 1 Introduction

Policymaking is traditionally grounded on the assumption that humans make decisions that, in economic terms, maximise their utility, meaning that these preferences consist of seeking the highest benefit for themselves (Kahneman and Thaler, 2006).

This assumption has been challenged by research in behavioural science, which showed that humans often make decisions that do not maximise utility (e.g., their relative preference for one option changes across different choice contexts). The field of behavioural science encompasses the study of human behaviour and the design of strategies to change it. It draws on research and methods from various disciplines, including behavioural economics, psychology, cognitive science, management sciences, sociology, and neuroscience. One of the major insights from behavioural science is that, in daily life, humans possess limited attention, memory, and computational capacities while dealing with a flood of information to process (Blanco, 2017, Simon, 1990). As a result, humans often rely on simplified behavioural strategies and attention management strategies (i.e., heuristics) to reduce task complexity, which can result in decisions that may not be welfare-improving (i.e., biases) (Tversky and Kahneman, 1973).

In the last two decades, policymakers have started incorporating these concepts into their work and designing policy measures with the support of behavioural science insights. The added value of behavioural science in policymaking is to diagnose problems better (e.g., What prevents people from behaving sustainably or saving money for their future?), to design solutions more precisely (e.g., How can policies be designed in a way that helps people act more sustainably or save more money?) and, most importantly, to increase the predictability of policy outcomes (e.g., What can I expect if I implement a particular intervention to promote sustainable behaviour or personal savings in a specific context and population group? What is the chance that it will backfire?) (OECD, 2017, 2019).

Given that people are embedded in specific cultural and geographical contexts, this prediction problem is challenging (Hallsworth et al., 2018). Discrepancies often arise between the perceptions of policymakers regarding potential effective policies and the real-world outcomes, or what works in practice. When designing for behavioural change, it is not enough to know that a policy solution has had a significant impact in another context or another country. Relying solely on insights from prior publications or assumptions about human behaviour can occasionally lead to misguided conclusions about the effects of a policy.

Instead, investigating the outcomes of a policy solution before scaling it up using experimental and observational methods indicates whether a policy is likely to work in a given policy context. Impact evaluations (i.e., experimental or quasi-experimental methods that enable a robust interpretation of statistical relations between variables) are common practice in economics and statistics (Angrist and Pischke, 2010). Yet, experimental and observational methods have not been extensively used in policymaking (OECD, 2020b).

Causal or context-specific knowledge generated by experimental or observational methods can be used to amend and tailor the design of a policy by taking into account the specificities of a given policy context, leading to more efficient use of public resources, as well as greater accountability and transparency given the evidence collected, essential aspects of good public governance (OECD, 2020b).

While there are many known ways to investigate the outcomes of a policy solution in the field, it can be challenging to put that into practice as there are many decisions to be made when planning for policy design and evaluation. Best practice guidelines are often tailored to specialised researchers and, therefore, hard to apply for practitioners. Moreover, there is a tendency to refer to the same methods using different names in different fields (e.g., clinical psychology, political science, or behavioural economics), making it harder for practitioners with diverse backgrounds to communicate with each other.

Responding to these challenges, this working paper provides a guideline for policymakers and practitioners with an easy-to-navigate map and straightforward information to support more informed decisions when designing policy solutions and evaluating their outcomes in the field.

This publication is for those who:

- Are looking for guidelines to compare different methods to generate the data necessary to evaluate a given policy solution;

- Are interested in a critical, comparative assessment of these methods and their requirements;

- Seek to develop a common language to discuss, in applied and actionable terms, the relative advantages and disadvantages of different methods.

This publication is not for those who:

- Are looking for a theoretical discussion about alternative methods for testing and collecting evidence;

- Need guidance on developing and designing a specific behavioural intervention;

- Need guidance on analysing and interpreting data.

# 2

## The case for experimentation in policymaking

With governments worldwide facing increasingly complex and wicked economic, environmental, and social challenges (OECD, 2020c), policy evaluations collected through experimental or observational methods can help demonstrate that government decisions and policies are informed by sound evidence, set realistic expectations, and effective spending of public resources, thus improving accountability (OECD, 2020b). Evidence-informed policymaking (EIPM) can help to build and maintain trust and promote fair outcomes for citizens (Brezzi et al., 2021). The OECD Council Recommendation on Public Policy Evaluation<sup>1</sup>, adopted in 2022, explicitly recommends that countries conduct policy evaluations to inform decision-making.

While policy challenges are often global (as they deal with human behaviour in various settings and culturally and geographically diverse contexts), policy solutions must consider local circumstances and respect cultural specificities. There is thus a pressing need to develop methods that enable the collection and analysis of data, rich in contextual knowledge and broad in outreach, as recently demonstrated during the COVID-19 pandemic (OECD, 2020c; Lancet, 2020), when the pandemic challenged citizens' trust in public institutions and governments worldwide (Brezzi et al., 2021).

Integrating behavioural science in policymaking is fundamental to addressing complex policy challenges. First, exploring the psychological and cognitive mechanisms behind human behaviour helps to understand why some policies work and others do not (OECD, 2020a). Second, it informs policy design, development, and delivery, improving the policymaking processes (Gofen et al., 2021). Third, behavioural evidence is crucial to understanding and learning how to increase the benefits and minimise policies' unintended consequences or spillover effects (OECD, 2020c).

Applying behavioural science and experimentation in public policy is also essential to better understand the effects and implications of biases in policymaking (Gofen et al., 2021; OECD, 2019). Researchers, government officials, policymakers and other decision-makers in public administrations are, like any other individual, prone to biases in decision-making (Drummond et al., 2021). Research shows that they may expect people to think similarly to themselves, leading them to inaccurately predict how people will behave and overestimate the degree of public support for a policy (Hallsworth and Egan, 2018).

By providing a deeper understanding of human cognition and the motivators of human behaviour, behavioural science can also help explain why the same policy solution may be received differently by different social groups in different contexts (Imai and Strauss, 2011). Decision-making research repeatedly showed that people's choices are highly contextual and are affected by individual differences and preferences (March and Simon, 1993).

Moreover, behavioural science comes with a body of experimental knowledge from social sciences and statistics on evaluating the impact of policy measures before implementing them on a larger scale. Piloting

---

<sup>1</sup> OECD Council Recommendation on Public Policy Evaluation  
<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0478>.

an intervention can provide valuable information on whether it creates the intended effect in a specific context and whom it affects (Al-Ubaydli et al., 2021; List, 2011).

Policymakers often need clarification about the method to use for collecting evidence. Usually, their work starts with a policy challenge (e.g., doctors prescribing too many antibiotics, causing dangerous antimicrobial resistance in hospitals). Their first task often is to collect evidence and analyse the evidence to understand why the policy challenge occurs, and what is causing it (OECD, 2019). Once the target population and behaviour causing the problem has been identified, and the challenge is better understood, can policymakers advance to designing a set of potential solutions (e.g., sending monthly reminders to the doctors and providing them with feedback on their progress) that derive from experience, literature research, or observation. In contrast to frameworks that incorporate behavioural science into public policy and typically recommend employing these methods at specific stages of the policy process, this working paper delves further into the discussion of these methods. It suggests their utilization regardless of the stage within a policy cycle, emphasising their relevance whenever evidence collection is required in policymaking.

A range of behavioural methods can improve understanding of challenging situations or test solutions, and the choice of method for collecting evidence in a policy setting can significantly affect the policy design and evaluation. Each has its pros and cons. For example, to better understand a rural area's need for portable solar energy sources (e.g., solar lamps), policymakers could set up a study to observe how people in a specific part of a country organise their working days. Alternatively, they survey a broader area to identify the most needed devices. Finally, they could conduct a field experiment to assess the impact of different pricing strategies of solar energy. All these methods deliver value, but they answer different questions.

Moreover, even the labelling of methods differs across academic fields, further hampering the integration of evidence into policy decisions (Ruggeri et al., 2020). Against this backdrop, the authors propose a framework for policymakers to transparently choose a method and communicate the reason for their choice.

# 3 How to choose a fit-for-purpose method

## Why use a map?

Determining the most suitable experimental or non-experimental approach is not a straightforward task. It involves a multitude of competing factors and questions that need careful consideration when applying behavioural sciences. Even seasoned scientists often dedicate considerable time to deliberating over methodological aspects and specifics. Additionally, certain projects may warrant the application of multiple methods.

Policymakers need suitable evidence on the safest and most feasible route to a policy solution given personnel, time, and cost constraints (e.g., the solution must be ready in a month, or there is limited access to hospital doctors). Some methods require more time and funds to be run, both of which are often scarce in practice, and this lack of time and funds thus makes some methods impractical for specific policy settings. Moreover, using some methods may require baseline data or the ability to collect several waves of data over time (more on this in Section 4).

Responding to the need of practitioners for a comprehensive yet straightforward guideline and overview, we present an innovative framework that supports selecting the most popular and frequently used methods that best fit the policy challenge, considering time and resource constraints. The framework, presented as a map, builds on five key questions (see the Figure 3.1) that, when asked in sequence, provide a step-by-step guide for choosing the most appropriate method. The questions were developed based on a thorough review of practice-oriented research literature and decades of experience in behavioural public policy. They form the cornerstones of a simple map for comparing, selecting, and using the different methods.

## Figure 3.1. Five key questions guiding the choice of a method in the map.

### The 5 key questions guiding the choice of a method in the map

- **Q1:** Do you want to test if a solution works?
- **Q2:** Do you want a solution to be applicable to global or local contexts?
- **Q3:** Do you want to compare changes across different groups?
- **Q4:** Do you want to measure changes over time?
- **Q5:** Do you want to test a causal relationship between a solution and its effects?

Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

### Box 3.1. Observational and experimental methods

The framework focuses on two main classes of methods: observational and experimental. Observational methods are used to explore human behaviour in natural settings, gain more knowledge, and develop hypotheses about which behaviours emerge in interactions with different environmental and socio-political factors. They often form the basis for the development of tailored policy solutions. Experimental methods play a crucial role in establishing a systematic approach to either supporting or disproving a hypothesis. Usually, hypotheses deal with understanding the cause-and-effect relationship between two main aspects. The first aspect involves actively making changes, which could mean introducing a new policy to tackle a novel challenge or adjusting an existing policy or its settings. The second aspect consists of the measurement and observation of something specific – in this case, the behaviour and actions of people. By utilising experimental methods, researchers aim to shed light on whether the changes in the first aspect have a discernible impact on the second aspect, helping to confirm or disprove their initial hypothesis.

Both observational and experimental approaches have merits and downsides (see Section 4 for a discussion on the pros and cons of the methods). At this stage, it is important to stress that whether a method helps inform policymaking depends on the evidence a policymaker seeks. Observational methods should be preferred when policymakers want to understand the drivers and barriers of behaviours in a specific policy context (e.g., Are all doctors prescribing too many antibiotics or only some of them? When and why are they engaging in this behaviour?). They support the exploration and definition of hypotheses for later experiments. Experiments should be preferred when policymakers want to identify a causal relationship but do not need to (or aim to) explain why a specific causal relationship exists (e.g., What is the effect of prompting doctors to prescribe fewer antibiotics by sending out a reminder once a week on their prescribing behaviour?). Usually, a combination of both methods works best for designing and demonstrating effective policy solutions.

Nevertheless, this guideline does not imply a one-to-one mapping between a specific method and a phase of the policy process. It rather encourages to select and combine different methods for different purposes. For example, policymakers can use experiments to explore new questions and observations to better understand an experiment's outcome.

### How to use the map

The map of the “7 Routes to Applied Behavioural Science: Experimentation and Observation” was further elaborated by the authors based on an innovative framework and map ideated, researched, and developed by Castro Soto, Wagner and Emmerling (2023). It offers a logical and visual guide for choosing a suitable method to collect evidence based on five key questions to be answered sequentially. Since usability is a top priority for policymakers, the visualisation comes as a map, connecting “routes” with different “stations” like the familiar design of metro maps. The five key questions correspond to different stations on the map, and the methods correspond to seven different routes, shown as lines on the map.

The map can be used in two ways, depending on whether one is evaluating and comparing different methods or already prefers a specific method. In the former case, it can be read from left to right, replying to the five key questions sequentially. In the latter case, the routes corresponding to the chosen method can be followed to learn the requirements and restrictions for that method.

The following section, looks at each of the five questions, illustrated as stations from left to right on the map, demonstrating the decision-making process that policymakers might face when implementing an evidence-informed approach. The following chapter will cover a detailed description of the seven methods.

### Box 3.2. Internal and External Validity

**Internal validity** refers to how accurately a study demonstrates a clear cause-and-effect relationship between different factors within a specific group or situation. It measures how much the observed effects can be attributed directly to the factor being studied, and not influenced by other things. Imagine a test of a new medicine – the internal validity would be strong if the study was designed in a way that could confidently show that any changes observed were due to the medicine itself, rather than caused by factors like the patients' personal traits, the placebo effect, or errors in measuring.

On the other hand, **external validity** looks at how well the results of a study can be applied to different situations, groups of people, or settings. It explores whether the findings hold true beyond the specific circumstances of the study. For instance, if a study involving college students reveals a particular behaviour, the external validity might be limited if we're uncertain whether the same behaviour would apply to people of different ages or cultural backgrounds. To improve external validity, researchers often use a wide variety of participants that accurately represent the larger group they're interested in, and they might conduct studies in real-life settings rather than in controlled lab environments.

## 5 key questions for choosing a fit-for-purpose method

### Q1: Do you want to test if a solution works?

The first key question asks whether a policymaker already has a solution in mind (and wants to test whether this solution works) or if they want to start by exploring and better understanding a specific context. For example, a policymaker could test a solution that worked well in a different geographical context and investigate its transferability to another context.

This question separates two different approaches (corresponding to the different stations in the map): when the response to the question is “*Yes: Evaluate change through manipulation,*” experimental and quasi-experimental methods (i.e., RCT, A/B Testing, Difference-In-Difference, and Before-After studies) are the proper routes to choose (see Box 3.3 for a discussion of the difference between experimental and quasi-experimental methods). When the response to the question is “*No: Explore further through observation,*” observational methods (i.e., Longitudinal, Correlational, and Qualitative studies) are recommended. For a more in-depth explanation and a visual representation of each method, see Section 4.

Simply put, experimental and quasi-experimental methods aim to test whether a solution (e.g., a specific behavioural intervention or something actively manipulated) affects what is measured and what needs to change (e.g., people's behaviour). Thus, they require a solution and the possibility of testing it on a sample. On the other hand, observational methods investigate how the target behaviour relates to further environmental and individual variables. Thus, they require no candidate solution but access to data already collected from the policy context (either qualitative or quantitative).

Knowing the answer to the first question is crucial before considering the methods more deeply. For example, if the policy goal is to increase energy efficiency, there are multiple ways to affect people's choices. It is hard to tell which way is best without understanding the impediments preventing people from consuming less energy.

The first step to addressing this challenge might be to run a qualitative study in the form of semi-structured interviews to ask people what their experience is when they receive their energy bill, whether they know if they are consuming more or less than the average per capita in their neighbourhood, and other relevant information. Through such interviews, one may discover that some people do not have a good understanding of their energy consumption because they have difficulty interpreting the numbers in kW/h on the bills.

### Box 3.3. Experiments, quasi-experiments, and natural experiments

Whereas researchers conducting experiments in laboratories can control for external factors more easily, conducting experiments in a real-world policy setting is more challenging. Such experiments, referred to as field experiments, often involve multiple factors that cannot be controlled and might affect the experiment's internal validity: its ability to draw reliable conclusions on the cause-effect relationship between policy solutions and behaviour (Shadish et al., 2002). On the other hand, field experiments have a much higher external validity: the capacity to generalise the experiment's findings to other populations and broader policy settings.

Researchers distinguish between "true" experiments, quasi-experiments, and natural experiments. The main difference among these terms is the level of experimental control that the policymaker has on the variables that lead to changes in behaviour. For example, to test whether a healthy eating campaign affects people's behaviours, one would compare people's eating behaviour in the same context: those exposed and those not exposed to the campaign. In practice, this is impossible since one cannot control exposure to the campaign in the same area.

In a perfect world, all regions would be equal in all aspects, and policymakers could decide which regions to expose to the campaign and which to have as a comparison. However, in the real world, when selecting regions to test a campaign, there will inevitably be differences in wealth, availability of healthy food, and other crucial variables related to outcome measures (i.e., healthy eating).

In **true experiments**, policymakers randomly assign participants or regions to either the campaign or non-campaign groups. True experiments are thus ideal for large samples, as the chance that the two groups of regions would be comparable on average increases among other things with the sample size.

In **quasi-experiments**, policymakers would actively select with which participants or in which region to run or not to run the campaign. In this case, they would need to balance the two groups, i.e., make sure that all types of individuals in the study population are equally represented in the sample, for the two groups to be as comparable as possible.

Finally, in **natural experiments**, the choice or scenario in question might have arisen outside of the researchers' control, spontaneously or due to external factors, such as policies adopted by certain regions. In such cases, policymakers do not actively control the experimental changes. Instead, they observe and analyse changes that happen independently, without their direct involvement or influence.

With this information, policymakers could start thinking about a solution, for example, changing the information shown on people's electricity bills. The bills could include information on how much one spent compared to others (as in the study of (Allcott and Mullnaitan, 2010), and such differences could be expressed not only in kW/h but made more concrete (e.g., by calculating how many km one can drive an electric car with that many kW/h as in the field experiment by Affective Advisory 2022, see Box 4.3).



With a defined solution such as this, it would be time to move from observational to experimental methods, such as manipulating the electricity bill's information and testing whether that would increase energy efficiency. The following key questions will help indicate which experimental method to use.

**Q2: Do you want a solution to be applicable to global or local contexts?**

The second key question is whether we seek a generalisable solution across different cultural, geographical, and socio-economic contexts or a more specific one for a local population and context. Ultimately, we are interested in identifying policy solutions that work globally. Yet, while policymakers deal with global problems, solutions also require sensitivity to the local context (in cultural or technological terms). Time and resource constraints can sometimes favour a greater focus on smaller and more local implementations. For example, a medical drug that cures a specific disease common across several geographic areas is a global solution. Yet, whether this drug is administered via a pill, an injection, or a suppository depends on local circumstances. Hence, both global and local solutions need attention.

### Box 3.4. Designing for Generalisability

In designing experiments, ensuring generalisability is crucial for drawing meaningful conclusions and making applicable inferences. Achieving generalisability requires careful consideration of various factors, including (but not limited to):

- **Sample Size:** A larger sample size increases the likelihood of obtaining reliable results that can be generalised to a larger population. With a larger sample, random variations and individual differences are more likely to balance out, leading to more robust findings. Conversely, a small sample size may limit the generalizability of results, as it may not adequately represent the target population, potentially leading to biased or misleading conclusions. See Box 3.5 to better understand the interplay between sample size and effect size.
- **Sample Characteristics:** The diversity and representativeness of the sample are also crucial to ensure broad applicability. A homogeneous sample may limit generalisability to specific subgroups, while a more diverse sample allows for broader generalisations. By considering demographic factors such as age, gender, ethnicity, and socioeconomic status, researchers can enhance the external validity of their findings and their potential to be generalised to the broader population.
- **Choice of Stimuli and Settings:** The stimuli used (e.g., the specific visual or verbal cues, or the environmental changes designed to prompt specific behaviours) should be relevant and ecologically valid, reflecting real-world scenarios as closely as possible. By using stimuli that resonate with the target population, researchers can enhance the external validity of their findings. Similarly, the settings in which the experiment is conducted should mirror real-life contexts to ensure that the results can be applied to similar situations outside the experimental setting.

This question is essential, as it directly impacts the study's necessary resources and scope of impact (i.e., broad, including different populations and contexts, or specific and limited to fewer populations and contexts). As a rule of thumb, the more critical generalisability, the larger the sample size needed (see Box 3.4 for a more in-depth discussion on the factors that affect generalisability and see Box 3.5 for a more in-depth discussion on the statistical consideration for determining a specific and appropriate sample size based on our expectations).

RCTs and Difference-In-Difference studies are usually the preferred methods when seeking global solutions. These methods scale well and offer robust findings based on a large, representative sample of a broader population. At the other extreme, qualitative studies are preferred when local solutions are the main interest, when exploring a new context, or when striving to identify the best implementation method for a known solution. A/B Testing and Before/After studies could also help study local contexts quantitatively, especially when the available project resources are not abundant.

However, many real-life situations are in between these extremes. Often, we would like to scale the analysis of a qualitative study, but we lack access to the resources needed to test qualitative findings on a broader sample. Nearly all methods can be used for the intermediate situations, but their results cannot be generalised easily. In all intermediate cases, and irrespective of the representativeness of the available sample, an adequate sample size should still be planned to achieve statistical significance, see Box 3.5).

In the previous example of testing different electricity bills to increase energy efficiency, if the focus is on small and medium enterprises (SMEs) in a specific country, selecting a subset of companies that is representative enough to run the study would be necessary. Testing the intervention on only a limited number of companies would make it challenging to assert the validity of the findings for all other industries in the country (by chance, these industries could be all small or come from the same sector).

It is worth mentioning that the desirable sample size depends on the stage of the project. When a solution has yet to be designed, running structured interviews with individuals can help identify what prevents them from saving more energy. In this early stage, it is neither possible nor desirable to collect an extensive sample: it would require too many resources to conduct such interviews at a larger scale, so it is important to focus on depth rather than numbers. In a later stage, this relationship might change in favour of a larger sample size for experimentation.

### Box 3.5. Sample size, power, and effect size

If a policymaker wants reliable evidence on whether their solution works, how big a sample do they need? This critical question in experimental design comes down to the interplay of three related factors: (1) how confident we want to be that an observed effect is not just due to chance (i.e., **the significance level**); (2) how confident we want to be that we will find an effect if there really is one (i.e., **the statistical power**); and (3) how big we expect the effect to be (i.e., **the effect size**).

Before we even start gathering data, a policymaker needs to pick a rule to judge if their solution works. This rule—the **significance or alpha level**—is like a threshold. It decides how confident we need to be before saying our solution has an effect. Normally, this threshold is set at 0.05. This means we are willing to accept a 5% chance of being wrong.

In the same way, we need to know the **statistical power** before we start gathering data. This is how good we are at detecting real effects. To ensure we are confident in our findings, we calculate the right number of participants using a **power analysis**. To do this, we also need to estimate the **effect size**, which shows us how much the solution actually changes behaviour.

In technical terms, the effect size is the normalised change in behaviour between two groups (typically, a control and an intervention group). “Normalised” means that it is possible to compare effect sizes across different studies measuring very different things, like the money put in savings accounts or the amount of vegetable portions eaten in a day. A large effect size indicates that the potential of the tested intervention to affect behaviour is quite high so larger effect sizes are more desirable. The relationship between effect size and sample size are generally inverse: smaller effect sizes require larger sample sizes, and vice versa.

Here's the tricky part: Larger sample sizes are not always better. First, lots of data can sometimes lead us to the wrong conclusion if it only comes from one type of group (i.e., they magnify a potential sampling bias). Secondly, more data allow us to find small, inconsequential effects. For example, in comparing two interventions with the same effect, tested in two separate studies A and B, where A had 150 participants and B 15000, both having the same significance level of  $p=0.05$ , intervention A would have a much bigger effect size than intervention B. To counteract these effects we should decrease our significance level to, e.g., 1% or even 0.1 %, when working with a bigger sample size.

Source: Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). *Big data and large sample size: a cautionary note on the potential for bias*. *Clinical and translational science*, 7(4), 342-346. Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). *Sample size, power, and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies*. *Biochemia medica*, 31(1), 27-53.

### Q3: Do you want to compare changes across different groups?

The next key question is whether it is desirable and whether resources are available to test the effect of a solution by comparing the behaviour of a treatment group (that has been exposed to the change, e.g., a new policy solution) with the behaviour of a control group (that has not or has yet to be exposed).

In principle, a control group allows one to deduce that a behavioural change is observed because of a policy solution, not something else. In other words, a control group provides a counterfactual, or what could have happened if a solution had not been implemented, thus providing a more solid basis for comparison. This phenomenon is similar to the testing of drugs in which the effect of a new treatment procedure or substance is compared to a placebo. While a counterfactual is necessary to infer a causal relationship between the solution and its effect, it is not a sufficient condition—more to this in question 5.

This third question is fundamental to deciding which experimental method to use. It separates the experimental methods into two classes: If the question is answered with "Yes", at least two groups are required. The resulting methods are RCT, A/B Testing and Difference-In-Difference Studies. On the other hand, when the answer to this question is "No", only Before/After Studies qualify as suitable experimental methods. In addition to Before/After Studies, all observational methods can be run without a second group.

In the energy efficiency example, if the majority of individuals in the sample decrease their consumption after receiving an updated version of the electricity bill, which includes information about how much they are consuming compared to other households and making the electricity numbers more concrete, how can one be sure that this decrease in consumption is due to the newly designed solution and not simply to seasonal fluctuations in overall energy consumption? Therefore, measuring energy consumption before and after implementing the solution with a Before/After study is not sufficient.

To be sure about the effect, data needs to be collected in the same months for other industries that do not receive the new energy bills but keep receiving the old ones instead (the control group). Only when there is a higher decrease in energy consumption in the treatment group compared to the control group can one reliably conclude that the intervention had an effect.

#### ***Q4: Do you want to measure changes over time?***

The fourth key question is whether one wants to detect a change from a specific point in time or compare differences in behaviour regardless of the timeframe. Measuring policy solutions' medium- or long-term effects is especially helpful in detecting undesirable and unintended policy effects that may occur with a delay. For changes over time, one must perform a baseline measurement (i.e., the ex-ante state of the world that one aims to change with a manipulation, e.g., a new policy solution) before introducing the solution.

This measurement is crucial for statistical analyses as it allows one to control for individual differences that could lead to incorrect conclusions. Suppose, for example, the individuals in the control group have a lower baseline measurement than those in the treatment group. In that case, they might also have a lower post-intervention measurement, and the comparison with the post-intervention measure in the treatment group would be unfair. To eliminate this bias and compare fairly, one could subtract both groups' baseline from the post-intervention measurement.

Moreover, this question is critical because it separates experimental and observational methods further. Among the experimental methods, one could use either Difference-In-Difference or Before/After studies to measure changes over time. If one is interested in changes regardless of a timeframe, either RCTs or A/B Testing studies could be used (although, in general, both methods also benefit from conducting a baseline measurement). Longitudinal studies are ideally suited to compare changes over time among observational methods. The other methods give an in-depth view of behaviour within a more limited timeframe.

In the energy efficiency example introduced in the previous questions, measuring the change in energy consumption from before the intervention (i.e., with the standard energy bills) to after the intervention (i.e., with the modified bills) was crucial. The baseline measurement was the energy consumption in the months before the beginning of the intervention. It allowed for considering pre-existing differences in energy consumption across SMEs.

### Box 3.6. Labelling the methods

The five key questions in the proposed framework provide a hierarchical structure where questions that come earlier in the sequence have more weight than those later to help differentiate and label methods.

All methods can be improved by adding features like baselines or randomisation. However, this should not affect their label. For example, one could plan a study with both a baseline to track changes over time and a randomised allocation of participants to a control and a treatment group to infer causality. In such a case, it might be unclear whether to use the RCT or the Diff-in-Diff label. According to the proposed hierarchy, however, since the question "Do you want to measure changes over time?" (Q4) comes before "Do you want to establish a causal relationship between a solution and its effects?" (Q5) and since more weight is given to earlier questions, we would call this a randomised Diff-In-Diff design. Likewise, a study with randomised allocation of participants to two different treatment groups and without a control group is labelled as a randomised A/B study and not as a RCT.

### **Q5: Do you want to establish a causal relationship between a solution and its effects?**

The fifth and last key question is whether one wants and has the resources to establish a solid and reliable cause-effect relationship between the policy solution (i.e., what is actively changed) and the measured change in behaviour (i.e., the effect of the policy on behaviour).

If the answer is "Yes," randomisation should be used. Randomisation is a statistical technique by which participants, i.e., the subjects of an experiment, are allocated to either a treatment or a control group by chance. As mentioned under question 3, having a control group is a necessary condition to infer causality but it is not sufficient. Randomisation is necessary when one needs or wants to control for potential confounding variables (i.e., variables related to both the group allocation and the measured behaviour) to infer causality. If individuals in the treatment group are significantly healthier than the ones in the control group, for example, and one wants to test the effect of a solution to increase healthy behaviour, the conclusions would be biased from the start.

Randomisation also helps to overcome the problem of self-selection, i.e., that participants volunteering to be part of the treatment group already differ significantly from the ones in the control group (e.g., the ones volunteering for healthy eating programs are also the ones that are healthier in the first place).

Yet, running an RCT also involves considerations and evaluations of the ethical implications of randomisation<sup>2</sup>. Despite the above-mentioned advantages, randomisation may not be a practical solution in cases where people in need would be restricted from getting the benefit of an effective intervention.

This question is only relevant for experimental methods: Technically, only RCTs require randomisation. At the same time, whenever there are resources to do it, it is advisable to use randomisation also in Difference-In-Difference and A/B Testing. Before/After studies would not allow randomisation as they do not have a control group.

In the energy efficiency example, imagine the policymaker wants to conclude that showing individual businesses what they consumed compared to themselves in the past and/or similar industries (i.e., the behavioural intervention) leads to reduced energy consumption (i.e., the behavioural change). This is an example of a causal statement.

<sup>2</sup> For more information on how to responsibly apply behavioural science in public policy, see the OECD's publication *Responsible by Design – Principles for the ethical use of behavioural science in government* (2022) at <https://oecd-opsi.org/publications/bi-gpps/>.

One would avoid the risk of confounding factors by randomly selecting and assigning individual businesses from different regions to either control or intervention groups. For example, if one region is significantly hotter in the summer leading to higher AC usage, assigning all businesses from that region to the control group would lead to an unfair comparison.

# 4 Seven routes to experimentation

After visiting the five key questions in the previous section, each method is considered more in-depth in the map, the methods are represented by seven metro routes, which are referred to in this section by their colour and name on the map. This chapter follows the methods from top to bottom. While this selection of methods should not be taken as an exhaustive list of all methods for investigating policy outcomes, it was chosen to be the most essential based on thorough literature reviews. It should not be considered a ranking since all methods can provide value for policymakers in different circumstances. Each method has pros and cons. Ultimately, their value depends on the problem policymakers intend to solve and the resources at their disposal.

## Experimental Methods

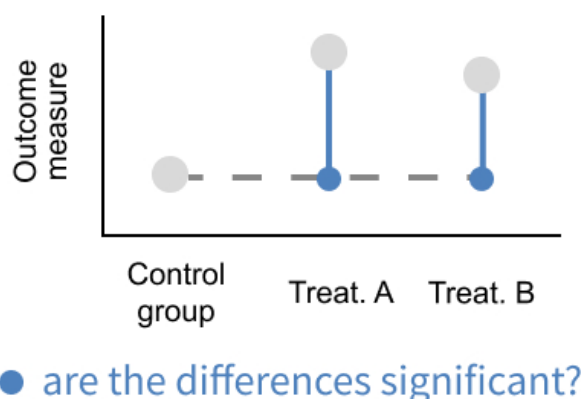
Within the experimental methods, one can differentiate between methods that look for effects at a certain point in time or over a period of time. The former are represented by the red and orange routes (RCTs and A/B Testing studies), and the latter are represented by the yellow and green routes (Difference-In-Difference and Before/After studies).

● **Randomised Controlled Trial (RCT) studies** (represented by the red line in the map) are an experimental method used to draw a reliable conclusion on the causal effect of a specific intervention or treatment. In an RCT, one should always ensure the existence of a comparable control group (i.e., a group that does not receive the treatment). Crucially, subjects should be randomly assigned to the groups so that the comparison between the control and treatment groups is fair (e.g., the two groups should have, on average, the same degree of migraines before the treatment and similar demographics that could affect migraines levels) (Glennester and Takavarasha, 2013).

An example is when one wants to know whether a particular drug is effective against migraines and thus needs to control for confounding factors, such as individual differences among people and the placebo effect. However, sometimes policies and strategies may involve multiple interventions, in which case an RCT is not necessarily able to capture multiple causalities between the interventions and their effects.

RCTs tend to have high internal validity when they are well-designed and implemented accordingly, yet, depending on the sample's representativeness, they do not necessarily always have a strong external validity (Kennedy-Martin et al., 2015). One way to improve the external validity is to have a representative sample of the target population. RCTs usually require more comprehensive resources: They tend to be relatively more expensive and require longer preparation. Moreover, more statistical and experimental expertise is usually required to set up the design in the right way.

Figure 4.1. Graph of a Randomised Controlled Trial (RCT) study.



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

#### Box 4.1. A Randomised controlled case study: Behavioural prompts to increase early filing of tax returns: a population-level randomised controlled trial of 11.2 million taxpayers in Indonesia

Last-minute tax filing is typical in several parts of the world. It can negatively affect tax compliance and the efficiency of tax collection systems (Persian et al., 2022). To encourage taxpayers to file their taxes earlier, the Behavioural Insights Team (BIT) and the Indonesian Directorate General of Taxes partnered to test the effect of 7 different email interventions.

The sample included 11.2 million individuals registered for online tax filing. Each personal income taxpayer was allocated to one of the seven conditions. These seven conditions included a 'pure' control group which received no email, a control group which received a reminder email similar to the previous email reminder, a simplification condition receiving a simplified version of the control email but reminding to file taxes early, a National Pride condition appealing to help to contribute to the nation by paying taxes, a Guidance condition giving guidance on how to file taxes and emphasising how early tax filing avoids problems, a Planning condition asking a taxpayer to indicate a date to file taxes and sending an additional reminder email two days before the indicated date, and a Guidance + Planning condition, combining the Guidance and Planning conditions.

All the reminder emails improved early and overall tax filing, compared to the 'Pure' Control group. Depending on the email, the results indicated a statistically significant increase of 0.8 to 2.1 percentage points in early tax filings. Apart from the National Pride condition, all the other email conditions led to a significant increase in early tax filing compared to the Control group. The email prompting taxpayers to plan when to file their taxes led to the highest increase in early tax filing.

By measuring the effects of the variations in the emails with a randomised experiment, they could establish a causal relationship between the interventions and their impacts.

Source: OECD OPSI Behavioural Insights Projects (2022), <https://oecd-opsi.org/bi-projects/behavioural-prompts-to-increase-early-filing-of-tax-returns-a-population-level-randomised-controlled-trial-of-11-2-million-taxpayers-in-indonesia/> (accessed 03-04-2023).



● **A/B Testing studies** (represented by the orange line in the map) are experimental methods used to compare two different treatments or manipulations against each other. A/B tests might be advantageous over Randomized Controlled Trials (RCTs) in cases where resources are relatively limited. This is because conducting an A/B Test doesn't strictly necessitate a control group or rigid randomization as two groups receiving different treatments can simply be directly compared to each other. For instance, two different versions of a website could be presented to separate user groups to determine which version generates more donations. Conducting A/B tests is often quicker and more cost-effective than running RCTs. However, due to their design, A/B Testing studies generally exhibit lower internal validity (see Box 3.2). As a result, A/B tests can be less reliable when attempting to establish a causal relationship between policy changes and behaviour. For instance, if the two websites are offered to individuals in different cities and one group yields more donations than the other, the observed effect could potentially stem from other factors influencing people in those diverse cities (e.g., one of the cities being wealthier or more philanthropic than the other).

Figure 4.2. Graph of an A/B Testing study.



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

### Box 4.2. An A/B Testing case study: Enhancing the design of a webpage to increase the number of consumer product safety reports in Canada

The Department of Health in Canada ran a randomised A/B Testing study (as explained in Box 3.6, A/B Testing studies, even if conducted with randomisation, they are not comparable to RCTs since they lack a control group) to test different enhancements to the Consumer Product Safety webpage (Experimentation Works, 2019). This policy aimed to increase the number of reported consumer incident reports.

During a series of qualitative interviews, Health Canada had identified several reasons for the low number of reported consumer incident reports: low web visibility, difficulty filling out a report, low transparency of the progress, and lack of a response after submitting a report.

Therefore, Health Canada decided to run an A/B test to test different changes to the website, including a button that made it easier to submit a report form, simplified instructions on how to fill out a report, and an explanation of why one should submit a report.

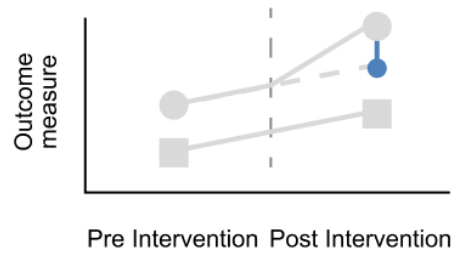
There was a total of 4591 web page visits during the three months the experiment ran for – 2592 visits on the existing page and 1999 visits to the modified web page, which included the link to the consumer product incident form. While 27% of those who visited the original web page accessed the report form, 61% of those who visited the enhanced web page continued to the consumer product incident report. This 34% difference was statistically significant, indicating that the clearer instructions and improved webpage design had succeeded in increase the number of reported consumer incident reports.

Source: Government of Canada. (2019). *Experimentation Works: Experimenting with visual design*. Retrieved: <https://www.canada.ca/en/government/publicservice/modernizing/experimentation-works.html>.

● **Difference-In-Difference (Diff-in-Diff) studies** (the yellow line in the map) are experimental methods comparing changes over a specific period. They are particularly suitable when seeking to reliably assess the impact of an intervention over time with relatively high reliability. These studies require a minimum of two groups: a control group that does not receive any treatment or intervention, and an intervention group that does receive the treatment or intervention under investigation. The underlying assumption is that both the control and treatment groups exhibit similar or parallel trends before the introduction of the intervention. As an example, when investigating the effectiveness of a specific treatment in reducing stress, it becomes necessary to account for individual variations among participants (such as their initial stress levels) or external factors such as environmental changes to ensure that these variations are similar or follow parallel trends.

To know whether the intervention had an effect, after the potential influence of individual differences and external factors are controlled for, it is necessary to compare the difference in stress levels before and after the intervention between the control and treatment groups (e.g., stress levels before and after the intervention). Only if this difference is more pronounced in the treatment group than the control one can reasonably conclude that the reduction in stress is likely attributed to the treatment itself, rather than to other factors that may have changed over time (Angrist and Pischke, 2010).

Figure 4.3. Graph of a Difference-in-Difference (Diff-in-Diff) study.



● is the difference significant?

Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

### Box 4.3. A diff-in-diff case study: Energy efficiency in Switzerland

As part of the Swiss energy strategy 2050, the Swiss Federal Office of Energy (SFOE) partnered up with Affective Advisory to test and evaluate behavioural interventions to increase energy efficiency in small and medium-sized enterprises (SMEs).

As a behavioural intervention, the team chose to test a modification of electricity bills for 400+ Swiss SME customers an electricity provider in Switzerland. This intervention was inspired by seminal studies of Allcott and colleagues (e.g., Allcott et al. 2011), which showed how, by including in electricity bills of private households in the US how much energy they were consuming compared to their neighbours or themselves in the past, they could effectively help them reduce their energy consumption. Not only had this study not been replicated in Switzerland before, but it also had yet to be previously replicated globally with SMEs.

The electricity bills were redesigned using insights from qualitative research, including among others social and individual comparisons, salient prompts, calls-to-action, reframing of information (e.g., a translation of kWh in driven kilometres) In total, there were two types of interventions, that consisted of including 1) information about how much was consumed compared to the previous year and compared to similar SMEs of the same industry, and 2) an info-sticker and/or box with a call-to-action asking recipients to increase their energy efficiency (i.e., "Increase energy efficiency now – here is how").

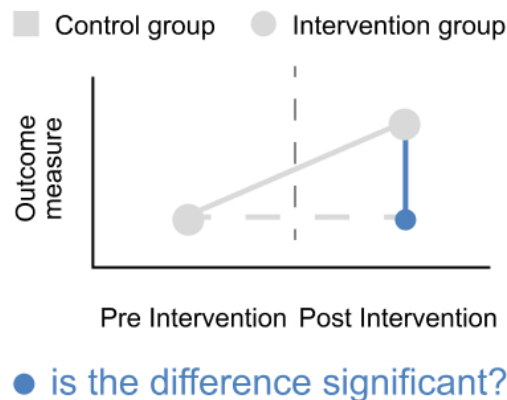
According to a 2x2 design structure, SMEs were randomly assigned to four groups, corresponding to the following bills: one with no comparison and no info-sticker, one with only the comparison, one with only the sticker, and one with both the comparison and the sticker. Energy consumption data were collected three months before the intervention to use as a baseline measurement and six months after the beginning of the intervention. The total experiment spans a timeframe of nine months.

The data was analysed using a diff-in-diff regression model, testing the change in energy consumption from baseline to the intervention period across the four experimental groups. The findings of Allcott and colleagues, which demonstrate increased energy saving behaviour as a result of social comparisons with a large sample of households, may not translate directly to energy conservation efforts in SMEs. Further research is needed in this area.

Source: Emmerling, T., Paul, A.F., and Seyffardt, D. (2021). *Behavioural Insights in Energy Policy: Behavioural science-informed potentials and interventions for increasing energy efficiency and the use of renewable energy in Switzerland's industry and services sectors*. Affective Advisory <https://affective-advisory.com/projects/ckw>.

● **Before-After studies** (represented by the green line in the map) are an experimental method comparing changes that compares changes before and after an intervention, albeit in a less controlled manner compared to methods outlined above. They can be particularly suitable when resources are limited, as this method does not require a control group. To conduct a Before-After Study, one must gather data from a single study group both before and after an intervention or treatment, and subsequently compare the two sets of measurements. The underlying assumption is that, in the absence of the intervention, the outcome for the study group would have remained consistent both before and after the intervention (Glennester and Takavarasha, 2013). However, due to this design, drawing definitive causal conclusions becomes more challenging: one can never exclude that a measurement could have changed because of a third variable that also changes throughout time (e.g., seasonal phenomena, daily fluctuations, or particular events).

Figure 4.4. Graph of a Before-After study.



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

## Observational Methods

### Box 4.4. A before-after case study: Digital health in the UK

Stallard and colleagues set up a Before-After study to evaluate BlueIce, a smartphone app designed to help young people manage distress and the urge to self-harm.

They recruited 40 young people with a history of self-harm. They assessed them first at baseline (before using the BlueIce) and throughout three main phases: 1) a familiarisation phase, 2) a post-familiarisation phase (after two weeks), and 3) a post-use phase (at 12 weeks). Participants' assessment included: A behaviour-screening questionnaire, a standardised measure of depression and anxiety, and self-reports of self-harm, app helpfulness, and safety.

The results were overall positive: there were no calls to the emergency numbers during the 12 weeks of using the app, their clinician withdrew no one due to the increased risk of suicide, and 73% of those who reported having recently self-harmed also reported reduced self-harming behaviour.

Moreover, there was a significant reduction in the depression and anxiety measured, and its users rated the app as extremely useful.

This study also shows how, by not including a control group, one can never be sure whether the changes observed over time were due, in this case, to the app use or other factors changing over the observation period. Nonetheless, this could be an excellent way to test an intervention, provided that appropriate risk management systems are in place (in this case, by involving the patients' clinicians).

Source: Stallard P, Porter J, Grist R A Smartphone App (BlueIce) for Young People Who Self-Harm: Open Phase 1 Pre-Post Trial *JMIR Mhealth Uhealth* 2018;6(1):e32 doi:10.2196/mhealth.8917

Within the observational methods, the light-blue line, representing longitudinal studies, looks at behaviour in depth and over time. The dark blue and pink routes, representing correlational and qualitative studies,

look at behaviour in depth but at a specific time. While longitudinal and correlational studies mainly deal with quantitative data, qualitative studies generally deal with verbal, visual and observational data.

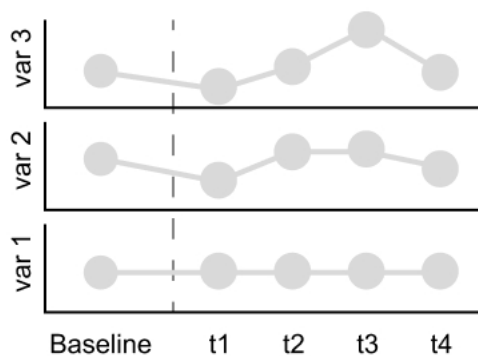
These routes belong to the “observational” realm. They are good when exploring the relationship among variables in all their complexity and natural occurrences. Therefore, it is less important to establish and quantify a causal link or control the effects of certain factors.

● **Longitudinal studies** (the light-blue line in the map) are a most often long-term-oriented resource-intensive observational method. They can help track the changes and monitor the evolution of one or two groups across multiple points over extended time periods. For example, suppose one is interested in understanding eating behaviour from early to late adolescence to modify existing school eating and nutrition policies. In that case, one needs to collect data in different schools over a few years (in several waves).

The complexity of a longitudinal study depends not only on how many individuals are followed and how long but also on the number of measurements one wants to track. For example, one could collect several physical measurements, such as BMI and glucose levels, and psychological measurements, such as different cognitive tests or depression, anxiety, and life-satisfaction questionnaires.

Sometimes, longitudinal studies could be set up as longitudinal experiments. In those cases, they would take the form of Difference-In-Difference studies, including a control group and possibly randomisation as well, but entail not just two but several measurement points over time and potentially more than just one measurement of the solution’s effect.

**Figure 4.5. Graph of a Longitudinal study.**



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

#### Box 4.5. A longitudinal case study: Physical and mental health in the US

To study the effects that the COVID-19 pandemic had on people's lifestyle and mental health, Giuntella and colleagues (Giuntella, Hyde, Saccardo, & Sadoff, 2021) used a longitudinal dataset consisting of both biometric and survey data from several cohorts of US college students. This dataset came from a wellness study that began before the pandemic and continued throughout when the courses moved from being on campus to online. The students received wearable devices (Fitbits) to collect biometric data and completed surveys about their well-being and use of their time. These data thus allowed the authors to conduct a longitudinal study on how physical activity and mental health evolved during the pandemic compared to baseline pre-pandemic levels.

In the first part of the study, they analysed the data and documented a significant disruption of the pre-pandemic levels of physical activity, sleep, and mental health.

In the second part of the study, they implemented a behavioural intervention in which they incentivised a subsample of the original cohorts (N=205) and randomly assigned them to two groups: A control group and an intervention group. Participants in the intervention group were incentivised to walk more than 10,000 steps. The randomisation considered the pre-intervention depression scores so that the two groups had similar scores distributions.

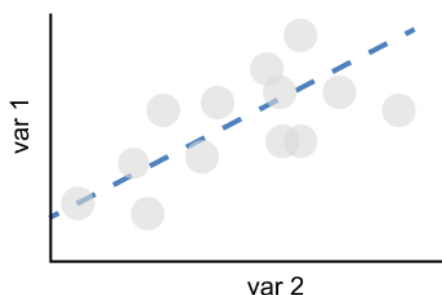
While the intervention significantly impacted physical activity, increasing participants' average steps by about 2,300 steps in the intervention group, there was no effect on the depression scores measured at the end of the intervention. Moreover, by continuing to track participants for a month after the end of the intervention, the authors found that the intervention effects declined after only one week to the control group's levels.

This study thus illustrates how longitudinal data can help understand behaviour in response to specific events or the long-term effects of behavioural interventions. They might be suited explicitly for testing how these effects evolve, even after interventions end, and when there is an established field set-up for data collection.

Source: Giuntella O, Hyde K, Saccardo S, Sadoff S. Lifestyle, and mental health disruptions during COVID-19. *Proc Natl Acad Sci U S A*. 2021 Mar 2;118(9):e2016632118. Doi: 10.1073/pnas.2016632118. PMID: 33571107; PMCID: PMC7936339.

● **Correlational studies** (represented by the dark-blue line in the map) are an observational method to explore relationships between two or more variables without engaging in experimental manipulation. They prove valuable when dealing with extensive quantitative data (e.g., health data, sales, voting preferences, demographics, online questionnaires) and the objective is to assess how specific variables co-vary. An example could be investigating the correlation between meat consumption and other lifestyle choices in relation to the likelihood of developing cancer (Becker et al., 2015). Correlational studies do not generally involve introducing changes or manipulations by researchers or policymakers. Instead, they delve into whether existing observable effects exhibit correlations or not. It is crucial to remember that correlations do not inherently imply causation. For instance, individuals who consume meat may also have distinct habits that correlate with a higher risk of cancer.

Figure 4.6. Graph of a Correlational study.



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

#### Box 4.6. A correlational case study: The role of political ideology in energy conservation in the US

Costa and Kahn (2013) sought to test the hypothesis that liberals are more likely to respond to energy conservation nudges compared to conservatives. Initial evidence had indicated that, compared to conservatives, liberals were more likely to engage in sustainable behaviours such as purchasing green products (Costa and Kahn, 2013; Kahn 2007, Kahn, and Morris, 2009) and restraining from purchasing products that have a negative impact on the environment (Costa and Kahn; Kotchen and Moore, 2008).

To test this hypothesis, the authors used data that had been collected for another study carried out by a California utility district, which included data on the participants electricity usage in response to a Home Energy Report (HER) and information about households' ideology, socioeconomic, and demographic factors (Costa and Kahn, 2013). A HER-report provides information about each household's monthly energy consumption relative to their neighbours and energy-saving tips.

The authors merged this data with the respondents' political party registrations, household donations to environmentalist organisations, household participation in renewable energy programs, and data on the specific characteristics of the residential communities where the respondents live (Costa and Kahn, 2013).

Based on theory, the authors formulated three hypotheses about the direction of the expected causal relationship between: 1) the effect of receiving the HER on energy consumption, 2) who is more likely to accept the treatment given their political partisanship, and 3) who values the report the most, given their political ideology. To test these hypotheses, the authors ran three correlational regression analyses; statistical tests to evaluate whether there exists a correlational relationship between two variables. The results from these analyses suggested that a nudge's effectiveness does indeed depend on a household's ideology.

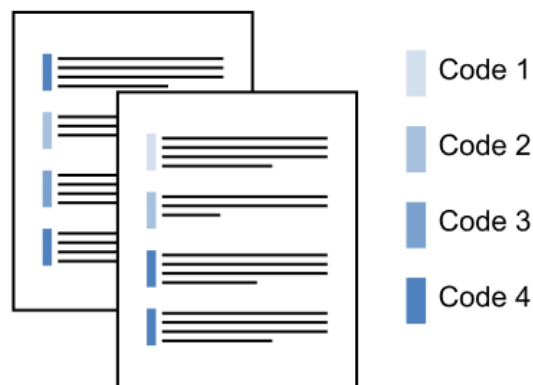
Source: Costa, D. L., & Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3), 680-702. Kahn, (2007). "Do Greens Drive Hummers or Hybrids? Environmental Ideology as a Determinant of Consumer Choice." *Journal of Environmental Economics and Management*, 54, 129–145. Kahn and Morris (2009). "Walking the Walk: The Association Between Environmentalism and Green Transit



*Behavior.* *Journal of the American Planning Association*, 75, 389–405. Kotchen and Moore (2007). "Private Provision of Environmental Public Goods: Household Participation in Green-Electricity Programs." *Journal of Environmental Economics and Management*, 53, 1–16.

● **Qualitative studies** (indicated by the pink line in the map) are observational methods employed to gain a deeper understanding of behavioural drivers and contexts. They excel when working with smaller sample sizes that involve potentially novel and intricate phenomena, which existing models might struggle to explain (such as a new infectious disease). Qualitative studies typically involve talking with and observing the behaviour of a specific target group. Therefore, these studies are often preferable: 1) before starting to gather quantitative data, to gain a more precise understanding of a situation or context to design solutions and identify the variables that are likely to affect the phenomenon of interest, or 2) when quantitative data from an experiment is available, but certain aspects remain inexplicable (e.g., difficult-to-explain outliers) and further contextualization of the findings is required. When employing qualitative methods, maintaining impartial and unbiased handling of information can at times be challenging (though this holds true for quantitative data as well). Moreover, participants might tend to provide responses that are perceived as socially desirable and reflective of their perception of the issue at hand. Qualitative studies are ill-suited for establishing causality (Glennester and Takavarasha, 2013).

Figure 4.7. Graph of a Qualitative study.



Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

The most prevalent qualitative research methods are interviews and focus groups (Gill et al., 2008). There are three primary types of interviews: structured, semi-structured, and unstructured interviews, each varying in their level of rigidity. Structured interviews involve a set of pre-determined questions and offer the least flexibility in terms of adaptation. Conversely, unstructured interviews have minimal to no guided questions. On the other hand, a focus group is a guided group discussion, providing a means to gain deeper insight into collective beliefs, social dynamics, experiences, and perspectives on a specific topic (Gill et al., 2008). Other commonly used qualitative methods include document analysis, ethnography and (existing) data analysis. Survey methods can also take stock of a situation, i.e., to collect data rapidly and cost-effectively, for instance, on the behaviours of a target group.

### Box 4.7. A qualitative case study: Drought management in Australia

To support Australian farmers in the face of severe drought, the Australian Government has implemented programs to support farmers, yet the uptake of these programs remained low. The Department of Prime Minister and Cabinet established the Joint Agency Drought Taskforce (JADT). They teamed up with the Behavioural Economics Team of the Australian Government (BETA) to better understand why farmers were not seeking help through these programs and how to communicate with farmers about the assistance (Perlesz et al., 2019).

The authors conducted a total of 19 semi-structured interviews. The interviewees included providers and coordinators of the assistance provided by the Government and farmers who were and were not currently receiving assistance. The providers and coordinators of the government assistance were recruited using JADT's networks and through snowball sampling. A third party recruited the farmers from the regions affected by drought. All interviews were recorded and transcribed, and a thematic analysis approach was conducted to identify patterned meanings in the interview answers.

The primary behavioural barriers identified by service providers were pride and shame, as asking for government assistance could have been seen as a failure. Among other things, farmers were found to get information from several channels on the assistance programs, yet to transform information to action, trusted advocates providing information and help in informal settings and face-to-face conversations were found crucial. In this context, exploring the barriers and enablers with the help of qualitative methods enabled them to provide nuanced insights into farmers' and providers' motivations, translating into precise policy action recommendations.

Source: Perlesz, L., Betros-Matthews, D., Truong, A. and Cotching, H., (2019). *Better Support for Farmers during Drought*. Retrieved: <https://behaviouraleconomics.pmc.gov.au/projects/better-support-farmers-during-drought>.

## Combining Methods

As discussed above, no two policy contexts are the same. Different situations may require different methods to address challenges as quickly, efficiently, and sustainably as possible. As a result, the process of collecting evidence for evidence-informed policymaking may involve employing a combination of the seven methods (**i.e., a mixed-method approach**) to efficiently identify behavioural challenges in more detail, to design effective policy solutions more precisely, and to increase policy solutions' predictability. The case study presented in Box 4.8 below exemplifies the use of mixed methods – both qualitative and quantitative – to gather evidence to increase the transparency of online platforms across four European countries.

#### Box 4.8. A mixed methods case study: Transparency on online platforms in four European countries

Online platforms provide consumers easier access to goods and services while challenging consumer protection and market competition (Lupiáñez-Villanueva et al., 2018). To explore and test how behavioural science can help to transform online platforms to be more transparent and fairer through regulation, Lupiáñez-Villanueva et al. (2018) conducted a literature review and a qualitative study to understand the behavioural drivers related to the usage of online platforms.

The literature review analysed the level of interest, awareness, and trust in online platforms. A "Think Aloud Online Task", which included three tasks related to information searches, a purchase situation simulation, and an assessment of user reviews, was conducted to complement this review on how consumers use online platforms. In total, 40 respondents from Germany, Poland, Spain, and the UK participated in various tasks online, accompanied and observed by an expert social researcher, to whom the respondents simultaneously explained their experience with online information search, purchases and user reviews while undertaking the task.

In the second part of the study, three online discrete choice experiments were run on a sample of 4800 participants from Germany, Poland, Spain, and the UK. The discrete choice experiments<sup>3</sup>, framed as a purchase decision in a mock-up e-commerce website related to booking a restaurant, purchasing a smartphone, and booking a hotel, tested the effect of different information content, visuals, and ranking ordering on the respondents' choices. Each experiment had a sample of 1600 participants, 400 from each of the four countries.

The results from the literature and qualitative task revealed that an average user is typically not that concerned about platform transparency and that people believe these platforms operate in the users' best interests. Insights from the experiments found that information on the ranking criterion and the order in which search results are presented significantly affect product selection. Using both qualitative and quantitative methods helped to understand why consumers behave the way they do and helped quantify what factors relating to information and visuals presented on online platforms influence a consumer's choice of a product the most.

Source: Lupiáñez-Villanueva et al., 2018. *Behavioural study on the transparency of online platforms*. Van Bavel, R., and Dessart, F.J., 2018. *The case for qualitative methods in behavioural studies for EU policymaking*. Publications Office of the European Union: Luxembourg.

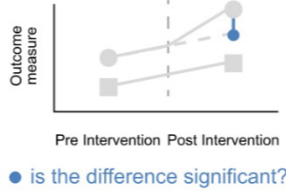
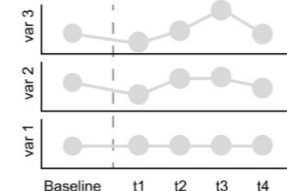
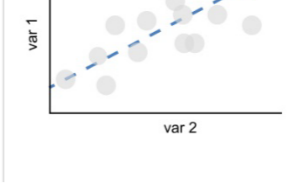
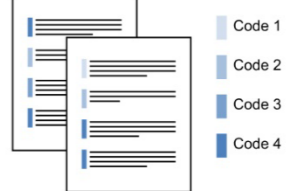
### Pros and cons of the 7 methods

The previous sections of this paper introduced key questions to consider when aiming to generate practical and relevant behavioural science evidence, depending on the problem's nature. It then provided a more in-depth introduction to the seven methods, highlighting key criteria and purposes. This section elaborates on this idea by emphasizing that every method has its own set of pros and cons, and it underscores that a rigid distinction between a "good" or "bad" method is neither useful nor preferable for policymakers. Rather, the suitability of a method depends on factors such as goals, priorities, and limitations.

The table below summarises these concepts and complements the map with further details on the pros and cons of each method.

<sup>3</sup> A discrete choice experiment is a way to elicit preferences without asking the preferences of individuals explicitly.

Figure 4.8. A table summarising the main goals, priorities and limitations of the seven methods.

Your preferred method is:	Randomised Controlled Trial (RCT) studies	A/B Testing studies	Difference-In-Difference (Diff-In-Diff) studies	Before-After studies	Longitudinal studies	Correlational studies	Qualitative studies
...when your goal is to:	Investigate the effects of one or more policy solutions against a control group.	Compare the effect of two policy solutions.	Investigate the effects of one or more policy solutions over a period of time.	Compare the effect of a policy solution between two points in time.	Better understand how different variables associated with a policy challenge develop over time.	Better understand which of the variables involved in a policy challenge are related to each other.	Investigate which/how variables are involved in a policy challenge, and/or the starting points for solution designs.
...when your priority is to:	Provide a reliable test of the cause-effect relationship between solution and effects.	Provide a fast(er) and cheap(er) comparison of two solutions' effects.*  * This is a rough estimate of the duration of an average A/B study, and might not apply to all.	Provide an estimate of a candidate cause-effect relationship between a solution and its effects over time.	Provide a fast(er) and cheap(er) comparison of the effect of a solution over time.*  * This is a rough estimate of the duration of an average Before-After, and might not apply to all.	Provide a broad, quantitative overview of how different variables associated with a policy challenge develop over extended periods of time.	Provide a broad, quantitative overview of how different variables associated with a policy challenge are related to each other.	Provide a rich and deep account of the factors behind a policy challenge, ideally formulating a testable hypothesis.
...when you can handle:	<ul style="list-style-type: none"> <li>High costs (research funds, time, and experienced personnel, creation of a control group).</li> <li>Not investigating a relationship over time.</li> <li>Not explaining why this relationship exists.</li> </ul>	<ul style="list-style-type: none"> <li>Not as reliable as RCTs (no control group).</li> <li>Not investigating a relationship over time.</li> <li>Not explaining why this relationship exists.</li> </ul>	<ul style="list-style-type: none"> <li>Not very reliable in assessing a cause-effect relationship unless randomisation is also performed.</li> <li>High costs (research funds, time, and experienced personnel).</li> <li>Not explaining why this relationship exists.</li> </ul>	<ul style="list-style-type: none"> <li>Not being as reliable as Diff-In-Diff.</li> <li>Not investigating more than one solution.</li> <li>Not explaining why this relationship exists.</li> </ul>	<ul style="list-style-type: none"> <li>High costs (research funds, time, and experienced personnel).</li> <li>Not explaining why this relationship exists.</li> <li>Not testing cause-effect relationships.</li> </ul>	<ul style="list-style-type: none"> <li>Needing access to a large data base.</li> <li>Not explaining why this relationship exists.</li> <li>Not testing cause-effect relationships.</li> </ul>	<ul style="list-style-type: none"> <li>Not being as broad as longitudinal and correlational studies.</li> <li>Not for estimating cause-effect relationships.</li> </ul>
Graphics:							

Source: Elaboration of the authors: Varazzani, C., Emmerling, T., Brusoni, S., Fontanesi, L., and Tuomaila, H., (2023), "Seven routes to experimentation: A guide to applied behavioural science methods," OECD Working Papers on Public Governance, OECD Publishing, Paris.

# 5 Conclusions

Real-world data on how humans behave and make decisions in specific policy contexts is crucial for designing, implementing, and evaluating policies. Behavioural science methods are increasingly integrated into evidence-informed policymaking, enabling more human-centred and scientifically robust solutions and driving more successful and trusted policy measures. Integrating behavioural science into policymaking has become more widespread as policymakers seek to design more effective and efficient policies. Both experimental and observational methods have proven helpful in gaining valuable insights into how people respond to different policy interventions.

Behavioural science-informed policymaking implies choosing the appropriate methods to evaluate policy measures and solutions. No straightforward guideline exists that allows practitioners, i.e., non-expert researchers, to select and compare methods in practice. While countries are increasingly investing in evidence-informed policymaking, there is an urgent need to develop common standards for defining methods to gather evidence and a more standardised approach to labelling methods for more accessible communication. Having a common language to communicate about methods across jurisdictions can foster evidence collection across borders.

Against this backdrop, this paper proposes an innovative guideline and map as a first step towards a more accessible and standardised approach to applying behavioural methods in policymaking. It proposes a creative and easy-to-access guideline and map as a first step towards a more accessible and standardised approach to applying behavioural methods in policymaking. It includes five "key questions" that enable the choice between a set of experimental and observational methods, depending on the core problem at stake, as well as on time and resource constraints. Moreover, it covers seven routes to experimentation and observation outlining each method's criteria and distinguishing factors in a simple overview.

As part of the Horizon 2020 Programme, the Observatory for Public Sector Innovation (OPSI) at the OECD has worked since 2015 together with the European Commission to promote and support innovation in the public sector. In the context of these efforts, the visual aid and guidelines presented in this paper will support policymakers in finding suitable experimental and observational methods to test and explore the effect of cross-border innovations and policies.

As policymakers explore new ways to design more effective policies, using fit-for-purpose methods will become increasingly important. The proposed map and its guidelines provide a valuable resource for policymakers seeking to incorporate behavioural science and methods into evidence-informed policymaking. It supports in selecting the most appropriate experimental and observational methods for specific policy context by providing a simple and easy-to-navigate framework for putting science into action.

# 6 References

Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C. L., & Suskind, D. (2021). How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural public policy*, 5(1), 2-49.

Allcott, H., & Mullainathan, S. (2010). Behavior and energy policy. *Science*, 327(5970), 1204-1205.

Angrist, Joshua D., and Jörn-Steffen Pischke. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24.2 (2010): 3-30.

Blanco, F. (2017). Cognitive Bias. In J. Vonk, and T.K. Shackelford (Eds.), *Encyclopedia of Animal Cognition and Behavior*. New York: Springer.

Brezzi, M., et al. (2021), "An updated OECD framework on drivers of trust in public institutions to meet current and future challenges", *OECD Working Papers on Public Governance*, No. 48, OECD Publishing, Paris, <https://doi.org/10.1787/b6c5478c-en>.

Chetty, R. "Behavioral Economics and Public Policy: A Pragmatic Perspective." *The American economic review* 105.5 (2015): 1–33. Web.

Castro Soto, L., Wagner, J., and Emmerling, T. (2023), 7 Routes to Applied Behavioural Science: Experimentation and Observation. Sevenroutes ([sevenroutes.com](https://sevenroutes.com))

Costa, D. L., & Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3), 680-702.

Drummond, J., D. Shephard, and D. Trnka (2021), "Behavioural insight and regulatory governance: Opportunities and challenges", *OECD Regulatory Policy Working Papers*, No. 16, OECD Publishing, Paris, <https://doi.org/10.1787/ee46b4af-en>.

Emmerling, T., Paul, A.F., and Seyffardt, D. (2021). Behavioural Insights in Energy Policy: Behavioural science-informed potentials and interventions for increasing energy efficiency and the use of renewable energy in Switzerland's industry and services sectors. [Affective Advisory \(affective-advisory.com\)](https://affective-advisory.com).

Gandy, K., Persian, R., Gibbons, D., Watson, J. and Akbari, R., 2019. Encouraging Earlier Tax Returns in Indonesia. *Policy Brief, London, UK: The Behavioural Insights Team*.

Gill, P., Stewart, K., Treasure, E. et al. Methods of data collection in qualitative research: interviews and focus groups. *Br Dent J* 204, 291–295 (2008). <https://doi.org/10.1038/bdj.2008.192>

Giuntella O, Hyde K, Saccardo S, Sadoff S. Lifestyle and mental health disruptions during COVID-19. *Proc Natl Acad Sci U S A*. 2021 Mar 2;118(9):e2016632118. doi: 10.1073/pnas.2016632118. PMID: 33571107; PMCID: PMC7936339.

Glennerster, R., & Takavarasha, K. (2013). Running randomized evaluations. In *Running Randomized Evaluations*. Princeton University Press.

Gofen, A., Moseley, A., Thomann, E., & Kent Weaver, R. (2021). Behavioural governance in the policy process: introduction to the special issue. *Journal of European Public Policy*, 28(5), 633-657.

Government of Canada. (2019). Experimentation Works: Experimenting with visual design. Retrieved: <https://www.canada.ca/en/government/publicservice/modernizing/experimentation-works.html>.

Hallsworth, M., Egan, M., Rutter, J., & McCrae, J. (2018). Behavioural government: Using behavioural science to improve how governments make decisions.

Hallsworth, M., and M. Egan (2018), The Illusion of Similarity, [The illusion of similarity | The Behavioural Insights Team \(bi.team\)](#).

Hansen, P., Larsen, E., & Gundersen, C. (2022). Reporting on one's behavior: A survey experiment on the nonvalidity of self-reported COVID-19 hygiene-relevant routine behaviors. *Behavioural Public Policy*, 6(1), 34-51. doi:10.1017/bpp.2021.13.

Imai, K., & Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1), 1-19.

Kahn, Matthew E (2007). "Do Greens Drive Hummers or Hybrids? Environmental Ideology as a Determinant of Consumer Choice." *Journal of Environmental Economics and Management*, 54, 129–145.

Kahn, Matthew E. and Eric Morris (2009). "Walking the Walk: The Association Between Environmentalism and Green Transit Behavior." *Journal of the American Planning Association*, 75, 389–405.

Kahneman, D. and Thaler, R.H., 2006. Anomalies: Utility maximization and experienced utility. *Journal of economic perspectives*, 20(1), pp.221-234.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and translational science*, 7(4), 342-346.

Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S., & Johnston, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*, 16, 1-14.

Kotchen, Matthew and Michael Moore (2007). "Private Provision of Environmental Public Goods: Household Participation in Green-Electricity Programs." *Journal of Environmental Economics and Management*, 53, 1–16.

The Lancet (2020), "COVID-19: a stress test for trust in science", *The Lancet*, Vol. 396/10254, [https://doi.org/10.1016/S0140-6736\(20\)31954-1](https://doi.org/10.1016/S0140-6736(20)31954-1).

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic perspectives*, 25(3), 3-16.

Lupiáñez-Villanueva, F., Gaskell, G., Tornese, P., Vila, J., Gómez, Y., Allen, A., ... & Veltri, G. A. (2018). Behavioural study on the transparency of online platforms. *Brussels: Office for Official Publications of the European Commission*.

March, J. G., & Simon, H. A. (1993). Organizations revisited. *Industrial and Corporate Change*, 2(3), 299-316.

OECD (2017), *Behavioural Insights and Public Policy: Lessons from Around the World*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264270480-en>.

OECD (2019), *Tools and Ethics for Applied Behavioural Insights: The BASIC Toolkit*, OECD Publishing, Paris, <https://doi.org/10.1787/9ea76a8f-en>.

OECD (2020a), *Building Capacity for Evidence-Informed Policymaking: Lessons from Country Experiences*, OECD Public Governance Reviews, OECD Publishing, Paris, <https://doi.org/10.1787/86331250-en>.

OECD (2020b), *Improving Governance with Policy Evaluation: Lessons From Country Experiences*, OECD Public Governance Reviews, OECD Publishing, Paris, <https://doi.org/10.1787/89b1577d-en>.

OECD (2020c), *Mobilising Evidence for Good Governance: Taking Stock of Principles and Standards for Policy Design, Implementation and Evaluation*, OECD Public Governance Reviews, OECD Publishing, Paris, <https://doi.org/10.1787/3f6f736b-en>.

OECD and MBRCGI (2023), *Embracing Innovation in Government: Global Trends 2023 (Preliminary Report)*, OECD Publishing, Paris, <https://oecd-opsi.org/publications/trends-2023/>.

Perlesz, L., Betros-Matthews, D., Truong, A. and Cotching, H., (2019). Better Support for Farmers during Drought. *Behavioural Economics Team of the Australian Government*. Retrieved: <https://behaviouraleconomics.pmc.gov.au/projects/better-support-farmers-during-drought>.

Persian, R., Prastuti, G., Bogiatzis-Gibbons, D., Kurniawan, M. H., Subroto, G., Mustakim, M., & Sutherland, A. (2022). Behavioural prompts to increase early filing of tax returns: a population-level randomised controlled trial of 11.2 million taxpayers in Indonesia. *Behavioural Public Policy*, 1-20.

Ruggeri, K., Linden, S., Wang, C., Papa, F., Afif, Z., Riesch, J., & Green, J. (2020). Standards for evidence in policy decision-making. *Nature Research Social and Behavioural Sciences*.

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power, and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, 31(1), 27-53.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin, and Company.

Silverman, D. (2019). Interpreting qualitative data. *Interpreting Qualitative Data*, 1-568.

Simon, H. A. (1990). Bounded rationality. In *Utility and probability* (pp. 15-18). Palgrave Macmillan, London.

Stallard, P., Porter, J., & Grist, R. (2018). A smartphone app (Bluelce) for young people who self-harm: open phase 1 pre-post trial. *JMIR mHealth and uHealth*, 6(1), e8917.

Tversky, Amos, and Daniel Kahneman. "Availability: A heuristic for judging frequency and probability." *Cognitive psychology* 5.2 (1973): 207-232.

Van Bavel, R., & Dessart, F. J. (2018). The case for qualitative methods in behavioural studies for EU policy-making. *Publications Office of the European Union: Luxembourg*.