OECD Working Papers on Public Governance No. 87

Unleashing the policy potential of rigorous impact evaluation and randomised trials

Stephane Jacobzone
Peter Bowers
Laurence Dynes
Harry Greenwell
Silvia Picalarga
Eleanor Williams

https://doi.org/10.1787/5c135873-en

OECD

Australian Government
The Treasury

Australian Centre for Evaluation
ace

OECD Working Papers on Public Governance n°87

Unleashing the policy potential of rigorous impact evaluation and randomised trials

By
Stephane Jacobzone (OECD)
Peter Bowers (ACE)
Laurence Dynes (OECD)
Harry Greenwell (ACE)
Silvia Picalarga (OECD)
Eleanor Williams (ACE)

Working Papers describe preliminary results or research in progress by the authors 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 Public Governance Directorate, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Abstract

Rigorous impact evaluations, including randomised trials, can provide governments with valuable insights into whether policies and programmes achieve their intended outcomes. When employed effectively, they offer an evidence base that goes beyond assumptions and precedent, supporting better resource allocation and more effective services for citizens. However, despite their potential, various technical and political barriers mean that such evaluations remain underused.

This report explores how governments can overcome these barriers and deliver high-quality evaluations to support policy development. It sets out the main evaluation methods available, ranging from randomised trials to quasi-experimental approaches, and explains the conditions under which each can be applied. The report explains ethical issues and highlights how such concerns can be addressed through careful design and stakeholder engagement. It also explores options for ensuring that evaluations remain cost effective, including through greater use of administrative data, alignment with policy priorities, and partnerships with wider networks. Finally, the report discusses the potential for artificial intelligence (AI) to contribute to impact evaluation, and the value of international co-operation and peer learning to build capacity, share methods, and upscale effective programmes.

Keywords: evaluation, monitoring, evidence-informed policymaking, randomised control trials, randomised trials, impact evaluation

JEL codes: A1; B49; C52; C9; H53; H11

Acknowledgements

This working paper was prepared jointly by the OECD and Australian experts. The experts in the OECD Public Governance Directorate include Laurence Dynes and Silvia Picalarga, policy analysts under the supervision of Stéphane Jacobzone, Senior Advisor. In the Australian Centre for Evaluation (ACE), experts include Peter Bowers and Harry Greenwell, under the leadership of Eleanor Williams, Managing Director of ACE. The authors are grateful to OECD colleagues Claudio Alberti, Andrew Blazey, Matthieu Cahen, Carlos Hinojosa, Anne Lauringson, and Matija Vodopivec, and to ACE colleague Andrew Mallos for their valuable comments. The paper also benefitted greatly from the insights offered by the speakers at a workshop organised in February 2025, at OECD premises, including Dr. Andrew Leigh, Assistant Minister for Productivity, Competition, Charities and Treasury Australia, Prof. David Halpern, Founding Director of the Behavioural Insights Team, United Kingdom, Prof. Michael Sanders, Kings College London, Douglas Sutherland, Head of Division, Economics Department, OECD, Prof. Tito Boeri, Bocconi University, Prof. Martina Björkman Nyquist, Stockholm School of Economics, Prof. Dan Levy, Harvard Kennedy School, and Dr. Paolo Paruolo, Head of the Competence Centre on Microeconomic Evaluation, European Commission, Joint Research Centre. The paper benefitted from editorial assistance from Andrea Uhrhammer and Meral Gedik.

Table of contents

FIGURES

TABLES

BOXES

Abbreviations and acronyms

ACE	Australian Centre for Evaluation
AI	Artificial Intelligence
AIDA	Artificial Intelligence for Development Analytics
BAC	Blood Alcohol Content
CC-ME	Competence Centre on Microeconomic Evaluation
CORE	Children Online: Research and Evidence
DAC	Development Assistance Committee, OECD
EEF	Education Endowment Foundation
EGM	Evidence gap-maps
EIPM	Evidence-informed Policymaking
ESIC	Evidence Synthesis Infrastructure Collaborative
ETF	Evaluation Task Force, United Kingdom
EU	European Union
FCDO	Foreign Commonwealth and Development Office, United Kingdom
GEI	Global Evaluation Initiative
HIV	Human Immunodeficiency Virus
HM Treasury	His Majesty's Treasury, United Kingdom
IGF	General Inspectorate of Finance, France
IMF	International Monetary Fund
INPS	National Social Security Institute, Italy
IT	Information Technology
J-PAL	Abdul Latif Jameel Poverty Action Lab
LLM	Large Language Model
MDE	Minimum Detectable Effect
NGO	Non-governmental Organisation
OECD	Organisation for Economic Co-operation and Development
RCT	Randomised Controlled Trial (also known as "randomised trials" in some contexts)
SDG	Sustainable Development Goals
TRACT	Trustworthiness in Randomised Clinical Trials
UN	United Nations
UNDP	United Nations Development Programme
USD	United States Dollars
WWN	The What Works Network, United Kingdom

Executive summary

Rigorous impact evaluations, and randomised trials in particular, have significant potential to improve decision making for policy purposes, particularly in terms of improving the efficiency and effectiveness of government interventions and of public spending. Impact evaluations are defined here as evaluations that seek to quantify the impact of a programme by exploring what would have happened in the absence of that programme (the "counterfactual"). In many cases, impact evaluations also explore the efficiency and effectiveness of a programme implementation in achieving these impacts. Despite the opportunities presented by high-quality impact evaluations, they are often underutilised. Reasons for this include assumptions that policy programmes already function effectively, a lack of resources and relevant skills, and an under-appreciation among policy advisers and decision makers of the benefits of impact evaluation.

There is therefore a need for experts to overcome several barriers that limit the use of impact evaluation generally, including randomisation and experimentation. There is also a need to address the common challenge of lack of political demand. Increasing the uptake of evaluative practices generally involves promoting cultural change and building a "test and learn culture". This involves developing policy environments where identification of unsuccessful programmes is seen as a step in the right direction rather than a setback, where evaluation is used throughout policy processes, and where learning-by-doing is actively encouraged. There is also a need to address the technical barriers to the supply of rigorous impact evaluation, including data access and lack of appropriate skills. Strengthening feasibility of impact evaluation means ensuring the evaluation approach chosen is proportionate to the programme and building effective data and information technology infrastructure. Programmes can also be designed to help foster a common language between researchers and policymakers, including through secondment programmes and training.

Artificial intelligence (AI) has significant potential to improve the efficiency of evaluation processes. AI tools can increasingly facilitate synthesis activities, allowing for evidence synthesis at a much faster rate than would otherwise be possible. These tools can automate time-consuming processes such as data collection, cleaning and analysis. They also have potential to support the use of randomised trials and quasi-experimental methods, by supporting participant selection processes, and assisting with measuring outcomes. Furthermore, they can help with the communication of evaluation results. However, use of AI may also come with some risks. If the data provided to AI tools is skewed or incomplete, this can result in erroneous predictions that reinforce previous mistakes. Inadequate data governance can also be an issue in the implementation of AI tools, particularly when it comes to sharing data across government units, as public administrations have access to sensitive data and have the obligation to protect the privacy and confidentiality of such data.

For impact evaluation to yield its full value in government, there is a need for findings to be effectively embedded into programme design, service delivery and ongoing decision-making. Preconditions include incorporating programme evaluation more broadly into legal and policy frameworks and establishing institutions that oversee the evaluation cycle and provide support to ministries. It is also important to engage relevant stakeholders throughout the policy process and effectively communicate results.

Government budget processes provide an important opportunity to use impact evaluation findings. Several OECD countries have taken measures to improve the impact of evaluation on budget allocation decisions.

Broader initiatives can support the development and use of impact evaluation at the global level, such as contributing to evidence gap maps, evidence synthesis, and knowledge management. Evidence gap maps provide a visual overview of existing evaluations within a sector and evidence synthesis brings together existing research and evaluation evidence to provide a holistic picture of existing knowledge. While these evidence products are well-established in some sectors such as health and international development, they are rarer elsewhere.

Effective delivery of impact evaluation involves consideration of the more appropriate methods and the trade-offs for different analytical approaches. Checklists and guidance documents can support such choices in the evaluation design phase and confirm the feasibility of an evaluation as well as the factors to consider when selecting the method. This includes checking whether relevant data exist, whether the sample is of sufficient size, and whether a credible counterfactual can be created. It is also valuable to develop a logical framework, or programme logic, highlighting the main features of the intervention, as well as how the results of the evaluation would likely be used. Evaluators should also identify and mitigate the risk of bias by preparing a pre-analysis plan and inviting peer review.

One reason impact evaluations—and randomised trials in particular—are not always widely employed is that they can be time and resource intensive. This paper discusses some strategies for managing these costs. This is also a reason for evaluation efforts to be prioritised.

Ethical issues need to be considered carefully when conducting a randomised trial, or other forms of impact evaluation. The paper outlines the ethical justification for using randomisation to exclude a control group from receiving a programme (at least temporarily). Randomisation is particularly justified when there is genuine uncertainty about a programme's effectiveness, cost-effectiveness, or scalability. Where resources are limited, it can also be fairer to ensure equal access across participants. Concerns about the deliberate exclusion of programme participants from a potentially beneficial policy can also be addressed by ensuring that participants can choose whether to consent to participate in the trial, offering the control group current standards of quality of care, and employing designs such as waitlists and staggered programme roll-out that delay rather than deny access. Engaging stakeholders by clearly explaining the purpose and ethics of randomisation and responding to concerns helps build trust in the process.

Evaluators and public officials should prioritise outcomes for the people and communities affected by policies, rather than protecting the policies or programmes themselves. Evaluations conducted by those within implementing agencies may be biased towards preserving programmes, even when results show limited impact. This can harm the very people the programmes are intended to benefit by allowing ineffective programmes to endure. Overcoming these challenges requires upholding evaluation independence and rigour, developing people-focused theories of change, using participatory methods, and clearly communicating findings, including when interventions do not work.

Finally, there are significant opportunities for greater co-operation at the international level. These include peer learning and exchange as well as knowledge hubs, including some that are academically led. There are also options to engage with a range of existing international initiatives, including those led by foundations and the not-for-profit sector. The OECD is also active in promoting evaluation systems through its Recommendation on Public Policy Evaluation, which supports governments in embedding robust evaluation practices and offers evaluation criteria and quality standards for evaluation under the auspices of its Development Assistance Committee and Public Policy Evaluation Experts working group.

1 Introduction

Sound public policy, grounded in rigorous evidence, is essential for effective government that can deliver services that meet citizens' needs and ministers' objectives. A strong foundation in evidence also supports accountability, increasingly driven by targets, missions and goals, which form the "what" of the political promise. Such a foundation is essential to maintain the effectiveness of government, and to ensure that governments can deliver on their promises. As noted by David Halpern, President Emeritus of the United Kingdom's Behavioural Insights Team: "*For mission-driven government, this includes open-minded use of the existing evidence on what works, active experimentation, data mining of variance, evaluation and use of feedback*" (Halpern, 2024[1]). These elements are crucial for building an "effective catalytic state", able to promote effective public interventions and help address market failures, while also mitigating the risk of government failures and reducing the potential for interventions that do not work. Such evidence requires capacity for measuring and monitoring the impact of public interventions.

Despite this, use of evaluation evidence in policymaking remains fragmented in many countries. Where evaluation is used, the insights generated are frequently underutilised or sidelined. This highlights the need for a "who"; that is, the institutions and individuals responsible for effective delivery, and the capacity to experiment and evaluate. There is also a need to ensure that there is appropriate demand for evidence, by stimulating attention at the political level and communicating in ways that can be understood by citizens and politicians. These elements are recognised in the OECD Recommendation on Public Policy Evaluation and its related toolkit, which calls on countries to strengthen their evaluation systems by looking at both supply and demand for evaluations (OECD, 2022[2]) (OECD, 2025[3]). Such features strengthen the links between evaluation, decision-making, and implementation, delivering more efficient and cost-effective policies and providing tangible outcomes for citizens.

While many different types of evaluation are valuable for government and public policy, the focus of this paper is impact evaluation, defined as evaluations that seek to quantify the impact of a programme by exploring what would have happened in the absence of the programme (the "counterfactual"). These counterfactual-based impact evaluations include randomised trials and a range of quasi-experimental approaches, such as regression discontinuity design, difference in difference approaches and control-on-observable methods.

There are also a range of other impact evaluation approaches (such as contribution analysis, realist evaluation, or process tracing) for answering related questions about programme impact. In addition, this paper acknowledges other important evaluation approaches, including implementation and process evaluations, and economic evaluations (see Box 1 for definitions).

Regardless of analytical approach, ideally impact evaluations are planned before a programme is implemented so that data collection can be developed to track outcomes throughout implementation. Impact evaluations may also be conducted on a pilot version of the programmes before it is adopted at scale. Alternatively, an impact evaluation may be designed and undertaken after the programme has been delivered noting that this may limit the data and analytical approaches available.

The paper seeks to draw from the leading initiatives in the field of impact evaluation, and follows a workshop held by the Australian Centre for Evaluation (ACE) and the Organisation for Economic Co-operation and Development (OECD) which drew on leading international expertise to explore the conditions required for rigorous impact evaluation to be embedded in the policy process. From an OECD

perspective, this workshop complemented efforts to support countries in implementing the Recommendation on Policy Evaluation, reinforcing the institutionalisation, quality and impact of evaluations and ongoing work of the OECD Development Assistance Committee (DAC) Network on Development Evaluation to strengthen evaluation policy and practice for sustainable development, including the use of rigorous impact evaluation (OECD, 2022[2]) and the work of the Public Policy Evaluation Experts working group. The workshop offered practical insights to support countries in scaling up high-quality evaluation practices and ensuring that they are fit for policy use. It discussed how governments and institutions can better generate, access, and apply high-quality evidence to inform policy decisions, particularly through experimental and quasi-experimental approaches such as randomised trials (see Annex A).

The paper is divided into three parts. The first part examines how to mobilise high-quality impact evaluations and evidence syntheses and support translation of findings into delivery and professional practice. The second part discusses how to enable innovative and impactful evaluation, making the best of the design phase, identifying relevant policy questions and balancing technical rigour with feasibility. The final part discusses broader opportunities for international cooperation in promoting the generation and use of high-quality impact evaluation.

2 Using evidence from impact evaluation to inform policy decisions

Despite its importance and relevance, as highlighted above, and the growing recognition of the value of impact evaluation in developing government policies and programmes, the consistent and effective use of evaluation evidence in decision making remains fragmented (OECD, 2020[4]). This section discusses the value of impact evaluation evidence in informing policy decisions and the barriers to this occurring.

2.1. The value of impact evaluations for policy

Impact evaluations matter for two main reasons. First, they help policy makers identify programmes that work to achieve their intended outcomes. Second, they help policy makers determine the extent to which these programmes work, and whether they are the most cost-effective way to achieve a given outcome. This helps both budgeting and planning decisions.

2.1.1. Impact evaluation helps identify the causal effect of a programme

Counterfactual-based impact evaluations are valuable to policy decision-making because they help policymakers estimate the causal effects of a policy or programme (Gertler et al., 2016[5]). In other words, they help achieve one of the most important but difficult goals in the creation of policy evidence: separating correlation from causation. They do this by comparing the outcomes for programme participants with what their outcomes would have been had they not participated in the programme (the “counterfactual”). The challenge for these impact evaluations is how to construct a credible counterfactual.

Impact evaluation covers a range of experimental and quasi-experimental approaches (see Table 1 in Chapter 3). For example, in his speech to the 2025 workshop Dr. Andrew Leigh illustrated the importance of causal evidence by describing a randomised trial that tested the impact of providing free reading glasses to adults in Bangladesh whose jobs require attention to detail, such as weavers and tea pickers. While reading glasses are an often-neglected intervention in developing countries, the randomised trial found that workers who received the glasses earned 33% more than those who did not receive glasses (Sehrin et al., 2024[6]).

The workshop had a strong focus on experimental methods, including randomised trials, which can produce compelling evidence when they are feasible, ethical and well conducted. However, recent advances in econometric methods, and wide access to large administrative datasets, have also expanded opportunities to use well-designed quasi-experimental approaches.[1] It is important to keep in mind that a balanced approach, with a focus on fit-for-purpose methods for each context, can help to reap the maximum benefits from randomised trials, and quasi-experimental methods using rich administrative data when they are available. (OECD/EC, 2025[7])

[1] See for example a discussion in the area of active labour market policies. (Card, Kluve and Weber, 2018[91]).

It can be tempting to rely on expert opinion, anecdotal evidence or common sense for making policy decisions. A range of different evaluative methods can provide more robust insights, depending on the question on hand. For example, process evaluation can be useful when the focus is on the implementation of an intervention and its fidelity to program design, while summative evaluations can be valuable in cases where the aim is to consider how a more well-established programme has evolved, and whether it is sustainable, efficient and effective. However, in cases where the aim is to test whether a clearly defined intervention achieves outcomes as intended, impact evaluations – and randomised trials in particular – are an effective tool.

Box 1. Different types of evaluation for different purposes

Policy evaluation can be defined as the structured and evidence-based assessment of the design, implementation or results of a planned, ongoing or completed public intervention. It assesses the relevance, coherence, efficiency, effectiveness, impact and/or sustainability of a policy based on its objectives (OECD, 2022[2]).[2] From this definition, it is clear that evaluations can take place in different moments of the policy cycle (*ex ante*, mid-implementation and *ex post*) and assess different aspects of the intervention.[3]

There are various different types of evaluation, each answering different questions. All can be valuable depending on the evaluation question of interest. One way to categorise the different types of evaluation and the questions they seek to answer is as follows (HM Treasury, 2020[11]).

- **Process evaluation** (or implementation evaluation): Is the intervention being implemented as intended? Is the design working? What is working and why?
- **Impact evaluation** (or outcome evaluation): What changes have occurred? Can changes in these outcomes be attributed to the policy in question? Did these changes vary for different groups? What is the scale of these changes and to what extent are they attributable to the intervention?
- **Economic evaluation** (or value-for-money evaluation): What are the costs and benefits of the intervention? Do the benefits outweigh the costs?

As this taxonomy illustrates, impact evaluation is an evaluation method that assesses the degree to which the intervention meets its higher-level goals and identifies the causal effects of the intervention. Impact evaluations may use experimental, quasi-experimental and non-experimental approaches (OECD, 2023[12]). Its significance lies in providing empirical data on whether the intervention has achieved its intended outcomes, thereby offering insights into its effectiveness and efficiency (Gertler et al., 2016[5]).

For the purpose of this paper, the issue of distributional impact analysis is not covered as it has been addressed in other OECD work (OECD/KIPF, 2024[13])

A range of high quality impact evaluations have shown that intuition and expert opinion can be misleading. For example, Leigh (2018[8]) points to the 'Scared Straight' programme in the United States that involved sending teenage juvenile delinquents to jail for a day to be 'scared straight' by seeing how tough prison

[2] The Glossary of key terms in evaluation and results based management for sustainable development prepared under the auspices of the OECD Development Assistance Committee also provides some useful entry points.

[4] P-hacking is an umbrella term referring to various questionable research techniques that researchers can use to inappropriately increase the chances of finding a statistically significant result, often by conducting repeated analysis with adjusted parameters.

really is. The programme had been well funded in multiple states of the United States, received celebrity endorsement, and even had a movie made about it. However, a systematic review of seven randomised trials with data on re-offending rates found that teenagers that took part in the programme were actually more likely to commit a crime in later life. The programme was not only failing to achieve its intended outcome, it was actually achieving the opposite of what was intended (Petrosino et al., 2013[9]; Van der Put et al., 2020[10]).

In essence, impact evaluations are a highly effective tool to determine whether public policies are delivering their intended outcomes. However, robust causal evidence in many cases only becomes possible because people were willing to acknowledge uncertainty about the true impact of the programmes and conduct a rigorous impact evaluation.

2.1.2. Impact evaluations can provide clarity on cost-effectiveness

There may be situations where evidence exists that a particular programme is successful, but where the extent of this success is less clear. A well-designed impact evaluation can address this by producing a robust estimate of a programme's impact. This estimate can then be combined with the cost of the programme to calculate how cost-effective the programme is, which allows policymakers to compare various programmes that are intended to achieve the same outcome (Gertler et al., 2016[5]).

For example, consider the policy question of how to improve school attendance in a developing country context. One potential programme to achieve this outcome would be providing free school uniforms to students. It would be reasonable for a policy maker to consider it extremely unlikely that such a programme has a negative effect on school attendance, and so an impact evaluation might not be considered necessary to answer this question. However, an impact evaluation could still be useful for determining whether providing free school uniforms to students is a cost-effective method of increasing school attendance compared to other methods available. For example, what is a more cost-effective way of boosting school attendance: giving out free school uniforms, or paying families directly if their student attends school (conditional cash transfers)? Research by the Abdul Latif Jameel Poverty Action Lab (2020[14]) found that while both methods are effective, provision of school uniforms in Kenya was more cost-effective at increasing years of education per student than an unconditional cash transfer programme in Malawi.

Without a rigorous impact evaluation, this comparison would not be possible. In the absence of an impact evaluation, cost-effectiveness analysis relies on making assumptions about the impact of the programme or comparing to similar programmes that have had impact evaluations. These assumption-based methods can work in some contexts, but using the results of a rigorous impact evaluation is generally preferable (ACE, 2025[15]; Gertler et al., 2016[5]).

2.2. Overcoming barriers to use of impact evaluation evidence in decision making

This section examines some of the challenges to greater use of impact evaluation in decision making – including biases, technical and bureaucratic barriers, most notably cost issues – and offers practical approaches for overcoming these barriers. It starts by exploring technical and bureaucratic barriers such as limited data access and fragmented skills before turning to political barriers and cognitive biases such as overconfidence and resistance to negative findings.

Evaluation methods and uses may differ when evaluations are needed *ex ante*, to inform a programme that is currently being developed, compared to when they are conducted *ex post*. The approach *ex ante* may require agile modelling and forecasting techniques, as well as experimentation and rapid evaluation methods, while evaluation *ex post* may offer an opportunity to make use of thorough administrative data and to conduct extensive applied analysis.

2.2.1. Overcoming technical and regulatory barriers

Technical barriers to the supply of evaluations

While data access can pose an issue for all types of evaluation, impact evaluations also face the issue of resource intensity in the production of data. Such evaluations often require data collection at various points in time, meaning different data collection tools must be designed, distributed and analysed at different points in the process, while evaluators must also invest in participant tracking systems to ensure they are following the same people throughout. Impact evaluations also often require large sample sizes. Taken together, this can mean that there are financial barriers in to implementing impact evaluations are high (Green Climate Fund, 2021[16]).

In cases where impact evaluations rely on existing administrative data, there are also several technical hurdles. Data may often be required from a wide variety of sources, and different sources may use different identifiers, making data linkage challenging. For example, a World Bank case (2020[17]) highlighted that even in Chile and Croatia, where judiciaries had advanced systems able to generate granular data, significant challenges arose when merging this data with firm tax records (World Bank, 2020[17]).

Another technical barrier to conducting impact evaluations in government can be a lack of people with the relevant analytical skills. This barrier is particularly true for impact evaluations, which require quantitative and qualitative analytical skills that can be difficult to attract in the public administration. For such skills to be fostered, they must first be identified and mapped.

This was a key reason the Australian Centre for Evaluation was established in Australia, the Evaluation Taskforce established in the United Kingdom, and various other entities in other countries. By making impact evaluation an explicit government priority, such entities are better able to recruit staff with experience in impact evaluation and to support other government agencies to identify the skills that they need to build their internal capabilities. Central evaluation entities can then support government departments and agencies to deliver impact evaluations through partnerships, feasibility assessments, technical advice, training or resource development.

Box 2. The Australian Centre for Evaluation (ACE)

Launched in July 2023, ACE is a branch within the Department of the Treasury aimed at improving the volume, quality and use of evaluation across the Australian Public Service. The Centre was established as part of the Australian Government's response to a 2019 Independent Review of the Australian Public Service led by David Thodey. This review identified evaluation practices within the Australian Public Service as fragmented, and a barrier to evidence-based policymaking.

The role of ACE is multi-pronged. It has already started working with government agencies to strengthen their evaluation capabilities, draw on existing evaluation evidence, connect with external evaluation experts, as well as to explore options to better embed evaluation practices into budget proposals.

ACE champions the use and delivery of high-quality impact evaluations, including randomised trials, and it leads several flagship impact evaluations each year in partnership with government departments, across portfolios including health, employment, education and social services. Its early work has included a series of randomised trials within the government's online employment programme.

Agencies are not required to use ACE's services but are encouraged to partner voluntarily. The Centre is responsible for fostering professional networks and peer learning through initiatives such as the Impact Evaluation Practitioners Network, and the Evaluation Profession. ACE also considers other cross-cutting challenges related to evaluation such as ethical oversight and access to administrative data for research. This whole-of-government approach to evaluation is expected to foster an evaluation culture and ensure that evaluative practices become standard practice across the policy cycle.

One skillset particularly relevant for randomised trials is that of quality assessment – in other words, analysing the robustness of the methodology behind an experimental trial, by looking at its sample size, use of preregistration, availability of data and code, and any signs of p-hacking.[4] Without such skills, misleading results from low-quality randomised trials can lead to poor policy outcomes and erode trust in the method itself. Some countries have explicitly incorporated such skills into competency frameworks. For example, the Competencies for Canadian Evaluation Practice include considerations such as whether the evaluator outlines strengths and weaknesses, determines the appropriate methodology for each evaluation question, indicates the data sources and appropriate tools, and assesses the validity and reliability of collected quantitative data (Canadian Evaluation Society, 2019[18]). In the United Kingdom, the Evaluation Task Force runs a five-day evaluation academy model which aims to upskill 40 future evaluation trainers in each cohort, including on methodological skills, allowing them to then set up training courses in their wider analytical community (OECD, 2025[3]).

Regulatory and data-related barriers to the supply of evaluations

Access to high quality administrative data can support more effective and efficient evaluations, particularly those using linked or integrated datasets, allowing researchers to study the tangible impacts of policy programmes across multiple outcomes. However, in many countries access to administrative data can be challenging. One of the most significant barriers is excessively stringent enforcement of data privacy regulations. While such regulations are critically important to protect inappropriate access to personal data and to preserve trust in data-holding government institutions, often their interpretation is so strict in a risk adverse bureaucratic environment that even researchers within government can have difficulties accessing the data for legitimate purposes. Even when data access is possible, verification processes can take time, meaning that research used for policymaking can be reliant on old data, blocking opportunities for real-time evaluation (OECD, 2025[19]). Furthermore, data ownership and access are often fragmented, with multiple different organisations responsible for data collection and storage, making it difficult for researchers to know where to look. Several countries are already making efforts to mitigate these issues. For example, in the Netherlands', while different government entities own various datasets, the central statistics agency has complete access, providing a single point of reference to researchers. While in Italy, the Fostering Open Science in Social Science Research project aims to create an Italian Open Science Cloud, providing a common platform of data sources so that researchers have a single point of access (Agenda Digitale, 2024[20]).

Technical barriers to the uptake of evaluations

One major obstacle to the uptake of evaluation is the issue of timing. Impact evaluation projects often operate on longer timelines, which are dictated by the time required for a programme's outcomes to emerge. By contrast, policymakers often require rapid results to inform immediate decisions. This mismatch means that by the time evaluation findings are available, the policy window may have closed, or priorities may have shifted. Even within government, evaluations and other impact assessments can often be produced too late in the legislative process to meaningfully influence decisions, thus serving more as a validation of political choices (if the result is positive) rather than a tool for exploring alternative options. In order to mitigate this issue, it is important to integrate evaluations into the policy cycle and start considering them during the policy planning stage. Over a longer timeframe, the accumulation of evidence from the monitoring and evaluation process can still be valuable and can ultimately contribute to broader evidence

[4] P-hacking is an umbrella term referring to various questionable research techniques that researchers can use to inappropriately increase the chances of finding a statistically significant result, often by conducting repeated analysis with adjusted parameters.

syntheses. Even if it is not used to influence immediate policy decisions, such evaluations can provide valuable lessons learned and points of consideration which policymakers can take on board going forward.

A related issue can be the lack of compatibility in the approach and incentives of researchers and policymakers. Researchers and evaluators often feel they are not able to properly engage with the policymaking process, while policymakers feel that they do not know how research and evaluation can best be utilised in this process. This is exacerbated by the fact that connections between these two groups, where they exist, are often informal and based on personal connections. This issue can be remedied through increase institutionalisation of the evaluation function in government. Embedding researchers in public offices can be one approach here, as it allows both parties more opportunities to communicate and thus better understand how they can help each other. During the workshop, several such examples were shared. For example, in Italy, the National Social Security Institute (INPS) developed a programme for researchers to conduct evaluations using INPS data. Another example is the use of Industrial PhDs, that fund applied doctoral-level researchers within government, as practiced in France. In some cases, issues also arise from misalignment of incentives, where policy work is not valued as highly as other types of academic work, and thus researchers may avoid it. Institutionalisation can also be beneficial here, leading to incentives more targeted towards the uptake of impact evaluations and the use of their results.

Discrepancies between the approaches of researchers and policymakers can be bridged in several ways. One option is to introduce partnerships with universities, and to even host researchers within government organisations to give them access to data and to relevant policy questions, which was highlighted at the workshop. Other options include strengthening the role of knowledge brokers within government. These can be individuals or organisations which function at the interface of evidence production and policy and are well versed in the communication styles of both. While the approach of knowledge brokers can vary, it frequently involves synthesising existing research results in such a way that their policy benefit is made clear, working with researchers to help them better understand the policy potential of their work, and offering training to both policymakers and researchers to help them better understand each other (OECD, 2025[21]; OECD, 2024[22]; OECD/EC, 2025[23]).

Another important point to ensure that impact evaluations are used and contribute to policy is to strengthen knowledge management, which has the potential for reducing the barriers to entry for uptake of evidence and information. One key component of knowledge management is the development of adequate infrastructure for the tracking, storing and sharing of evaluation evidence. This can allow decision makers to more readily access such information, understand what it was previously used for, and recognise what its potential could be. This can take the form of well-publicised centralised information repositories, as well as frameworks for evidence logging. Effective communication of the results of impact evaluations also matters, through clear writing and storytelling skills, which can help evidence stick in the minds of decision makers.

2.2.2. Addressing the lack of political demand

One of the main challenges faced by evaluators may be the lack of political demand for impact evaluation. Decision makers may not always see the need for thorough and complex evaluations, especially given time lags before evaluation results become available, and the tight deadlines within the policy-making cycle, as discussed above. These challenges may also be due to lack of dialogue, mutual trust and understanding between elected officials, senior government officials, and evaluation experts. These actors should be able to find the most cost-effective way to deliver on their mandate and ensure policy performance.

As touched on above, it can be easy for policymakers to make assumptions of programme effectiveness in the absence of evidence produced through impact evaluation. However, even when evidence gaps are highlighted and evaluations fill these gaps, results may not necessarily be utilised if they contradict existing strongly held beliefs and political convictions.

In some cases, there can also be a reluctance to conduct a robust evaluation where limitations to programme outcomes are anticipated. Fears from the programme team or other stakeholders that a rigorous impact evaluation will find their programme is not as effective as they had hoped are understandable as many large-scale interventions fail to show impact when rigorously tested.

A reluctance to rigorously evaluate therefore may reflect the desire for proponents of a programme or policy to emphasise its anecdotal effectiveness in order to get it adopted or keep it in place. This can also lead to the downplaying of impact evaluation results which show such programmes and policies to be ineffective. During the workshop, Dr. Andrew Leigh highlighted the United States' 21st Century Community Learning Center Initiative, which commenced in 1994 and expanded in 2002, and which saw children attending centres for after-school programmes. While several randomised trials found that attending the centres raised a child's likelihood of being suspended, and that there was no evidence that the programme improved academic outcomes, the programme was continued[5] (The Hon Dr Andrew Leigh MP, 2025[24]) (U.S. Department of Education, 2005[25]).

There are parallels in the health sector, which has a more embedded approach to testing interventions. Most clinical trials in medicine find the hoped-for treatment is not successful, which provides important information. The field of medicine has discovered numerous effective drugs and treatments by identifying many more that do not work, allowing the field to focus on the ones that do. Dr Andrew Leigh argued in the workshop that a similar approach should be taken in social policy, where the goal of evaluation is 'to find interventions that work', both by identifying those interventions that achieve their intended outcomes and ruling out those that do not. Adopting such an approach to public policy may require cultural change within some government departments and non-profit organisations that deliver social programmes. If a programme is found to not be working as hoped, the organisation or government branch running it should not be held at fault. Instead, they should be applauded for having the humility to admit they cannot be certain the programme they are administering is as effective as they would like it to be.

Furthermore, the results of impact evaluations need to be interpreted carefully. In some cases, a disappointing result may be enough to conclude that a programme or policy should be discontinued, while in others, the result may mean that the programme or policy design should simply be revised. This will depend on why the programme or policy 'failed'. Process evaluation, incorporating programme, survey or qualitative data, can often help shed light on this question, which is why mixed-method impact evaluations are often the most informative.

2.2.3. Building a 'test and learn' culture

Developing a test-and-learn culture can help make evaluation and experimentation more attractive. One way to develop such a culture is by taking an incremental approach to evaluation, making use of small impact evaluations at various steps of the policy process, and integrating this feedback into implementation processes. This incremental approach has multiple benefits. Firstly, as the evaluations are small-scale, any negative results are likely to be more acceptable than a negative evaluation of a programme as a whole. Furthermore, it allows time for trust to be built between policymakers, experts and evaluators. This supports the learning-by-doing approach discussed above, allowing researchers to test small adjustments and refine their approach over time. Finally, it makes it easier to embed evaluation into programme design from the very start, as opposed to it being left to the final stages of implementation.

The development of independent (or semi-independent) institutions with links to government can also help to bring about cultural change. The OECD Quality Standards for Development Evaluation underline the

[5] Although this was the subject of some discussion. ttps://www.researchgate.net/publication/304658267_The_National_Evaluation_of_the_21st-Century_Community_Learning_Centers_A_critical_analysis_of_first-year_findings

role of independence of evaluators vis-à-vis stakeholders in the context of evaluating development aid (OECD, 2010[26]). For example, ACE (see Box 2) has developed partnerships with other departments in the government where it uses experimental, quasi-experimental and other evaluation approaches to support internal teams in analysing the impact of departmental programmes. Its work is subject to transparency requirements, with the intention for all completed evaluations to be made publicly available on their website and a central library, thus limiting the scope for political influence. The United Kingdom's Evaluation Task Force has also developed a centralised evaluation registry to make the publishing of evaluation findings more routine.

The United Kingdom model is also supported by a network of What Works Centres that receive coordination support through the Evaluation Task Force at the Cabinet office, who also liaise with analytical and evaluation units of the central departments. An example of a well-established What Works Centre is the Education Endowment Foundation (EEF), which provides valuable tools that support teachers and school leaders who are making decisions around how to improve learning outcomes. The EEF synthesises the findings of primary studies, including an overview of the strength of the evidence supporting different education strategies and an assessment of the scale of learning outcomes. This toolkit is updated on a regular basis (Education Endowment Foundation, 2025[27]). The model of independent What Works Centres also provides the added benefit of maintaining continuity across political cycles.

Box 3. The United Kingdom's What Works Network

The What Works Network

The What Works Network (WWN) was launched by the United Kingdom Government in 2013 in order to mobilise and make accessible the evidence on 'what works' both to decision makers and practitioners. It is composed of nine Centres, each working in different areas of social policy. The Centres regularly meet. A stocktaking study on the Centres by University College London found that they conducted a wide variety of work – most notably, building a more comprehensive evidence base, raising awareness regarding the need for use of evidence, and influencing policymakers to consider evidence more effectively. To do this, they often conduct syntheses of research findings, as well as translating findings into briefings to make it more accessible. They have developed sophisticated communication methods. The Centres have criteria for standards of evidence, some internally created and some externally developed, to ensure that evidence used in policymaking is sound

The Evaluation Task Force

The Evaluation Task Force (ETF**)** is a joint Cabinet Office-HM Treasury unit established in 2021, created in response to concerns about the lack of robust evaluation in government spending. The ETF was designed to address this gap by ensuring that evidence and evaluation sit at the heart of spending decisions. Inspired by medical testing, the ETF brings an experimental, evidence-led approach to public policy, working to embed a culture of learning and accountability across departments. The team consists of 15 evaluation specialists and provides targeted advice and support to both HM Treasury spending teams and individual departments. Beyond advising on spending proposals, the ETF promotes stronger evaluation practices across government, playing a coordinating role in the United Kingdom's network of What Works Centres. It manages the Evaluation Accelerator Fund, a GBP 15 million initiative to support new evaluation activity in priority policy areas. As of 2023, it had funded 16 projects, including a trial monitoring prison wastewater to detect drug use, and an evaluation of one-off payments to 18-year-olds leaving care.

Source: (UCL, 2018[28]) (OPSI, 2023[29])

In governments where evaluation is not consistently regarded as a valuable skill, programme evaluations are less likely to be implemented. Learning-by-doing, as discussed above, can be one method for fostering an evaluation culture, i.e. taking a more forgiving approach to making mistakes that allows more junior staff to engage directly with datasets and methodologies early on, and thus learn evaluation-relevant skills more quickly. It could also be encouraged through developing teams that include a variety of professions (i.e. economists, psychologists and data analysts), thus allowing for a wider range in perspectives on evidence quality and how evaluations should inform programmes.

2.3. Evidence synthesis and identifying evidence gaps

In order to determine which evaluations should be prioritised, it is helpful to understand the existing evidence landscape and the gaps that exist. Evidence gap maps are thematic tools designed to provide a visual overview of existing and ongoing impact evaluations and systematic reviews within a specific sector or sub-sector, highlighting the types of programmes evaluated and outcomes measured. Use of such tools can be valuable in implementing the OECD's Recommendation on Policy Evaluation, which calls on countries to make use of evidence synthesis methodologies to aggregate evaluation findings and assess them in a systematic manner. This helps focus the analysis where evaluations are most needed, as well as coordinate efforts and avoid overlaps. Evidence gap maps can be a particularly useful methodology due to their ability to provide a bird's eye view of all available evidence in a given sector, or on a particular policy topic, helping identify areas where evidence is lacking.

Box 4. The experience with evidence gap maps in international development

Evidence gap maps (EGM) can be used in international development, as seen with '**3ie**', the International Initiative for Impact Evaluation. These tools present a visual overview of existing and ongoing studies or reviews in a sector or sub-sector, demonstrating the types of programmes being evaluated and the outcomes measured. This evidence is mapped onto a framework, graphically highlighting areas where few or no impact evaluations or systematic reviews exist and where there is a concentration of impact evaluations but no recent high-quality systematic review.

EGM methods draw on the principles and methodologies developed for systematic and transparent evidence review, including consultation of relevant decision makers and stakeholders, particularly at the start of the mapping, to discuss the scope, questions and framework, and at the end of the mapping to review draft findings. EGM then involves a systematic search to identify relevant published and unpublished impact evaluations and systematic reviews, including both completed and ongoing studies and reviews. The resulting map includes characteristics of the evidence base, and is populated with studies, links to their summaries and, when available, full-text reports. In the case of systematic reviews, EGMs also include a confidence rating of how the review was conducted to help decision makers gauge how much to rely on that evidence.

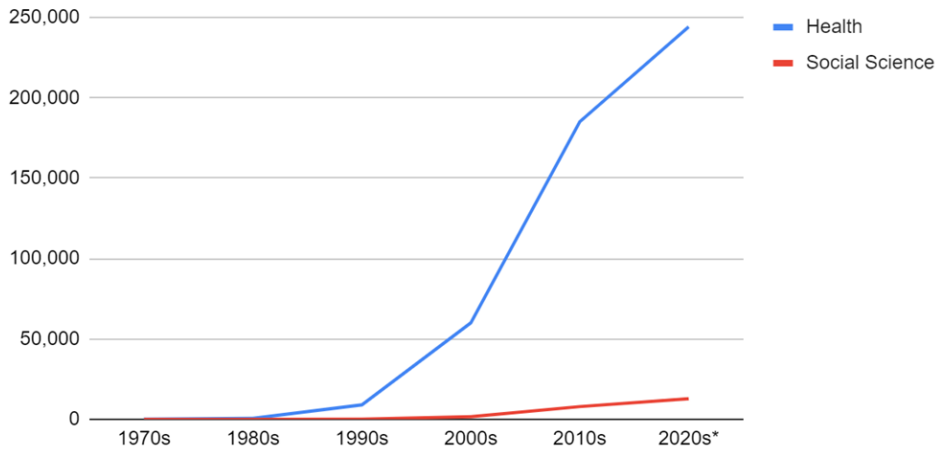
Source: Evidence gap maps: a starting point for strategic evidence production and use, 3ie Working Paper 28. Uttley L, Quintana DS, Montgomery P, Carroll C, Page MJ, Falzon L, Sutton A, Moher D. The problems with systematic reviews: a living systematic review. J Clin Epidemiol. 2023 Apr;156:30-41. doi: 10.1016/j.jclinepi.2023.01.011. Epub 2023 Feb 14. PMID: 36796736.

Evidence gap maps are well established in areas with the most experience in evaluation, such as health and international development (see Box 4). However, as shared by David Halpern at the workshop, routine production of evidence remains thin outside these sectors, particularly regarding the use of experimental and quasi-experimental designs (see Figure 1). Furthermore, there is often little correlation between spending levels and production of evidence through research and development (R&D). For example, in

the United Kingdom, large social departments, such as the Department of Work and Pensions, spend only 0.24% of their budgets on research evaluation and analytics, whereas the smaller Department for Business Energy and Industrial Strategy spends close to 4%, and the Foreign Commonwealth and Department office spends slightly less than 3% (BIT, 2024[30]). This finding was echoed by other workshop participants, who mentioned that countries often tend to spend more on evaluating foreign aid than some of their domestic activities.

Figure 1. Production of evidence, health vs social sciences

Campbell Collaboration analysis of Web of Science data, forecast as of 2024, Number of RCTs



Source: (BIT, 2024[30])

There would be clear value in taking stock of existing evidence to better assess evidence gaps, both at national level and international level through international cooperation (see Part III). Often, domestic path dependency tends to prevail, where past investments in research and evaluation drive new and future investments, given existing expertise, skills and capacity to mobilise funding. This may explain why some issues, such as labour markets and education, tend to be heavily researched, while new or emerging issues, such as the challenges of public safety, do not benefit from the same level of investment. One new area which has seen significant investment is the green transition. This area has benefited from additional investments or special efforts, such as those undertaken by the OECD in the area of Carbon Mitigation approaches (Box 5).

As highlighted by the OECD example of the Inclusive Forum of Carbon Mitigation Approaches in Box 5,[6] there are two main challenges to identify and map what has already been done to get a sense of what works, and to ensure broader sharing of existing evidence through better knowledge management. Sharing of evaluations more openly can create opportunities for broader engagement and can help create broader consensus between researchers and policy makers on the weight and implications of evidence. While across OECD countries policy evaluations are usually public by default (61%) these evaluations are not

[6] The need for robust mechanisms for monitoring climate commitments and evaluating the impacts is also acknowledged in (OECD, 2025[93]).

always collected in unified and user-friendly evaluation portals (42%)[7] (OECD, 2025[31]). The challenge is therefore to ensure that within government, there is capacity to discuss and learn from evaluation results in a safe space in order to maximise the gains from a policy decision, continuously improve policies and programmes, and minimise any negative impacts.

Box 5. The OECD's Inclusive Forum for Carbon Mitigation Approaches

As part of the Inclusive Forum for Carbon Mitigation Approaches, the OECD has been providing a comprehensive overview of the effectiveness of climate policies, systematically reviewing, gathering, assessing and comparing evidence, then identifying gaps and drawing lessons for policy design and implementation. This forum brings experts from across policy communities and fields of expertise, which is critical for building a multi-disciplinary approach. As the number of policy evaluations related to climate mitigation has roughly quadrupled over the past decade, the forum's challenge was to identify how to synthesise such a large and fast-growing body of work. Using a systematic evidence synthesis approach, the forum first selects which studies to include, before setting up protocols for how to assess the evidence and extract findings and working to make the results comparable. This has helped to identify summary statistics and distribution of values and elasticities of about 400 estimates from around 200 studies.

The work of the forum has revealed that major evidence gaps remain, potentially due to publication bias. While most of the evaluations that have been identified relate to emissions trading schemes, little is available on regulation or government procurement programmes. While North America and Europe are well represented, Africa is less so. In terms of sector, energy and industry are well represented, while agriculture is not. This already gives a sense of where additional research and effort might be most valuable. Putting everything on a common scale helps identify outliers and questionable results. Additional work entailed using kernel density graphs to show the spread of results with colour coding to highlight quality variations.

One of the challenges is to foster the use of such information and to communicate its value. As many countries are conducting evaluations at national level, these are often not shared or accessible internationally. There would therefore be value in bringing them together as part of repositories, as an international public good, so countries can learn from one another, as discussed in section 4.

The OECD Recommendation on Policy Evaluation calls on countries to make evaluations available, promoting transparency. However, as previously mentioned, uptake of this component of the Recommendation remains uneven. This is often the case for evaluations in policy domains that make use of sensitive information, such as defence. Another potential reason for this is the risk of unfavourable evaluation results being poorly received. This can be mitigated by ensuring that evaluations are conducted where decision makers and other stakeholders want to learn about the effectiveness (or lack thereof) of a policy or programme, and where they clearly understand that there is a possibility of a disappointing result before the evaluation is commenced. Furthermore, when results are shared, they should be compared to benchmarks for similar programmes implemented in comparable settings. For some of them, a statistically significant effect of 2-3% will also be practically significant for policy makers, while some stakeholders might feel that the number is lower than they expected.

[7] Data are from the 2023 OECD Survey on Public Policy Evaluation. Responses were collected from 31 OECD countries, referring only to central/federal government practices as of 1 May 2023 and reflecting the country's own assessment of current practices and procedures. Respondents were country delegates with cross governmental co-ordinating functions on policy evaluation.

Beyond this, there is a need to be able to take stock at national level and identify the relevant set of evaluations in each country. This challenge was highlighted through a recent multi-country project at European level, with countries identifying the lack of such knowledge management tools as an important barrier to effective use of evaluations at national level. In some countries, such as Latvia, repositories exist for publicly funded research, although these do not cover all policy evaluations conducted at national level. Another example comes from France, where the Court of Audit (*Cour des Comptes*) has created a platform of public policy evaluations with a database of all available evaluations at national level, with nearly one thousand evaluations currently available (OECD, 2025[3]). Other French institutions such as the General inspectorate of Finance have also created databases of inspection reports, which often have a significant evaluative dimension. These make use of artificial intelligence in systematically analysing existing evidence (see section below). The Australian Centre for Evaluation library is another example that provides easy access to evaluation evidence from across the country (Australian Centre for Evaluation, 2025[32]). In addition, in some policy areas, knowledge management platforms exist at the international level, such as the DAC Evaluation Resource Centre (DEReC) that counts over 4 000 evaluation reports from development agencies, and the CORE- Knowledge base on children & youth in the digital age at the European level (OECD, 2025[33]; CORE, 2025[34]).

2.4. Mobilising AI to support effective evaluation and evidence synthesis

Growing interest in the use of Artificial Intelligence (AI) across government and among evaluation experts and academia makes it important to reflect on how AI can be leveraged to support evaluation and evidence generation. AI can play a significant role in supporting policy evaluation and evidence synthesis by increasing the productivity of evaluators, facilitating synthesis activities, and by allowing researchers and policymakers to leverage a broader range of data than would be possible if processing manually. However, the use of AI for policy evaluation remains at early stages, particularly within government. Recent OECD (2025[35]) analysis of 200 AI use cases found that only five use cases related to evaluation, the lowest of the 11 government functions analysed in the report. Additional investments and guidance on how best to use AI tools could help in better understanding its potential.

OECD work on this topic has mapped some of the different areas in which AI can be used for policy evaluation and identified use cases (OECD, 2025[35]). More specifically, AI has the potential to accelerate and automate essential tasks, such as data collection and analysis, and to support evaluation design and implementation. It can also help build predictive systems and simulations to anticipate potential impacts before implementation. The Australian Centre for Evaluation has prepared a series of short training videos to help evaluators work through how to use AI to support tasks in their work. While some tasks are not yet systematically supported by AI, others have been employed for some years, particularly in supporting impact evaluations where counterfactual analysis was not possible due to data limitations (See for example (Abrell, Kosch and Rausch, 2022[36])).

AI has been applied to streamline the process of evidence synthesis, particularly in areas where policy decisions must draw from wide-ranging research studies. Through tools such as text mining, machine learning classifiers and automated extraction software, AI tools can support each phase of systematic reviews, from screening literature and extracting data to summarising findings and identifying patterns that would not be obvious to a human analyst. At the same time, early examples show the importance of supervised approaches, and the significant front-end investment required to establish the analytical framework and prepare data for machine learning (Franzen et al., 2022[37]).

Some AI tools have been in use for over a decade. Examples include “Rayyan”, an AI software that aids with rapid screening of titles and abstracts (Johnson and Phillips, 2018[38]), “Robot Reviewer” software that helps to assess the risk of bias (Jardim et al., 2022[39]), and “Covidence” which streamlines the process of conducting systematic literature reviews (Kellermeyer, Harnke and Knight, 2018[40]). These tools

significantly reduce the time needed to complete comprehensive reviews, making them more compatible with real-time or time-sensitive decision-making. Evidence syntheses also benefit from AI's ability to process both structured databases and full-text documents, widening the pool of accessible material for review and minimising manual input requirements. This has enabled organisations, such as the World Bank, to expand the size and scope of their evidence base when conducting programme reviews, which would have otherwise been constrained by resource and time limitations. AI-driven tools can also extend the analytical options available to government evaluators by filling data gaps (Lee and Lee, 2020[41]). Predictive techniques offer the flexibility to test different impact scenarios and policy designs before and after implementation, supporting both *ex ante* and *ex post* evaluation.

AI can potentially enhance understanding of why a programme works by enabling more efficient analysis of structured and unstructured data. For example, using Natural Language Processing-enabled text mining on policy documents allows the evaluator to more easily extract insights relevant for the evaluation. These tools reduce the time required for qualitative coding and can facilitate more consistent interpretation of evaluative data, offering rich information that can answer additional evaluation questions.

AI can also support different stages of randomised trials. For example, in clinical trials, AI could be used to select potential study participants without relying on costly manual review to predict the natural history of each participant, and to assess study end points in a data-driven method (Lee and Lee, 2020[42]). Large language models have been used to assess the trustworthiness of published randomised trials with some promising results (Au et al., 2025[43]). Indeed, when assessing trustworthiness in randomised trials (using the TRACT[8] checklist), results generated by ChatGPT and the human assessor had an 84% level of agreement (16/19) and substantially accelerated the qualitative assessment process (Au et al., 2025[43]).

Machine learning models are also being used to attempt to predict counterfactual outcomes in the absence of control groups, allowing analysts to simulate policy alternatives based on historical and synthetic data. An example can be drawn from recent evaluations of carbon pricing schemes, where machine learning algorithms were combined with economic modelling to estimate emissions reductions and cost implications in the absence of an *ex post* control group (Abrell, Kosch and Rausch, 2022[36]). Such models support quasi-experimental approaches such as matching, synthetic control, and instrumental variable estimation, improving the accuracy and policy relevance of evaluations conducted under data constraints.

Finally, AI can also support the dissemination and uptake of evidence by extracting information from multiple evaluations and reports. This can be done, for example, by developing chatbots that draw on quality reports (see Box 6 for French and Finnish examples). An example of such a chatbot is Artificial Intelligence for Development Analytics (AIDA), which was developed by the United Nations development Programme (UNDP), and using natural language processing to extract evidence from over 7 000 UNDP evaluation reports to reply to specific questions (Independent Evaluation Office, 2025[44]). The interest in AI for knowledge management in evaluation and evidence synthesis was also highlighted in a recent workshop developed by Wilton Park, the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom in partnership with Global Affairs Canada (Wilton Park, 2025[45]).

While AI can support the development of better evaluations and more robust evidence, governments face challenges in fully unlocking its potential. In particular, the practical use of these tools depends on the availability of high-quality data, as well as the capacity of evaluators to access and integrate AI methods into their workflows. Skewed or incomplete data (whether from the evaluation itself or from the data used to train the AI tool) can be a risk when using AI, as it can reinforce or exacerbate pre-existing outlooks,

[8] Trustworthiness in RAndomised Clinical Trials (TRACT) is the first checklist developed specifically to detect trustworthiness issues in RCTs. The checklist includes 19 items organised into seven domains that are applicable to every RCT: 1) Governance, 2) Author Group, 3) Plausibility of Intervention Usage, 4) Timeframe, 5) Drop-out Rates, 6) Baseline Characteristics, and 7) Outcomes (Mol et al., 2023[90]).

and make erroneous predictions. In some cases, the lack of transparency of certain AI tools can make it more difficult to understand and identify the rationale for AI-driven insights (OECD, 2025[35]). Government departments are also navigating the data privacy challenges of using open-access AI tools and the governance required to ensure safe use of government-held information.

In the same way that limited analytical skills can be an obstacle to conducting impact evaluations, skills gaps in AI can also present limitations for effective use and implementation of AI in government, which was a general finding of the overall OECD report (OECD, 2025[35]). In many cases, government employees do not feel equipped to use AI technologies. This can limit their ability to take advantage of the latest developments in AI, and may contribute to civil servants' reluctance to use AI to its full potential. Lack of understanding of when AI is best used can also lead human operators to over-rely on automation, leading them to accept results without scrutiny, in what is known as automation bias

Some OECD countries are developing AI hubs and Communities of Practice to better understand the role and implications of AI in government activities, with particular relevance to policy evaluation. For example, the United Kingdom has introduced an Incubator for Artificial Intelligence, and in France the data team in the General Inspectorate of Finances has established a team of experts to leverage AI tools for policy analysis (see Box 6). Similarly, structured guidance and training materials, as provided by the Australian Centre for Evaluation, may be relevant reference points for governments seeking to improve AI-supported evaluation practices. Finally, developing informal networks and communities of practice across government to share insights on AI and evaluation is an interesting practice. This is currently taking place in Canada where evaluation experts in different departments are meeting regularly to exchange their insights on use of AI for evaluation.

Box 6. Developing and leveraging AI for policy analysis in France and Finland

In 2019, the General Inspectorate of Finance (IGF), located in the Ministry of Finance in France, established a data science team composed of nine experts in econometrics and data science to expand the quantitative analysis of the Inspectorate. The team possesses strong skills in programming and leverage machine learning and large language models to conduct analysis and evaluations. In addition, it has access to confidential administrative data, allowing it to conduct robust evaluations that would not otherwise be possible. A recent example of AI use involved estimating the value of municipal real estate by leveraging machine learning models and private transaction data. In addition, the team built an in-house Retrieval Augmented Generation tool called Fragments which contains every IGF report since 2006 as well as the Supreme Audit Institution's (*Cour des Comptes)* reports. The tool can be used to ask questions on this knowledge base and provide answers that mobilise the vast amount of information in these reports.

In 2025, the Independent Development Evaluation Unit of the Ministry for Foreign Affairs developed *OpenEval.fi*, an AI-assisted tool that can be used to research information in evaluation reports on development policy and cooperation. OpenEval allows for better use of the information available on Finland's development co-operation and related issues, allowing both the general public and development experts to access information in a faster and effective way (OECD, 2025[46]).

For more information: (INSEE, 2025[47]; OECD, 2025[46]).

2.5. Supporting the translation of evaluation findings into service delivery and professional practice

To ensure that evaluation findings inform decision-making in government to make a positive contribution to citizens' lives, it is essential to embed them into the systems that shape policy, budgeting, and service delivery. This section explores how various OECD countries are strengthening the translation of evaluation results into concrete actions by ensuring a systematic approach to designing and conducting evaluations that are appropriate for intended users, and that are integrated into budget cycles.

2.5.1. Carrying out evaluation in a systematic manner

To inform policy delivery and practice, evaluations need to be carried out systematically and at a consistently high level of quality. For this to occur, the OECD Recommendation on Public Policy Evaluation recommends that countries embed evaluation into legal or policy frameworks (OECD, 2022[2]). This was done in Japan, for example, with the 2001 Government Policy Evaluations Act No. 86, the purpose of which was to promote the implementation of *ex ante* evaluation. *Ex post* evaluation was then incorporated into the legislation in 2017 (OECD, 2025[48]). This legislation provides an overview of what methods evaluators must use, as well as how they should be reflected in planning, budgeting, and inter-agency coordination. More recently, Spain implemented its own legislative framework, with Law 27/2022 aiming to institutionalise the use of evaluations. Like Japan's framework, it provides requirements that evaluations must be evidence-based and public and provide an overview of the intended evaluation process. Such legislative frameworks not only mandate the production of evaluations, they also provide guidance to ensure evaluations are conducted to similar standards across ministries (Government of Japan, 2001[49]) (Government of Spain Official State Gazette, 2022[50]).

While such legislative frameworks are helpful, further support is also needed to oversee the implementation of evaluations, ensure that they meet the required standards, and ensure that ministries have appropriate supports in place. In most OECD countries this role is performed by an institution which serves as an "evaluation champion". In the Netherlands, this is the responsibility of the Strategic Analysis Unit, established within the Ministry of Finance. This Unit oversees the entire evaluation cycle, and is responsible for providing advisory services, best practices, and guidelines to ministries, helping ensure that their evaluations are consistently of a high standard. Canada's Policy on Results requires that each government department establish an evaluation unit, thus providing a similar point of reference as the Strategic Analysis Unit at the departmental level, with the Treasury Board Secretariat taking on a cross departmental coordination function (OECD, 2025[3]). Similarly, in Portugal, PlanAPP, a centre of expertise for public policy evaluation, is responsible for coordinating an inter-ministerial network focused on foresight and evaluation practices (OECD, 2024[51]).

2.5.2. Ensuring evaluation evidence is useful and accepted by stakeholders

When designing an evaluation, the methodology selected should be commensurate with the task at hand. It is thus important to "right-size" evaluations to their policy context and ensure they are fit-for-purpose (OECD, 2010[26]). As different methodologies answer different questions, it is important to acknowledge the needs of the primary users in order to maximise the usefulness of evaluation results. Impact evaluations are useful for questions related to the causal effect of a programme in achieving its intended outcomes, but might not be the most suitable method if decision-makers are focused on other questions or if an impact evaluation would not be feasible or cost-effective (Peersman and Rogers, 2025[52]).

As highlighted in the OECD Recommendation on Public Policy Evaluation, governments should engage relevant stakeholders in the evaluation process from the outset. This can help ensure that any results are accepted by stakeholders, thus providing a higher degree of credibility and increasing its use. Such a process also helps create a sense of ownership, reducing the risk of negative or neutral results being

rejected or not utilised. A good example comes from Costa Rica, where the Ministry of Planning and Economic Policy has used a participatory approach to develop guidelines on stakeholder engagement and participatory evaluations in support of the evaluation work carried out by line ministries (OECD, 2022[2]) (OECD, 2025[53]).

Involvement of key stakeholders can be supported with structured and well-defined approaches that promote the independence and credibility of the final results. This involves first functional independence, where an evaluation team is able to understand the policy priorities and evidence gaps that will have implications for the objective and process of the evaluation without daily managerial interference. Second, it implies behavioural independence and neutrality, meaning the evaluators themselves do not have an entrenched conflict of interest. One solution for ensuring such independence is through outsourcing evaluations to external providers, as it reduces the risk of vested interests weighing in on the results. However, biases can also be present in such cases, particularly when considering the financial relationship between those commissioning and delivering the evaluation. Furthermore, external evaluators may have greater difficulties in having their evaluation results achieve impact as they are less familiar with policy processes and content (Boehmer, Fernandez Ordonez and Sharma, 2014[54]). As such, an effective middle ground can be to outsource evaluations but ensure that civil servants are actively involved in the process or to establish internal evaluation units with governance arrangements to achieve structural independence through separate reporting lines and oversight (OECD, 2020[4]). It is valuable to have ethical considerations included in the guidelines both for internal evaluations and the commissioning process of external evaluations.

Finally, once results are available, it is important to make sure that these are well communicated to a broad range of stakeholders and are owned by those involved in the evaluation process. While many evaluation experts are likely to be technically proficient, it is important that evaluation results are communicated in non-technical language so they can be easily understood by a wider range of stakeholders and political actors. Such efforts have been made by Spain's Independent Fiscal Institute, which developed an interactive tool known as the Observatory of Findings and Proposals, which aims to improve the availability of evidence, analysis and data to policymakers. The Observatory, which is accessible to the general public, includes a variety of published evaluations, their findings, and their subsequent proposals, as well as any monitoring carried out by the Ministry of Finance with regard to the implementation of these proposals. This allows policymakers to rapidly assess the evidence available for new propositions, and the extent to which existing propositions have been successful (IReF, 2023[55]).

2.5.3. Connecting evaluation with budgeting and resource allocation

Use of evaluations within the budget process is fairly common in the OECD, with nearly 70% of OECD countries incorporating policy evaluation findings into the budget cycle to some extent. However, the exact impact of such evaluations on budget decisions remains relatively limited. In order to achieve greater impact, governance mechanisms could be established that allow, or mandate, evaluations to be considered d as part of budget allocation decisions (OECD, 2020[4]).

Such a governance mechanism has been recently introduced in Italy, where, as of 2024, ministries must present their analysis and evaluation activities in an annual Evaluation Plan for the following three years. This document must be incorporated into the ministry's budget preparation, and aims to increase the value of public resources used while reducing inefficiencies. This aims for evaluations to become integral to budget allocation decisions. The requirement for annual updates of a three-year plan increases accountability, as evaluations must be planned in advance, and the removal of any such evaluations is likely to attract attention (OECD, 2025[53]). A similar initiative was implemented in Chile, where the Budget Directorate started to integrate spending reviews into the budget process by developing a dedicated methodology. The Ministry of Finance is directly involved in the development of spending reviews, and is

responsible for the link between spending reviews and the budget process (Ministry of Economy and Finance, 2024[56]) (OECD, 2024[57]).

Empowering ministries of finance with the evaluation process can be valuable, as they can be more receptive to negative evaluation findings than other ministries due to their financial and economic background and their attention to cost management, thus reducing the risk of the over-advocacy trap. Such an approach can be seen in Lithuania, where the Ministry of Finance, alongside the Office of Government, summarises the results of evaluations in preparation for budget negotiations. These summaries provide information on the progress achieved by the agency since the evaluation, as well as any implementation gaps (OECD, 2020[4]). Ministries of finance (or similar institutions) can also play a role in incorporating evaluations into the budget process by attaching evaluation requirements to funding. For example, in Canada, the Treasury Board, responsible for the operational management of government spending and administration, requires that its spending allocation decisions are informed by evaluation findings. As such, if a government organisation wishes to obtain expenditure authority for a policy or programme, it must state whether an evaluation has been conducted (and share its results if so) and whether the head of evaluation or performance measurement has been consulted in the development of the policy.

3 Delivering high-quality impact evaluations

This chapter discusses how governments can best deliver high-quality impact evaluations. This includes some of the key impact evaluation methods and how they identify the causal effect of a programme, how to keep impact evaluations cost-effective, and responding to stakeholder concerns about whether randomised trials are ethical, drawing in part on the February 2025 workshop discussions.

3.1. How the key impact evaluation methods work, and signs they may be feasible

Once there has been agreement to run an impact evaluation, the question becomes which approach and method to choose so that it is able to effectively generate relevant policy insights. This section discusses some of the main methodologies available to this end, considering both experimental and quasi-experimental methods. A key theme in this discussion is how each evaluation method addresses the need for a counterfactual – i.e. an estimate of what would have happened in the absence of an intervention. (For an overview of experimental and quasi-experimental methods, see the Magenta Book, United Kingdom Government guidance for policy evaluation, Annex A).

There are various advantages and disadvantages of a given impact evaluation method with the most appropriate method always depending on the context and the purpose of the evaluation. Timing is also critical – for example, randomised trials are very rigorous but must generally be rolled out before a programme is implemented. On the other hand, quasi-experimental methods can often be implemented after a programme has already been rolled out if the required data is available.

Table 1. Summary of empirical approaches

Category	Impact evaluation method	Signs that a method might be feasible
Experimental designs	Randomised controlled trials (RCTs)	It would be possible to randomise which individuals/schools/classrooms/businesses participate in a programme, and which do not
Quasi-experimental designs	Found or 'natural' experiments	There is a natural source of randomness in the world, e.g., a lottery
	Regression Discontinuity Design	There is a cut-off in eligibility, e.g., a selective school test
	Difference-in-Differences	There is a group or region that didn't receive the programme, and there is data on outcomes before and after the programme was introduced
	Instrumental variables	There is a variable that affects likelihood of treatment (i.e., programme participation) but does not affect the programme's outcomes in any other way
Control-on-observables designs	Matching	We understand the underlying phenomena well, can be confident that observed data explain the underlying dynamics, so that any unobserved data (e.g., motivation, emotional wellbeing) that affects programme participation are irrelevant to its outcomes
	Regression analysis	
	Time series event study	

3.1.1. Randomised trials

How randomised trials work

Randomised trials involve assigning some individuals, school classrooms, regions or businesses to participate in a programme (the 'treatment group') and others to not participate (the 'control group'). Since participation is determined by randomisation, the analyst can be confident that before the trial begins, the treatment and control groups are likely to be extremely similar in both observable characteristics (e.g., age, income, education level), and also unobservable characteristics (e.g., motivation level, support network). Because the groups are almost identical before the trial commences, any difference in outcomes between the two groups is due to the 'causal effect' of the programme itself. For guidance on randomised trials, see the Abdul Latif Jameel Poverty Action Lab's (J-PAL) online guidance (Gibson and Sautmann, 2023[58])

Signs that a randomised trial may be possible

An advantage of randomised trials is that they do not rely on a counterfactual comparison group happening to exist by chance. Instead, the programme administrator or researcher creates a counterfactual group using randomisation. They are therefore feasible where it is possible to randomise which programme recipients receive a programme and which do not. This randomisation could be at the level of individuals, schools, households, businesses, or regions.

Some stakeholders worry that deliberately excluding some people from participating in a programme could be unethical. This is an understandable concern and can sometimes be grounds not to proceed with randomisation.

While not always necessary, some common randomised trial designs that can help mitigate ethical concerns to randomisation include using a:

- **Waitlist control design:** This is useful when a programme is new. Everyone who would like to participate in the programme will have the opportunity to do so but some people might be randomised to come off the waitlist earlier than others.
- **Staggered rollout:** This is useful when a programme is new (or being expanded) and being rolled out sequentially to some regions or areas before others. In this case, the order of rollout could be determined by randomisation.
- **Oversubscription design:** In cases where a programme is already oversubscribed, some participants will already have to miss out because there is insufficient funding to meet demand. This can be an opportunity to randomise who receives the programme and who misses out.

For a full discussion of when a randomised trial may be ethical, see the section below, "Overcoming ethical concerns to randomised trials".

3.1.2. Regression discontinuity

How regression discontinuity works

Regression discontinuity studies are a quasi-experimental evaluation option possible in cases where there is a cut-off in eligibility for a programme or policy on a continuous variable. They use this cut-off to statistically compare people who are just below and just above the cut-off in order to find the effect of the programme or policy on a different outcome of interest. The threshold can cover a wide range of policy rules, including students with an entrance exam mark above a threshold getting admitted to a selective

school, families with a poverty index below a certain level getting access to certain government supports, or people above (or below) a certain age achieving eligibility for a programme.[9]

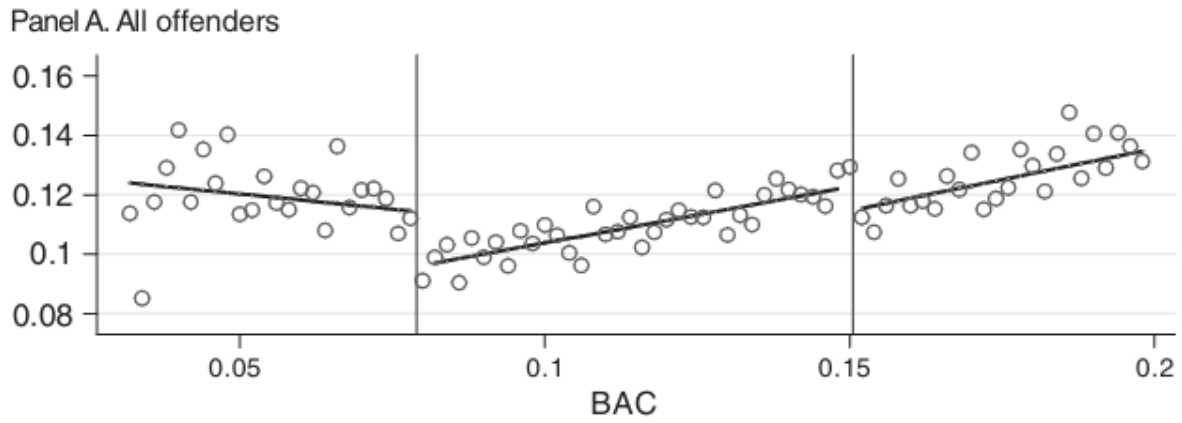
For example, Hansen conducted a regression discontinuity study to evaluate whether punishments for drink-driving are effective at reducing the chances that an individual who is caught is likely to drive drunk again (Hansen, 2015[59]). He used the continuous variable of a person's blood alcohol content level (BAC) when they are stopped for a drink-driving test in Washington state in the United States. In Washington, there is a policy cut-off of 0.08 where people receive fines if their BAC is above this level but do not if it is below (Hansen, 2015[60]).

Regression discontinuity uses the fact that people whose BAC levels are extremely close when they are stopped for a drink driving test are likely to be very similar on average in terms of relevant characteristics like how likely they are to break the law, and how safe they are at driving. This means that for people who are extremely close to the BAC level of 0.08, the fact some are punished, and others are not can be seen as being 'as good as random' with respect to other causal variables. Therefore, if there is a large difference in the likelihood of someone re-committing a drink driving offence after being stopped for a drink driving test, it is likely due to the effects of being punished rather than due to other factors.

The chart below illustrates Hansen's main result. The x-axis is a person's blood alcohol content level when they are initially stopped for a drink driving test, while the y-axis is the likelihood of the same individual re-committing a drink driving offense. As the chart shows, people whose BAC is just above the 0.08 level are substantially less likely to re-commit a drink driving offense, which suggests the fines and punishments Washington state has in place for discouraging drink driving have been effective at achieving that outcome.

Hansen also conducted a regression discontinuity study of the effects of having a BAC of 0.15 because, BAC levels above this cut-off receive even higher punishments. Again, people just above the 0.15 level are less likely to re-commit a drink driving offense, which again suggests the existing system of punishments are effective.

Figure 2. Effect of being over blood alcohol limit at drink driving test on likelihood of drink driving again in future



Source: (Hansen, 2015[59]). The x-axis is a person's blood alcohol content level when they are initially stopped for a drink driving test. The y-axis is the likelihood of the same individual re-committing a drink driving offense.

[9] For greater detail see Fougere and Jacquemet (2023[92])

Signs that a regression discontinuity might be possible

Regression discontinuity designs require a cut-off that changes the probability of individuals participating in a programme or being affected by a programme. Cut-offs can come in a remarkably wide range of areas including student scores, geographical proximity, a poverty index, or measures of health. Examples include:

- Students that score above a certain mark are eligible to enrol in a selective school (Lucas and Mbiti, 2014[61])
- A household poverty index score where people below a certain level are given extra supports (Filmer and Schady, 2009[62])
- Politicians that are elected to parliament when they receive above a certain share of votes (Eggers et al., 2014[63])
- Clinical guidelines state that people with white blood cells below a certain count receive HIV treatment (Venkataramani, Bor and Jena, 2016[64])
- Children of a certain age or school year receive a vaccine (Venkataramani, Bor and Jena, 2016[64]).
- Blood alcohol content levels above 0.08 are punished, while those below are not (Hansen 2015)

3.1.3. Difference-in-differences

How difference-in-differences works

Difference-in-differences involves comparing changes over time between a treatment group and a comparison group. For example, if one state in the United States increased the minimum wage while a nearby state with a near identical economy did not increase the minimum wage, the two states could be compared to estimate the effect of the change in the minimum wage on employment (Card and Krueger, 1994[65]). These methods became very popular during the late 1990s as econometric tools made progress, particularly as they can be used in a powerful way with administrative data. (See (Fougere and Jacquemet, 2023[66])

The difference-in-differences design relies on the 'parallel trends' assumption. That is, even though the level of the outcome variable may be different in the treatment and control group, the changes from year to year should be the same. Thus, the outcome variables for both groups move 'in parallel'. In the United States example, the method requires us to assume that if it were not for the policy change, the employment rates of each state would have increased or decreased by the same amount. It is not possible to directly verify this assumption, although it is possible to gather data on past or future trends that may make it seem more plausible.

Signs that a differences-in-differences may be possible

Differences-in-differences studies may be possible when:

- data is available before and after a policy change (or a program) is introduced,
- there is a comparison group that did not receive the program, and it is plausible that the trends for the comparison group are similar to those for the treatment group.

3.1.4. Control-on-observables methods (including regression, matching, and modelling)

How control-on-observables work

Control-on-observables methods include statistical regression, matching, synthetic controls, econometric modelling, and time series event studies. They typically involve statistically adjusting for observable differences between people who received the programme and people who did not. For example, if an

impact evaluation was examining the effect of participating in a job training programme on getting a job, the evaluators could statistically control for potential 'confounding variables' they have data on such as previous employment history, education level, age, race and gender. Time series event studies can also fall into this category but usually use aggregated time-series data rather than individual data. They use econometric modelling techniques to forecast a counterfactual for what would have occurred without a policy change.

The disadvantage of control-on-observables methods is that they require making the strong assumption that any unobserved variables (that the evaluator does not have data on) that may have influenced programme participation are irrelevant to the outcomes of interest. This is often a hard assumption to justify because it is often difficult to get data on many things that affect people's behaviour. In the example of the job training programme above, the following factors may all have some effect on an individual's likelihood of getting a job (but not be observed in data): their level of motivation, whether they have a supportive family, their professional network, and their day-to-day mental and physical health.

An advantage of control-on-observables methods is that they are often feasible to implement on existing data, including government administrative data. Rich administrative data is population-wide, which permits granular distributional and sub-group analysis (OECD/EC, 2025[7]), and largely removes the risk of attrition or recall bias that can create issues for randomised trials (OECD/EC, 2025[7]). For example, administrative data has been used to conduct counterfactual-based impact evaluation of various active labour market policies across several OECD countries (OECD/EC, 2025[7]).

Signs control-on-observables may be possible

Control-on-observables methods are possible when rich, detailed data are available on both people who have received a programme and people who have not. The method is more reliable in situations where the 'data generating process' in the real-world is very well understood, and there is high-quality data on all important factors in the data generating process. This will usually mean there is already a sophisticated theory describing how people end up participating in the programme, and how the factors that influence participation decisions could also influence outcomes of interest.

3.2. Making the best of the design phase

3.2.1. Establish a programme logic or logical framework and determine the evaluation questions

The first step in conducting an impact evaluation of a programme is often to clarify how the programme works, and how it is expected to achieve its intended outcomes. A common and effective way to do this is to create a 'programme logic' or 'logical framework'.

The Logical Framework Approach supports the design of an intervention. Its main output, the Logical Framework Matrix, summarises in a single framework the main characteristics and specifications of the intervention, including measurement indication, becoming a relevant tool of the monitoring and evaluation process. A logic model helps with evaluation by setting out the relationships and assumptions between what a programme will do and what outputs and outcomes it expects to deliver (Hayes, Parchman and Howard, 2011[67])

Once a logic model has been established, the next important challenge is to determine how the results of the evaluation will be used and what information about outcomes would be useful for decision makers. For this, the key is to determine the question that the evaluation will seek to answer to ensure its relevance, feasibility and impact. In a policy context, these questions should relate closely to the decision that policy makers will need to make.

There are various different types of evaluation, each answering different questions. A common grouping of the different evaluation methods is process evaluation (or implementation evaluation), impact evaluation (or outcome evaluation), and economic evaluation (or value-for-money evaluation). See Box 1 for a definition of the different methods. All these methods can be useful depending on the purpose of the evaluation. Impact evaluations tend to focus on a question along the lines: 'Does this policy or programme achieve the outcomes it was intended to achieve?' or similarly, 'What was the causal effect of the programme on the outcomes of interest?'

Note that the term 'impact evaluation' sometimes refers explicitly to counterfactual-based methods and sometimes does not. The focus of this report are counterfactual-based methods though the authors do not object to non-counterfactual methods also being described as impact evaluations.[10]

3.2.2. Preliminary checks to ensure feasibility

Once an impact evaluation method has been identified, the next step is determining if it is feasible. This will depend on the specific method, but the following three considerations are almost always relevant:

- **Data on the outcome of interest:** For an impact evaluation to be possible, there must be existing data – or it must be possible to collect new data – on the outcome of interest. Furthermore, for some impact evaluation methods, it is necessary to collect outcome data for a group that did not receive the policy or programme, in addition to those who did. It is also important to check the relationship of the outcome of interest with the priorities and understanding of decision-makers, so that results can be fully useful.
- **Sufficient sample size:** Any impact evaluation will need a sufficient sample size to generate precise results, particularly considering specific categories of interest. Researchers should conduct a 'statistical power' analysis, which determines the sample size needed for a particular minimum detectable effect (MDE), if the programme has such an effect. The MDE is the smallest effect size that the evaluation is likely to be able to draw meaningful conclusions about. It should ideally be set at the minimum effect size of policy interest. For example, if an education programme was considered not to be worth the effort if it results in less than a 3% improvement in test scores, then the MDE might be set at 3%. The power analysis might then conclude: 'In order to have a high likelihood of detecting a MDE of 3%, if there is one, the evaluation will need a minimum of 1 000 students in the sample.'

  Usually, issues of sample size will not be a problem when relying on administrative data, but when relying on survey data, or specific experimental design, or trial, it is an important consideration, given the costs of generating and accessing the data.
- **Feasibility of establishing a credible counterfactual:** In order to accurately separate causation from correlation, impact evaluation methods require a 'counterfactual' comparison group. For example, in randomised trials, the counterfactual is created by randomising some people to be in the 'control group' where they do not receive the programme or receive a different version of the program. In this case, a randomised trial will be possible if it is feasible and ethical to randomise some people to not receive the programme (see below for a full discussion of when it is ethical to randomise). Other quantitative impact evaluation methods rely on the counterfactual group already existing by chance. There are also qualitative impact evaluation methods that use theory to establish the causal effect of a policy.

[10] The Glossary of key terms in evaluation and results based management prepared under the auspices of the OECD Development Assistance Committee also provides some useful entry points.

3.2.3. Ensuring impact evaluations are rigorous

There are several things the impact evaluation community can do to help ensure their evaluations are rigorous:

- **Publicly register a pre-analysis plan before examining the data:** A pre-analysis plan (or 'trial protocol') outlines in detail how the trial data will be analysed. Registering this plan before looking at the trial data helps prevent 'p-hacking', where the researcher analyses the data in various different ways in order to find one way where results are statistically significant. There are various websites where pre-analysis plans can be pre-registered, including the American Economic Association's RCT Registry or the Open Science Framework or the Registry for International Development Impact Evaluation. The Australian Centre for Evaluation offers a template for creating pre-analysis plans.[11] During the workshop, Prof. Dan Levy highlighted the risks associated with such "p-hacking" and offered practical ways to address them.
- **Engage with peers to review the design and results of evaluation**: Peer review represents an essential way to ensure the quality of all evaluations, including randomised trials. Using a 2-stage model of peer review (like the Registered Report) represents a good way to ensure the quality of evaluation design. In the first stage the quality of the study design is assessed before the data is even collected (emulating the pre-analysis plan described above). During the second stage the reviewer compares the final analysis to the original plan, ensuring all planned analyses are reported, and assesses whether any deviations from the original plan were justified (Chambers and Tzavella, 2021[68]).
- **Where possible provide data to replicate the analysis**: To promote transparency and reproducibility, impact evaluations should, where possible, make their data and code publicly available. If privacy or ethical concerns prevent this, researchers might consider using synthetic datasets that replicate key characteristics of the original data. Additionally, sufficient detail about the intervention should be provided to enable replication in other settings or populations. This is essential from an Open data and Open Science perspective.

3.2.4. Balancing rigor with feasibility

When designing an impact evaluation for policy making purposes, it is critical to balance technical rigor with feasibility. This is relevant to many areas of evaluation design, but a simple example is the choice of statistical significance threshold. In academic research, the standard statistical significance threshold of interest is 5%. That is, results must have a 5% or less probability of being a false positive.

For impact evaluations to inform a policy decision, a decision will often need to be made based on the best evidence available, and there will often be a significant time constraint. In this context, a less stringent alpha (e.g., 10%) may be adequate.

Further, if a decision needs to be made without time for further evaluation, policy makers will need to make a judgement based on the balance of evidence, and may use results that do not meet the usual significance thresholds as one input among many into a decision.

3.3. Making impact evaluations cost-effective

Lack of funding may at times be given as a reason for not implementing or commissioning evaluations. However, rigorous impact evaluations do not have to be expensive, and lower-cost approaches are

11 Available at: https://evaluation.treasury.gov.au/toolkit/preregistration-and-pre-analysis-plans.

feasible in many circumstances. The Arnold Foundation in the United States has run 'low-cost RCT competitions' since 2013 where they have provided grants of USD 100 000, USD 150 000 or USD 300 000 to randomised trials for building high-quality policy evidence (Laura and John Arnold Foundation, 2018[69]). Similarly, the Paul Ramsay Foundation in Australia have provided grants of AUD 200 000 (approximately USD 130 000) for non-profits to evaluate the effectiveness of their programmes with a randomised trial (Paul Ramsay Foundation, 2024[70]). Such amounts may be considered as acceptable for middle to large OECD countries, even if for the smaller countries, both financial resources and availability of proper human expertise might be an issue. This section will therefore discuss the resource associated with randomised trials – data, IT, staff and sample size – and how they can be managed.

It is important to distinguish the costs of an evaluation with the costs of delivering a programme. In an academic setting, the cost of the evaluation will often include the cost of delivering the programme itself because the programme would not have been delivered if it was not for the research project. In a policy context, however, the programme has usually already been funded. The costs of the impact evaluation are only the incremental or additional costs associated with evaluation such as staffing, data collection, and any administrative or IT costs required to deliver the evaluation.

Excluding programme delivery costs, the most significant costs for an impact evaluation are often those associated with data collection. In a policy context, there might already be high quality programme data available, which will mean the costs of additional data collection can be avoided. If this is not the case, other administrative data, which are distinct from data collected for the programme, can often help. Administrative data can help keep costs down by removing the need to collect new data. It can also save costs by providing new control variables that can improve 'statistical power'. For example, in a recent randomised trial, the Australian Centre for Evaluation estimated that access to high quality linked administrative data could improve the statistical power of the design by up to 20%. In situations where recruiting participants or delivering an intervention is costly, this would represent a substantial cost saving.

The randomisation stage of a randomised trial often has no cost if the evaluation team is familiar with standard statistical programmes that can randomise participants into the treatment and control groups. However, in some cases an IT 'build' will be required to set up the randomisation infrastructure for the trial. This can have a significant cost, especially when working with large government platforms – for example, for employment services, social security payments, or tax administration. While the initial cost of IT builds can be significant, the randomisation infrastructure can also often be re-used for future trials. Building IT systems with the capacity to run randomisations from the beginning is another way to mitigate the costs of a randomised trial as incorporating these features later can often cost more.

While academic partnerships can be of significant benefit to an impact evaluation, Dan Levy acknowledged during the 2025 Workshop that the incentives of academics and policy makers are not always well aligned (Levy, 2025[71]). Academics want findings that are novel and also generalisable across different contexts. Policy makers want to answer a specific question or questions to inform a decision, so want timely results and are not primarily interested in whether results are novel or generalisable. If an impact evaluation is being conducted for the purpose of building policy evidence, priority should be given to the interests of policy makers, while also trying support any academic partners where possible.

Another way to manage the costs of an impact evaluation is to focus on study designs that have a higher level of statistical power as these will also generally have smaller sample size needs. The statistical power of an impact evaluation can be kept high by being disciplined in focusing on the evaluation question that is of primary interest, rather than questions that are of secondary interest (having multiple treatment arms, and having multiple hypothesis tests both decrease power). Power will also be kept higher by avoiding research questions that seek to identify very small effects that are of no practical significance. To illustrate this, consider as an example, a programme that is intended to increase student test scores. More power (and hence a larger sample size) would be needed to detect an increase in test scores of say 0.3% than would be needed to detect an increase in test scores of 3%.

McKenzie (2023[72]) outlines several further strategies for increasing the statistical power of a randomised trial, including avoiding imbalance between the treatment and control groups, measuring outcomes of interest accurately, collecting observations of participants at multiple points in time, and using outcomes of interest closer in the causal chain. Haskins and Feldman (2016[73]) provide some further suggestions for how to implement low-cost randomised controlled trials.

The costs of a randomised trial vary widely. For some randomised trials (e.g., behavioural insights trials) the only cost beyond team time will be the costs of an ethics submission which is generally free for researchers affiliated with a university. On the other hand, large, complex randomised trials can cost several million dollars – for example, the evaluation of *eHealth intervention in Australia* (Champion K. and Emma, 2023[74]).

While there can be a significant cost to conducting rigorous impact evaluations, there are also many ways to minimise these costs. Even in cases where high-quality evidence relevant for policy is expensive, it is important to remember that high-quality evidence is also valuable, as policy mistakes can also be very costly. This is true both for achieving policy goals, but also from a financial point of view. In OECD countries, the cost of impact evaluations will generally be a small share of the costs of delivering a programme. If a high-quality impact evaluation finds that the programme is not achieving what it is intended to achieve or is not as effective as it could be, there could be significant cost savings from improving or discontinuing the programme. Similarly, if the evaluation establishes that a programme is effective, it can ensure the longevity of a programme that has important impacts on well-being.

3.4. Overcoming ethical concerns to randomised trials

A potential barrier to the use of randomised trials is that some stakeholders may worry that it is not ethical to exclude the control group from receiving the intervention. This can sometimes prevent actual experimentation, due to rules for equal access to public services. For example, France had to change its constitution in 2003 to allow for experimentation in public services, through the possibility of introducing experimental clauses in laws and regulations. A complementary clause was also introduced allowing the use of experimentation by local governments. This has allowed for experimentations to be conducted in a large majority of public services.[12]

However, limiting the creation of evidence needed to inform a policy decision can also expose citizens to a significant welfare loss. Therefore, governments, NGOs and university ethics committees are increasingly viewing policy randomised trials as not just ethically acceptable, but in some cases ethically preferable. The value of using randomisation is outlined in this section.

3.4.1. The ethical justification for using randomisation and control groups

The ethical consequences of not using a well-designed randomised trial—when one would be feasible—can be serious

In many cases where a well-designed randomised trial is feasible, it will be the most rigorous way of determining how effective a programme is at achieving its objectives. Quasi-experimental and control-on-observable methods can be used as an alternative, but these tend to rely on assumptions that can be difficult to verify, and so are not always quite as rigorous. This can have serious implications on the lives of the people intended to benefit from the program.

[12] See: https://www.vie-publique.fr/fiches/20114-en-quoi-consiste-lexperimentation-legislative-locale

The story of Frances Kelsey (who worked at the United States Food and Drug Administration in the 1960s) illustrates the consequences of a less rigorous evaluation. Ms Kelsey was responsible for approving a new drug that had been approved in Canada and over 20 countries in Europe and Africa. The new drug was helping thousands of pregnant women with morning sickness. She insisted on wanting to see the results of a randomised trial before approving, which was not always required at the time. She was put under significant pressure, with arguments along the lines of her needlessly causing thousands of women to suffer and being heartless in her insistence on wanting to see data on a drug that had already been approved in many other countries. The drug Frances Kelsey insisted on seeing randomised trial data for was Thalidomide. She saved thousands of American children from being born with deformities (Rouhi, 2005[75]).

It can be tempting to dismiss lessons from medicine as not relevant to the evaluation of other kinds of policies or programs. There are, after all, almost no social policy programs whose negative consequences could be as bad as Thalidomide, and it is certainly less likely a non-medical programme could have as negative consequences as Thalidomide. However, it is entirely possible that a non-medical programme does not achieve the intended outcomes or leads to unintended outcomes. Adopting an impact evaluation method that is less rigorous and less likely to be able to discover shortcomings in a programme is also ethically problematic if it means improvements to the programme are not made and the beneficiaries of the programme consequently do not receive any benefit. This is especially true of government programs that may exist for many decades, and where money could be redirected to programs that are more effective at supporting the intended beneficiaries.

There might be genuine uncertainty about whether a programme is effective

Even when a programme seems promising or is based on a sound theory, it is not always clear that it will work in practice. Resolving this uncertainty by testing the programme in a rigorous way can be a good reason for conducting a randomised trial. If there is no solid evidence showing that a programme is effective, then it's not necessarily the case that people who are randomised to the control group are disadvantaged. In fact, if the programme turns out to be ineffective, or even harmful, then not receiving it might actually save participants time, effort, or inconvenience.

The history of randomised trials has shown that certainty about a program's effectiveness cannot be guaranteed by expert opinion, robust theory, or even prior evidence. History has shown that programmes that seem like common sense, can actually have no effect or even negative effects (Rehill et al., 2025[76])

In other circumstances, there could be evidence of a programme's effectiveness in a small, controlled study but not evidence of its effectiveness at scale or in a different context. Here, uncertainty can also arise from potential externalities, where a programme may appear beneficial when considered in isolation but its broader impact on non-participants and the wider system may be less clear from a general equilibrium perspective. Uncertainty could also remain around what the optimal programme parameters should be (e.g., the exact support model) or who the programme should be targeted at.

Even if a programme is known to have a positive effect, uncertainty might remain about its cost effectiveness

For example, Goldfeld et al. (2022[77]) conducted a rigorous randomised trial of a programme where women experiencing adversity received 25 home visits from a paediatric nurse and additional support from a social worker during pregnancy and the first two years of childhood. As a programme that provided relatively intensive support with a well-established treatment (paediatric nursing care), it could be argued that it was very likely the programme had some positive effect, and that the randomised trial was therefore not necessary. However, given the intensiveness of the support, it is plausible that uncertainty remained about the cost-effectiveness of the program. This is particularly true given any funding provided to this programme would not be able to fund other programs intended to support mothers experiencing adversity.

In short, randomised trials can be used to determine which of several programmes are most cost-effective and hence deserving of further funding. However, issues of implementation will have to be taken into account, as the effects on specific populations may vary.

Rigorous evidence can benefit many people both in the present and in the future.

It can be helpful to consider the number of beneficiaries of a rigorous impact evaluation. A randomised trial is not conducted primarily to benefit the people in the treatment group, or the people in the control group—which in most trials will together be a couple of hundred or a thousand people. Instead, the randomised trial is conducted to benefit the entire population of people who are likely to benefit from the evidence that is generated. This group is much larger. It would include all present and future participants in the program. The ethics of randomisation must also take account of the potential long-term benefits of putting future policy decision-making on a surer footing.

Randomisation can in some cases be a practical solution to a logistical challenge like limited funding or phased roll-out

In cases of limited resources or phased implementation, not everyone can receive the programme at one time. Sometimes a programme doesn't have enough funding to reach all eligible participants, or it needs to be rolled out gradually. In these situations, some people will inevitably miss out, regardless of whether a trial is being run. When this is the case, randomisation can be a fair way to allocate access. For example, an oversubscription design randomly selects participants from a pool of applicants to decide who gets to access the program. This can be seen as fairer than other methods of selection, which may rely on subjective judgments or arbitrary criteria while also providing the opportunity to learn about impact throughout the implementation. (Braga et al., 2025[78])

Practical ways to mitigate concerns about excluding the control group

In addition, the implementation approach can help mitigate concerns about randomisation. There are several ways that a randomised trial can be designed to minimise ethical concerns that participants will miss out on the hoped-for benefits of a policy or program. These include:

Many randomised trials use informed consent to give participants themselves the right to choose whether they would like to participate.

First, participants themselves have the opportunity to weigh the potential benefits and risks of being in the treatment group with the potential disappointment of being in the control group. They can then come to their own conclusion about whether they would like to participate. In some cases, a randomised trial can be conducted in an ethical way without informed consent if, for example, it would not be practical to obtain informed consent. But in general, practitioners of randomised trials take the approach of using informed consent unless there is a compelling reason not to do so.[13]

The control group can be provided with other services such as the current best standard of care available

For a new programme, the status quo may be just as good or better than the new programme which has not yet been evaluated. So the control group could be provided with the current best standard of care.

[13] See JPAL resource on ethical conduct which is itself derived from (Glennester R. and Power, 2019[79])

There are also other ways to mitigate ethical concerns about randomisation. For example, instead of randomising over the whole population, it might be possible to provide the treatment to the people that need it the most first, and then randomise for the rest of the population.

Randomisation doesn't always mean permanent exclusion

Sometimes, participants can be randomised to receive the programme later, rather than not at all. For example, in a 'waitlist control design', people are randomly selected to come off the waitlist in stages, meaning that everyone eventually gets access. This approach can work well when rollout takes time or when funding is released over several years. In other cases, a waitlist might be created specifically for the randomised trial. Similarly, a stepped wedge design involves rolling out the programme to different regions at different times. If some areas receive the programme earlier than others due to random assignment, this also constitutes a form of randomised trial. For example, a government programme in Australia involved providing energy upgrades to vulnerable, older people. Because the upgrades took time to install, it was not possible to roll the programme out to all eligible households immediately. So, in order to gather rigorous evidence, the order in which regions received the upgrades was randomised resulting in a stepped wedge randomised trial.

Encouragement designs can reduce ethical concerns about exclusion

An encouragement design is a type of randomised trial where all individuals are allowed to access the program, but only some are randomly selected to receive targeted outreach, such as emails or social media ads, encouraging them to participate. This way, no one is denied access, but the trial still tests the impact of the intervention by comparing outcomes between those who were encouraged and those who were not. The trade-off is that this approach usually requires a larger sample size. (Glennester R. and Power, 2019[79])

Engaging stakeholders throughout the evaluation is essential to achieve a shared understanding of the need for randomisation

Stakeholders might be sceptical about randomisation even when there is a clear rationale for it. For this reason, it is essential to explain the rationale for using a randomised trial and engage with any concerns community members may have. A comprehensive literature study highlights the value of engaging stakeholders in research to achieve impact. (Boaz et al., 2018[80]). Some recent Australian evidence suggests the public has a surprisingly positive view of randomised trials when it is explained to them. In a survey of the Australian general public, (Biddle, Gray and Hiscox, 2022[81]) found that after having the concept of randomised trials explained to them, 83% of people supported or strongly supported their use (Rehill et al., 2025[76])

3.5. Helping people, rather than protecting policies and programmes

Another obstacle to rigorous impact evaluations being conducted can be concerns among programme teams and stakeholders that the results might be disappointing, and that a programme that stakeholders know well and are passionate about is not as effective as they had hoped. By committing rigorous impact evaluation methods, both programme administrators and evaluators lose any ability to guide the evaluation toward the results they would like to see.

At the workshop, Professor Dan Levy explained he had encountered this issue several times before. He argued that the solution to is "to be an advocate for the people that you're trying to help, not for the programmes you're using to help them." This advice applies not only to external evaluators but also to programme administrators, community stakeholders, and in particular evaluators working within

implementing agencies. In such contexts, evaluation results can directly impact the legitimacy and even the existence of such programmes, creating strong incentives for staff to design evaluations, identify research questions, or select methods that portray the programme in a favourable light. As highlighted in Ravallion (2009[82]), it is much more difficult to publish a paper that reports unexpected or ambiguous impacts, and the wider knowledge system tends to privilege findings that confirm prior expectations. Ravallion also points out that managers of weaker programmes often have incentives to avoid evaluations altogether, since credible evidence of poor performance could threaten their activities.

One example demonstrating this issue, highlighted during the workshop, was the evaluation of a school construction programme. A rigorous impact evaluation found it had little to no effect on the three outcomes policymakers were hoping it would help address: school enrolment, attendance and learning. Faced with such results, a constructive approach would be to treat such null or negative findings not as failures, but as feedback. Such a case is made by Pritchett, Samji and Hammer (2013[83]), who emphasise that while a given design may not have worked in the particular context, it does not mean the idea itself has failed, but rather than there is need to adjust the design or implementation. They introduce the idea of structured experiential learning, which includes the need to create legitimate spaces for failure. It can be tempting, however, for policymakers or school principals to dismiss the results of the evaluation. Dismissing such results misses an opportunity to improve the effectiveness of the programme.

Disappointing results require evaluators to stand by their findings, and to be very clear about what did and did not work and the implications. This may also call for some key quality standards in policy evaluation, as called upon by the OECD Recommendation on Policy Evaluation and also developed by the OECD Quality Standards for Development Evaluation. This includes promoting the functional autonomy of evaluations, safeguarding both the final output and the evaluation process.

There is also often a fear that if a programme is discontinued, an alternative might not be readily offered to the people benefitting from it, leaving them with no good option. A specific focus on a given programme or policy may also prevent the sector as a whole being considered, where people might be served by competing or complementary programmes and policy makers need evaluations to offer a broader perspective to effectively feed into the policy cycle (Department of Planning Monitoring and Evaluation, South Africa, 2019[84]).

Another way to manage disappointment with the results of an evaluation, is to communicate clearly and appropriately. This might mean giving benchmarks for the effect sizes of other similar programmes. For example, an effect size of say 5% for a particular programme might seem small to some stakeholders. However, if stakeholders are first told that most similar programmes had effect sizes of 1-3%, then they might understand than effect size of 5% would be something to celebrate.

Discussions among participants during the workshop helped to identify some opportunities for evaluators to overcome this challenge. For example, the following options might be explored:

- Developing a theory of change that is focused on the individuals participating in the programme .
- Being clear from the beginning that the evaluation is about understanding what works for people, not protecting programmes, and making sure evaluation questions reflect this people and outcomes-centred approach.

Professional evaluators are often trained in developing credible theories of change. To support this, it can be addressed as part of a challenge function in the review process to ensure a clear, people-focused narrative. This should be connected to the evaluation questions. A related important factor is the use of participatory methods, which involve a wider group of stakeholders in the process. This helps avoid a purely top-down approach and can give a voice to the clients of the programme, ensuring that the evaluation really reflects the needs and experiences of the people affected.

3.6. Leveraging the value of mixed methods for impact evaluation

A criticism sometimes made of impact evaluations is that they over-emphasise quantitative information, which can tell you about what happened, but does not explain how changed was achieved.

In addition, focusing too heavily on quantitative findings may neglect individuals' lived experience, or participants' own views of whether they benefitted from receiving a programme. A related criticism is that impact evaluations sometimes over-emphasise the average effect of a programme, when the effect on a particular group-such as the most disadvantaged-might be of primary interest. These risks are genuine and are true of some evaluations, however properly implemented impact evaluations need not suffer from either of these issues.

A mixed methods approach is a good way to address these concerns, as it means incorporating qualitative research, for example from focus groups, surveys or interviews with participants or other stakeholders (Pluye, 2023[85]). Overall, a mixed methods approach is more likely to provide actionable insights for decision-makers as it incorporates qualitative and qualitative research to answer a broader range of questions. By combining both methods, evaluation can answer not only if the programme is effective but also why, for whom, and under what circumstances. Such information will be valuable for any future improvement of the programme.

There are also specific actions researchers can take to avoid their impact evaluation over-emphasising the average outcome. They can clarify with policy makers from the beginning whether the effects of the programme on a particular group are of primary interest and focus the evaluation on that cohort. They can ensure sufficient sample size in the cohort of interest using stratified randomisation in order to accurately calculate effects for that cohort.

For example, a 7-year randomised trial evaluation of the ‘Early Head Start’ programme in the United States not only found that the programme improved children’s learning on average. It also examined the impact for 27 different sub-groups including maternal age, birth order, race and ethnicity. This subgroup analysis showed that, for some sub-groups, the programme was more effective than the average effect. It also showed that the Early Head Start programme was ineffective for families that had multiple risk factors (Love et al., 2002, pp. 356-7[86]).

While it will often be worthwhile to consider these approaches, it must be acknowledged that the effect of a programme on ‘the average’ will often still be of significant interest to policy makers alongside impacts for specific cohorts. If the programme was extremely effective for the most disadvantaged quintile of the population but there was no effect on ‘the average’ then the programme must also have had a negative effect on some people in the population. If this were the case, policymakers would be interested to know and think about how best to target the intervention to people who would benefit.

4 Maximising opportunities for international cooperation

The development of effective evaluation capacity within government requires investment in a wide range of areas, including development of skills, provision of a supporting environment at the national level, and engagement with data holders, statistical offices, academia and researchers. The complexity of such tasks means that countries and organisations stand to gain significantly by sharing their experiences across borders.

The health sector was a pioneer in the use of evidence syntheses and systematic reviews, starting 50-60 years ago, with the development of health technology assessment. The systematisation of medical reviews under Cochrane is now over 30 years old. This practice expanded over time to the social sciences through academic networks. In 2000, the Campbell collaboration was established, with a focus on producing reviews of the evidence on the effectiveness of social interventions on an international level. Several of the Campbell programmes were highlighted during the workshop, including those on Countering Violent Extremism evidence, a five-country partnership of Australia, Canada, New Zealand, the United Kingdom and United States academics.

Creating an experimenting society and bringing a 'what works' philosophy to social policy presents an opportunity for significant population-level outcomes. This follows the path forged by the prominent social scientist Donald Campbell, who wrote 'Methods for Experimenting Society', over 50 years ago, as part of his vision for helping governments to produce better informed policies and social interventions, via research and evaluation.

Given the cognitive, technical and political barriers to supplying and using evaluations for decision making, continuous and sustained collaboration and momentum is required. Some of the efforts undertaken through expert and academic initiatives such as Campbell, can serve as inspiration for further international and intergovernmental cooperation. Collaboration can also be facilitated by working with philanthropic foundations, several of which have a strong track record of supporting robust evaluation, including the Arnold foundation, the Paul Ramsey Foundation, the Wellcome Trust, and many others in the area of international development. Additional organisations have also emerged, for example the Global Commission on Evidence, which began as a grassroot effort to improve use of research evidence, and in its latest 2025 annual update, focused on priorities such as formalising and strengthening domestic evidence support systems, enhancing and leveraging the global evidence architecture and putting evidence at the centre of everyday life, as mentioned by Dr. Andrew Leigh during the 2025 workshop.

International organisations such as the OECD can also play a valuable role in raising the quality and quantity of evidence in many of their sectoral areas of expertise. The OECD has an extensive history of generating robust policy evidence and conducting and sharing evidence reviews, for example through counterfactual evaluations of active labour market policies (OECD/EC, 2025[7]), reviews of evaluations systems for development cooperation (OECD, 2023[87]), and in the area of carbon mitigation, as highlighted above. It also supports the goal of promoting the institutionalisation of rigorous evaluation across public policy areas, increasing both quality and impact as per the OECD recommendation of the Council on Public Policy Evaluation.

The challenge for national governments is to find practical, actionable options to share evaluation approaches and innovative solutions. This includes drawing on leading practices, while recognising some of the broader challenges of evaluating interventions and policies in the social and economic field. Much of the existing literature on health and medical interventions focuses on narrowly defined, specific interventions whereas evaluations of social programmes or policies may focus on broader service systems, where the institutional aspects play an important role. As such, getting a broader understanding of the effectiveness of interventions from an international perspective holds significant potential. Many countries are implementing policies to this end, which include a range of interventions and other institutional aspects, often implemented both at national and subnational level.

There are already some promising international initiatives aimed at boosting international collaboration on evidence. These include the Global Commission for Evidence, the recent Global Evidence Report, which offers a blueprint for better international cooperation (see Box 7), and the Evidence Synthesis Infrastructure Collaborative initiative, supported by the Wellcome Trust.

Box 7. The Global Evidence Report, a blueprint for better international collaboration on evidence

The Global Evidence Report seeks to offer a roadmap for leveraging international cooperation. The goal is to facilitate broader cost sharing and promote greater efficiency in the provision of evidence for public services across national borders. It provides six strategic initiatives under three main priorities:

A. Increasing the generation and sharing of high-quality primary evidence
   1. Establish a shared evaluation fund
   2. Promote standard reporting and publication protocols

B. Advancing the quality and relevance of secondary evidence
   3. Conduct evidence gap maps across priority policy areas
   4. Promote meta-living evidence reviews, to synthesise this evidence into high quality comprehensive reviews

C. Boosting evidence adoption
   5. Strengthen international public service professional networks
   6. Conduct research into effective translation and adoption

Government leaders and experts from the United Kingdom, Canada, the United States and Australia contributed to this global evidence report, supported by the United Kingdom's Economic and Social Research Council. The report provides some cost estimates for these initiatives with the evidence gaps maps and the living evidence reviews being in the higher range, whereas common reporting, research on translation and adoption and policy and professional networks represent lower-cost investments.

Source: (Halpern and Maru, 2024[88])

4.1. Peer learning and exchange

One of the most accessible forms of cooperation is to promote peer learning and exchange. The challenge is to organise this across countries, while also building bridges within the academic community. Many international networks to date have been organised through academic and expert organisations, with fewer initiatives involving the participation of government officials. An increase in such initiatives would be of value – indeed, the Global Evidence Report (Box 7) recognises the need for public service professional networks.

Peer learning and exchange at the international level may help countries move a step further in developing such networks. Some OECD governments have decades of experience identifying evidence gaps, putting policies to the test and implementing the most effective programmes. However, further work is required to reform public administration to focus on finding effective interventions, and on building a body of programmes backed by strong, replicated, evaluation evidence of important lasting improvements in people's lives. This requires setting up specific functions with cross-cutting leadership, coordinating across departments and identifying the best and most relevant expertise at the international level.

Peer learning and exchange is organised at the OECD both for policy evaluation of countries' national practices and for development evaluation:

- The OECD network of policy evaluation experts meets to discuss and gather best practices, and is assisting with the implementation of the OECD recommendation on policy evaluation. In their respective countries, the experts that make up this network work with organisations that often have a cross-cutting mandate for promoting evaluation across government, disseminating standards and ensuring the right professional skills. In some countries, central organisations have the capacity to request and coordinate evaluation plans from ministries, but often these remain far from evidence gap maps. More focused in-depth discussions could still be organised among these experts to highlight and identify the capacity for developing evidence gaps maps and for sharing expertise in terms of how to implement and promote randomised trials, as well as robust quasi-experimental and theory-based approaches.
- The OECD Development Assistance Committee network on Development cooperation (EvalNet) promotes the use of evaluation for building a strong evidence base for learning, policy making and accountability, to achieve better development results. It involves the central and independent evaluation units of countries and collaborates closely with the evaluation units of partner countries, regional development banks, the World Bank, the IMF, UNDP and the UN Evaluation Group. EvalNet also works closely with evaluation associations and other evaluation networks.

s

In terms of research into effective translation and adoption, the OECD has also engaged with the European Commission and its Joint Research Centre on a multi country project over 2022-2024 (OECD/EC forthcoming), focused on promoting evidence-informed policy making (EIPM). While some of the focus was wider, encompassing the use of all academic research, the project investigated the issues of achieving impact, ensuring demand for evidence by policymakers, and creating the conditions for a fruitful policy making ecosystem. Many of the principles identified through this research are cross-cutting and include elements such as developing cross-governmental networks for science advisors, evaluations, and statistics, creating guidelines to promote incorporation of evidence into policy, organising checks and balances in evidence mobilisation processes, institutionalising EIPM at the centre of government, developing knowledge repository and knowledge management tools and promoting evidence champions. This project also highlighted that in many European countries, practices for creating knowledge repositories and knowledge management tools are often inadequate. Furthermore, it identified that in order to foster stronger international collaboration, countries must be in a position to mobilise the work and evaluations which they have conducted at national level.

4.2. Engaging in knowledge hubs

Governments and countries can also engage with knowledge-driven organisations operating in the field. The topic may depend on the kind of expertise and profile of respective organisations.

As previously mentioned, the Campbell Collaboration has been leading the field of systematic reviews for social policy, promoting evidence-based decisions and policy. This not-for-profit organisation has been established for over a quarter of a century, and publishes results in an open access journal, committed to

publishing systematic reviews. Campbell also coordinates evidence gap maps in a number of areas.[14] Its international development coordinating group was established jointly by the International Initiative for Impact Evaluation (i.e. see Box 4), and the University of Ottawa in 2011. Several Nordic countries which have a strong tradition of excellence in systematic reviews, have also been actively engaged with Campbell, including through organisations such as the Norwegian Directorate for Health and Social Affairs, and the Swedish National Board of Health and Welfare. However, as reviews are largely dominated by English speaking papers and journals, some of the results may not always be fully appreciated in non-anglophone OECD countries. Regardless, a more systematic mapping of what would be available in the Campbell space could represent a useful first step before undertaking evidence synthesis at national level.

Another relevant knowledge hub is the Competence Centre on Microeconomic Evaluation (CC-ME) established in May 2016 at the European Commission's Joint Research Centre. Their work aims to support European Union (EU) policy making through *ex post* causal evaluation and data driven microeconomic analysis, which is often closely related to the policy needs of EU Member states, particularly in the areas of spending and regulation. The CC-ME also supports networks and various communities of practice at European level, which offers an opportunity for practitioners working in government to get more familiar with quasi-experimental techniques and the handling of large micro datasets, as well as to exchange with peers. The centre also helps experts to navigate and identify the best conditions for using quasi-experimental methods and randomised trials. Increasing understanding of large microdata sets is valuable because administrative data can be used to make impact evaluations more powerful by both providing outcome data and providing control variables that can, for example, be used to look at different effects for different cohorts. The CC-ME maintains a knowledge hub on projects with admin data (See also Berton and Paruolo, (2025[89])).

In the area of development, the Global Evaluation initiative (GEI) is a global network working with developing countries to build stronger monitoring and evaluation (M&E) systems to improve how governments gather and use evidence. The GEI is supported by various national development cooperation organisations in Canada, Denmark, Finland and Germany, and is coordinated through the World Bank Group's Independent Evaluation Group in collaboration with the UN Development Programme's Independent Evaluation Office. By creating a global network for knowledge exchange and promoting best practices in evaluation, GEI ensures that data is effectively utilized to address current challenges and inform future solutions.

Overall, it is worth noting that there are often more resources available at the international level for evaluation experts working in the area of international development, than for domestic level programmes. This reflects a strong tradition established in the area of international development, related to accountability requirements from donors. However, this may also result in the international market for expertise and knowledge hubs to be slightly skewed towards international development issues, despite the domestic reform needs of OECD countries.

4.3. Engaging with existing initiatives

Finally, there is scope for countries to coordinate evidence agendas. As highlighted above, experts from the United Kingdom, Canada, Australia, the United States as well as academic experts have joined efforts in sharing needs and identifying common avenues for international cooperation.

[14] I.e. Ageing, Business & Management, Children & Young Persons Wellbeing, Climate Solutions, Crime & Justice, Disability, Education, International Development, Knowledge Translation & Implementation, Methods and Social Welfare

Another broader initiative is the Evidence Synthesis Infrastructure Collaborative (ESIC), initiative supported by the Wellcome Trust. This was conceived as an approach to deliver synthesised evidence to a wide range of audiences, with an ambitious international framework, results available in seven languages, and co-publications across five journals.[15] ESIC builds on the four-country collaboration and the Global Commission on Evidence, and broadens the framework, with support both from United Kingdom Research and Innovation and the Wellcome Trust,[16] with a broader focus on sustainable development goals (SDGs) to appeal to a range of countries. The goal is to develop repositories, tools and governance that can collate and make sense of evidence in complex areas such as climate change and healthy ageing.

The initiative set up five working groups, focusing on 1) demand-side engagement, 2) data sharing and reuse, 3) safe and responsible use of AI, 4) methods and process innovation, and 5) capacity sharing, with another set of discussions focusing on governance and change management. Each of the five working groups used a "double diamond" approach, first exploring widely through divergent thinking, before then focusing action through convergent thinking on "What's the solution". The approach included a SHOW ME the evidence approach with six features to reliably deliver evidence to those who need it:

- Support systems nationally that use many forms of research evidence to help address local priorities.
- Harmonise efforts globally to make it easier to learn from others around the world.
- Open-science approaches that make it the norm to build on what others have done.
- Waste-reduction efforts to make the most of investment in evidence support and research.
- Measured communications to clarify what is known from existing evidence.
- Equity and efficiency in all aspects of the work.

Over a hundred contributing authors supported this work, available in various languages and published in a variety of settings. This work also seeks to align with the United Nations (UN) evaluation efforts, including the UN Independent Evaluation offices active in various UN agencies, and coordinated by the UN SDG System-wide evaluation office.

This effort also highlights that at a sectoral level, in areas such as education or health technology assessment, evaluation efforts are also coordinated through professional organisations. The OECD as a multisectoral organisation involved in applied research and economics is also connected to many of these initiatives through its wide range of committees and working groups. The work gives significant attention to Living Evidence Syntheses, identifying a need for sharing methods and for sustained funding to access human and other resources, technologies and databases.

In terms of governance, ESIC frames many of the bridges that are necessary between the demand side - including policy makers, practitioners, intergovernmental organisations at regional and global level, and citizens - and the supply side, with evidence synthesis generators, academic institutions, networks and technology providers, as well as funders. On the practical considerations, the project offers a range of initiatives in terms of solutions and provides some costing estimates.

These international initiatives are helping build connections between countries and sectors. Overall, this is all part of a two-way process, where countries and experts need to create national systems for coordination, sharing expertise on evaluation and disseminating good practices, and then ensuring some connections, peer learning and knowledge sharing mechanisms at the international level.

[15] Journals included Campbell, Cochrane, collaboration for environmental evidence, guidelines international network, and JBI Evidence Synthesis

[16] See https://wellcome.org/news/evidence-synthesis-infrastructure-collaborative

5 Conclusion

High-quality impact evaluations and evidence syntheses are powerful tools for informing public policy decisions. They cover a range of techniques which can be used to mobilise quantitative methods, with a mix of experimental and quasi-experimental approaches which help to assess the impacts of policies and programmes. They can provide rigorous insights into what works, for whom, and under what conditions, thereby helping avoid policy failures and ensuring resources are directed where they can deliver the greatest benefit. However, a range of political, technical, and institutional barriers often hamper their widespread adoption These barriers can be overcome through better alignment of incentives, greater investment in evaluation capacity, smarter use of existing data, stronger partnerships across sectors, and the integration of evaluation into budgeting and decision-making processes. Great potential also exists with a variety of new tools and approaches, including artificial intelligence and the importance of international collaboration to scale effective practices. By embedding evaluation more systematically into policy processes, and overcoming barriers to its adoption, governments can make more informed decisions, better allocate resources, and deliver more effective outcomes to their citizens.

Unleashing the potential of rigorous impact evaluation has implications for a range of audiences:

- For senior policy and decision-makers, embedding rigorous impact evaluation (including randomised trials) into the heart of policy development can be valuable in leading to a shift from intuition- or precedent-based decision-making toward evidence-informed governance and decision-making. The benefits of this approach can be used as a basis to advocate for institutional reforms, such as incorporating evaluation into budget processes or establishing legal frameworks that require systematic evaluation across departments. There are several strong international examples that can serve as models for embedding evaluation into legislative and budgetary cycles. Policymakers can use the findings to build the case for cultural change within government, to value learning from failure and prioritise outcomes for citizens over the preservation of programmes.
- For evaluation professionals, There is potential to mobilise technical guidance and strategic resources through sharing some best practices and highlighting practical steps for designing and delivering high-quality evaluations, including checks for feasibility, summaries of key impact evaluation methods, and tips for ensuring impact evaluations are rigorous. It is also important to be aware of how to navigate ethical concerns, engage stakeholders, and communicate findings effectively. Evaluation units can draw on the findings to advocate for better access to administrative data and to build internal capability through training and peer learning.
- For budget and finance officials, rigorous impact evaluation findings can be used to ensure that public spending delivers value for money, in particular demonstrating how they can inform resource allocation decisions by identifying which programmes are most cost-effective. Several countries have embedded evaluation into their budget processes, offering practical models for linking evidence to funding decisions. Finance ministries have a key role in championing an evaluation culture across government, particularly by attaching evaluation requirements to funding approvals and supporting interministerial or cross-government coordination. Finance officials can use the paper to support the development of public financial management and governance mechanisms that require evaluations to be considered during budget planning, and to justify investments in evaluation infrastructure as a means of reducing long-term fiscal risk.

In summary, rigorous impact evaluation is not just a technical exercise, but can form a core part of accountable, effective, and citizen-focused government. By embracing innovative tools and approaches, governments can move beyond assumptions and anecdotes to make decisions grounded in evidence. For those shaping policy, allocating resources, and designing policies and programmes, evaluation can and should be embedded from the start to ensure continuous learning and improvement. To do this, there is a need to build better systems for rigorous impact evaluation, invest in the requisite skills, and foster a culture that drives evidence-informed approaches. In doing so, this unlocks the full potential of public policy to improve lives and deliver public value to citizens.

Annex A. The ACE-OECD International Workshop on Rigorous Impact Evaluation

In February 2025, the OECD and ACE co-hosted a high-level international workshop with senior policymakers, academics, and evaluators discussing how to strengthen evaluation practice and embed evidence more systematically into policy. Dr Andrew Leigh, Assistant Minister for Productivity, Competition, Charities and Treasury (Australia) stressed the value of experimental methods in identifying what works and what doesn't. He emphasised the need for humility in policymaking and for evaluators to prioritise the interests of citizens above defending existing programmes. David Halpern, President Emeritus of the United Kingdom's Behavioural Insights Team, provided the second keynote speech. He highlighted the UK's experience with What Works Centres and the Evaluation Taskforce, contributing a culture of experimentation. A panel discussion brought together Professor Tito Boeri (Bocconi University), Professor Martina Björkman Nyquist (Stockholm School of Economics), Professor Michael Sanders (King's College London), Doctor Paolo Paruolo (EC JRC) and Professor Dan Levy (Harvard Kennedy School) to share a range of perspectives on the barriers and enablers to the use of impact evaluation in decision-making.

Discussions highlighted the issue of frequent political discomfort with the perceived unfairness of randomisation whereby some people receive a promising intervention and others explicitly do not. In these cases, new methods can help leverage the value of large administrative datasets to understand the impact of new/pilot interventions. Experts discussed successes in opening up data and archives to researchers at administrative institutions, allowing for collaboration under strict on-site requirements, which built mutual understanding and facilitated rigorous evaluations with real-world impact.

The conversation also emphasised the importance of evaluation being considered from the outset of policy proposals, rather than being tacked on at the end, and drew attention to the need for broad support and institutional commitments to ensure evaluation can remain beyond political cycle. Other experts, highlighted the specific issue of a reluctance of teams to address failure, and advocated for embracing learning through failure. In some cases, social care trials can be linked to planning policy rollouts, and frontline workers' ideas can in many cases be more effective than centrally designed programmes. The expert panel also highlighted the tension between scientific and political approaches, stressing the need to promote a stronger culture of evidence, and to ensure political actors have guidance in understanding where evaluations are useful and where they are not.

Participants gathered in structured breakout sessions, with the following key insights in relation to barriers and enablers for high quality impact evaluation:

- Ensuring the quality of impact evaluations through adequate sample sizes, clear theories of change, robust implementation, ensuring that randomised trials are well designed and used in the right way.
- Engaging evaluators early, during the design phase, to ensure alignment with programme goals and political timeline.
- Effectively addressing ethical and design concerns amongst stakeholders.
- Promoting transparency by pre-registering evaluations and documenting changes to methods or analysis and sharing any design flaws when identified.
- Strengthening evaluation ecosystems by building both supply-side and demand-side capacity, as well as supporting data harmonisation.

Note: the workshop interactive breakout sessions were moderated by OECD experts, including Andrew Blazey (Public Management and Budgeting, Claudio Alberti (Development Cooperation Directorate), Anne Lauringson (Employment Labour and Social Affairs) and Carlos Hinojosa (General Secretariat/Evaluation and Internal Audit),

References

Abdul Latif Jameel Poverty Action Lab (2020), "Conducting cost-effectiveness analysis (CEA)", https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea (accessed on 14 July 2025). [14]

Abrell, J., M. Kosch and S. Rausch (2022), "How effective is carbon pricing?—A machine learning approach to policy evaluation", *Journal of Environmental Economics and Management*, Vol. 112, p. 102589, https://doi.org/10.1016/j.jeem.2021.102589. [36]

ACE (2025), *Cost-Effectiveness Analysis (CEA)*, https://evaluation.treasury.gov.au/sites/evaluation.treasury.gov.au/files/2025-07/guide-cost-effectiveness-analysis.pdf. [15]

Agenda Digitale (2024), "FOSSR: innovazione e Open Science per la ricerca sociale italiana [FOSSR: Innovation and Open Science for Italian social research]", https://www.agendadigitale.eu/cultura-digitale/fossr-innovazione-e-open-science-per-la-ricerca-sociale-italiana/? (accessed on 10 June 2025). [20]

Au, L. et al. (2025), "Using artificial intelligence to semi-automate trustworthiness assessment of randomized controlled trials: a case study", *Journal of Clinical Epidemiology*, Vol. 180, p. 111672, https://doi.org/10.1016/j.jclinepi.2025.111672. [43]

Australian Centre for Evaluation (2025), *ACE Evaluation Library*, https://evaluation.treasury.gov.au/about/ace-evaluation-library. [32]

Berton, F. and P. Paruolo (2025), *Data-driven learning in the EU*, Springer Cham, Professional Practice in Governance and Public Organizations. [89]

Biddle, N., M. Gray and M. Hiscox (2022), "Public support for Randomised Controlled Trials and nudge interventions in Australian Social Policy". [81]

Boaz, A. et al. (2018), "How to engage stakeholders in research: design principles to support improvement", *Health Research Policy and Systems*, Vol. 60/16, https://doi.org/10.1186/s12961-018-0337-6. [80]

Boehmer, H., X. Fernandez Ordonez and N. Sharma (2014), *Reflections on independence in evaluation*, https://nec.undp.org/sites/default/files/2021-07/Reflections%20on%20independence%20in%20evaluation%202013.pdf. [54]

Braga, L. et al. (2025), "Randomised Controlled Trials - the What, When, How and Why", *Journal of Pediatrick Urology*, Vol. 21/2, pp. 397-44, https://doi.org/10.1016/j.jpurol.2024.11.021. [78]

Canadian Evaluation Society (2019), "Competencies for Canadian Evaluation Practice", https://evaluationcanada.ca/files/pdf/2_competencies_cdn_evaluation_practice_2018.pdf (accessed on 16 June 2025). [18]

Card, D., J. Kluve and A. Weber (2018), "What Works? A meta analysis of recent Active Labour Market Program Evaluations", *Journal of the European Economic Association*, Vol. 16/3, pp. 894-931, https://doi.org/10.1093/jeea/jvx028. [91]

Card, D. and A. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", https://www.jstor.org/stable/2118030?seq=1 (accessed on 11 July 2025). [65]

Chambers, C. and L. Tzavella (2021), "The past, present and future of Registered Reports", *Nature Human Behaviour*, Vol. 6/1, pp. 29-42, https://doi.org/10.1038/s41562-021-01193-7. [68]

Champion K. and L. Emma (2023), "Health4Life eHealth intervention to modify multiple lifestyle risk behaviours among adolescent students in Australia: a cluster-randomised controlled trial", *the Lancet Digital Health*, Vol. 5/5, pp. E276-E287, https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00028-6/fulltext. [74]

CORE (2025), *A knowledge base on children & youth in the digital age*, https://core-evidence.eu/. [34]

Department of Planning Monitoring and Evaluation, South Africa (2019), *National Evaluation Policy Framework*. [84]

Education Endowment Foundation (2025), "Teaching and Learning Toolkit", https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit (accessed on 12 June 2025). [27]

Eggers, A. et al. (2014), "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races", *American Journal of Political Science*, Vol. 59/1, pp. 259-274, https://doi.org/10.1111/ajps.12127. [63]

Filmer, D. and N. Schady (2009), "School Enrollment, Selection and Test Scores", https://ssrn.com/abstract=1437950 (accessed on 11 July 2025). [62]

Franzen, S. et al. (2022), *Advanced Content Analysis: Can Aritificial Intelligence Accelerate Theory-Driven Complex Program Evaluation?*, IEG Methods and Evaluation Capacity Development Working Paper Series, https://documents1.worldbank.org/curated/en/400031645128516191/pdf/Advanced-Content-Analysis-Can-Artificial-Intelligence-Accelerate-Theory-Driven-Complex-Program-Evaluation.pdf. [37]

Gertler, P. et al. (2016), *Impact Evaluation in Practice*, World Bank, http://hdl.handle.net/10986/25030. [5]

Gibson, M. and A. Sautmann (2023), *Introduction to randomized evaluations*, https://www.povertyactionlab.org/resource/introduction-randomized-evaluations. [58]

Government of Japan (2001), "Government Policy Evaluations Act (Act No. 86 of 2001)", https://www.soumu.go.jp/english/kansatu/evaluation/evaluation_09.pdf (accessed on 12 June 2025). [49]

Government of Spain Official State Gazette (2022), "Ley 27/2022, de 20 de diciembre, de institucionalización de la evaluación de políticas públicas en la Administración General del Estado [Law 27/2022, of December 20, on the institutionalization of the evaluation of public policies in the General Administration of the State.]", https://digital.gob.es/content/dam/sgad/sefp/es/portalsefp/evaluacion-politicas-publicas/Ley.pdf (accessed on 12 June 2025). [50]

Green Climate Fund (2021), "Designing an impact evaluation in six steps", https://ieu.greenclimate.fund/blog/designing-impact-evaluation-six-steps (accessed on 10 September 2025). [16]

Halpern, D. (ed.) (2024), *Global Evidence Report, A blueprint for better international collaboration on evidence*, Behavioral Insights Team, UK Nesta, Economic and Social Research Council, https://www.bi.team/wp-content/uploads/2024/08/ESRC-Global-Evidence-Report-September-2024-1.pdf. [30]

Halpern, D. (2024), *The Catalytic State, a practical theory of government*, Behavioral Insights Team, International School of Government, the Policy Institute, King's College London, https://www.kcl.ac.uk/policy-institute/assets/the-catalytic-state-a-practical-theory-of-government.pdf. [1]

Halpern, D. and D. Maru (2024), *Global Evidence Report, a blue print for better international cooperation on evidence*, The Behavioral Insights Team, NESTA, UK Economic and Social Research Council, https://www.bi.team/publications/international-collaboration-evidence/. [88]

Hansen, B. (2015), "Punishment and Deterrence: Evidence from Drunk Driving", *American Economic review*, Vol. 105/4, pp. 1581-1617, https://doi.org/10.1257/aer.20130189. [59]

Hansen, B. (2015), "Punishment and Deterrence: Evidence from Drunk Driving", *American Economic Review*, Vol. 105/4, pp. 1581-1617, https://doi.org/10.1257/aer.20130189. [60]

Haskins, R. and A. Feldman (2016), "Low-Cost Randomized Controlled Trials", https://govinnovator.com/wp-content/uploads/2020/05/low-cost_rcts.pdf (accessed on 11 July 2025). [73]

Hayes, H., M. Parchman and R. Howard (2011), "A logic model framework for evaluatio nand planning in a primary care practice-based research network (PBRN)", *J. Am Boardd Fam Med*, Vol. 24/5, pp. 576-82, https://doi.org/10.3122/jabfm.2011.05.110043. [67]

HM Treasury (2020), *Magenta Book: Central GOvernment guidance on evaluation*, https://assets.publishing.service.gov.uk/media/5e96cab9d3bf7f412b2264b1/HMT_Magenta_Book.pdf. [11]

Independent Evaluation Office (2025), *Aida enters a new chapter*, https://www.undp.org/evaluation/news/aida-enters-new-chapter. [44]

INSEE (2025), "Courrier des statistiques N13 - 2025", https://www.insee.fr/fr/information/8546934?sommaire=8546949 (accessed on 10 September 2025). [47]

IReF (2023), "IReF reaffirms its commitment to Evaluation and presents an Observatory of findings and proposals", https://www.airef.es/en/news/la-airef-reafirma-su-compromiso-con-la-evaluacion-y-presenta-un-observatorio-de-hallazgos-y-propuestas/ (accessed on 12 June 2025). [55]

Jardim, P. et al. (2022), "Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system", *BMC Medical Research Methodology*, Vol. 22/1, https://doi.org/10.1186/s12874-022-01649-y. [39]

Johnson, N. and M. Phillips (2018), "Rayyan for systematic reviews", *Journal of Electronic Resources Librarianship*, Vol. 30/1, pp. 46-48, https://doi.org/10.1080/1941126x.2018.1444339. [38]

Kellermeyer, L., B. Harnke and S. Knight (2018), "Covidence and Rayyan", *Journal of the Medical Library Association*, Vol. 106/4, https://doi.org/10.5195/jmla.2018.513. [40]

Laura and John Arnold Foundation (2018), "Low-Cost Randomized Controlled Trials to Drive Effective Social Spending", https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/Request-for-Proposals-Low-Cost-RCT-FINAL.pdf (accessed on 11 July 2025). [69]

Lee, C. and A. Lee (2020), "How Artificial Intelligence Can Transform Randomized Controlled Trials", *Translational Vision Science & Technology*, Vol. 9/2, p. 9, https://doi.org/10.1167/tvst.9.2.9. [42]

Lee, C. and A. Lee (2020), "How Artificial Intelligence Can Transform Randomized Controlled Trials", *Translational Vision Science & Technology*, Vol. 9/2, p. 9, https://doi.org/10.1167/tvst.9.2.9. [41]

Leigh, A. (2018), *Randomistas: How radical researchers changed our world*, La Trobe University Press and Black Inc, Melbourne. [8]

Levy, D. (2025), "Address to OECD International Workshop on Rigorous Impact Evaluation Approaches including Randomised Controlled Trials". [71]

Love, J. et al. (2002), "Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start", https://eric.ed.gov/?id=ED472186 (accessed on 11 July 2025). [86]

Lucas, A. and I. Mbiti (2014), "Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya", *American Economic Journal: Applied Economics*, Vol. 6/3, pp. 234-263, https://doi.org/10.1257/app.6.3.234. [61]

Martinuzzi, A. (ed.) (2022), "Nurse home visiting to improve child and maternal outcomes: 5-year follow-up of an Australian randomised controlled trial", *PLOS ONE*, Vol. 17/11, p. e0277773, https://doi.org/10.1371/journal.pone.0277773. [77]

McCloskey, G. (ed.) (2019), *Balancing Risk and Benefit: Ethical Tradeoffs in Running Randomised Evaluations*, Oxord Handbooks, https://global.oup.com/academic/product/the-oxford-handbook-of-professional-economic-ethics-9780199766635?cc=fr&lang=en&#. [79]

Mckenzie, D. (2023), "Seven ways to improve statistical power in your experiment without increasing n", https://blogs.worldbank.org/en/impactevaluations/seven-ways-improve-statistical-power-your-experiment-without-increasing-n (accessed on 11 July 2025). [72]

Ministry of Economy and Finance (2024), "I Piani Triennali (The Three-Year Plans)". [56]

Mol, B. et al. (2023), "Checklist to assess Trustworthiness in RAndomised Controlled Trials (TRACT checklist): concept proposal and pilot", *Research Integrity and Peer Review*, Vol. 8/1, https://doi.org/10.1186/s41073-023-00130-8. [90]

OECD (2025), "Building capacity for evidence-informed policymaking in Belgium: Assessment and recommendations roadmap", *OECD Public Governance Policy Papers*, No. 71, OECD, Paris Cedex 16, https://doi.org/10.1787/223b01a8-en. [21]

OECD (2025), "Building Capacity for Evidence-Informed Policymaking in Governance and Public Administration in Belgium Findings from the Diagnostic Report & Needs and Gaps Assessment", https://bosa.belgium.be/sites/default/files/content/documents/EIPM%20Belgium%20Annex.pdf (accessed on 10 June 2025). [19]

OECD (2025), *DAC Evaluation Resource Centre (DEReC)*, https://www.oecd.org/en/about/programmes/dac-evaluation-resource-centre---derec.html. [33]

OECD (ed.) (2025), *Development Co-operation Tools Insights Practices*, https://www.oecd.org/en/topics/sub-issues/development-co-operation-in-practice/development-co-operation-tools-insights-practices.html. [46]

OECD (2025), "Governing for the green transition", *OECD Net Zero+ Policy Papers*, No. 13, OECD Publishing, Paris, https://doi.org/10.1787/5b0aa7d0-en. [93]

OECD (2025), *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*, OECD Publishing, Paris, https://doi.org/10.1787/795de142-en. [35]

OECD (2025), *Government at a Glance 2025*, OECD Publishing, Paris, https://doi.org/10.1787/0efd0bcd-en. [31]

OECD (2025), *Implementation Toolkit for the OECD Recommendation on Public Policy Evaluation*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/77faa4fe-en. [53]

OECD (2025), *OECD Regulatory Policy Outlook 2025*, OECD Publishing, Paris, https://doi.org/10.1787/56b60e39-en. [48]

OECD (2025), "Public Policy Evaluation Implementation Toolkit", https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/implementation-toolkit-for-the-oecd-recommendation-on-public-policy-evaluation_f24516be/77faa4fe-en.pdf (accessed on 16 April 2025). [3]

OECD (2024), "Building capacity for evidence-informed policymaking in Latvia: Assessment and recommendations roadmap", *OECD Public Governance Policy Papers*, No. 65, OECD Publishing, Paris, https://doi.org/10.1787/2b43bc19-en. [22]

OECD (2024), "Improving decision making through policy evaluation in Portugal", *OECD Public Governance Policy Papers*, No. 63, OECD Publishing, Paris, https://doi.org/10.1787/4f8ef34d-en. [51]

OECD (2024), "Selective Spending Reviews in Chile", https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/12/selective-spending-reviews-in-chile_9c102565/bb9194fe-en.pdf (accessed on 12 June 2025). [57]

OECD (2023), *Evaluation Systems in Development Co-operation*, https://doi.org/10.1787/a255365e-en. [87]

OECD (2023), *Glossary of Key Terms in Evaluation and Results-based Management for Sustainable Development (Second Edition)*, https://doi.org/10.1787/632da462-en-fr-es. [12]

OECD (2022), "Recommendation of the Council on Public Policy Evaluation", https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0478 (accessed on 12 June 2025). [2]

OECD (2020), *Improving Governance with Policy Evaluation: Lessons From Country Experiences*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/89b1577d-en. [4]

OECD (2010), *Quality Standards for Development Evaluation*, DAC Guidelines and Reference Series, OECD Publishing, Paris, https://doi.org/10.1787/9789264083905-en. [26]

OECD/EC (2025), *Counterfactual impact evaluations of active labour market policies: Lessons from using linked administrative data*, https://doi.org/10.1787/bea7860e-en. [7]

OECD/EC (2025), *Strengthening National Evidence-informed Policymaking ecosystems*, OECD/EC. [23]

OECD/KIPF (2024), *Addressing Inequality in Budgeting, Lessons from Recent OECD Experience*, https://doi.org/10.1787/ea80d61d-en. [13]

OPSI (2023), "The Evaluation Task Force", https://oecd-opsi.org/innovations/the-evaluation-task-force/ (accessed on 18 June 2025). [29]

Paul Ramsay Foundation (2024), "Experimental evaluation open grant round", https://www.paulramsayfoundation.org.au/news-resources/experimental-evaluation-open-grant-round (accessed on 11 July 2025). [70]

Peersman, G. and P. Rogers (2025), "Impact Evaluation", https://www.betterevaluation.org/methods-approaches/themes/impact-evaluation#:~:text=Impact%20evaluation%20might%20be%20appropriate%20when%20there%20is%20a%20need,the%20quality%20of%20the%20implementation. (accessed on 11 September 2025). [52]

Petrosino, A. et al. (2013), "'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency", *Cochrane Database of Systematic Reviews*, Vol. 2013/4, https://doi.org/10.1002/14651858.cd002796.pub2. [9]

Pritchett, L., S. Samji and J. Hammer (2013), "It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space". [83]

Ravallion, M. (2009), "Evaluation in the Practice of Development", https://documents1.worldbank.org/curated/en/693621468155368608/pdf/767890JRN0WBRO00Box374387B00PUBLIC0.pdf (accessed on 12 September 2025). [82]

Rehill, P. et al. (2025), *Randomised trials in Australian public policy: a review,*, The Australian Centre for Evaluation, The Australian Treasury, https://evaluation.treasury.gov.au/publications/randomised-trials-australian-public-policy-review. [76]

Revillard, A. (ed.) (2023), *Difference-in-differences Method*, Éditions science et bien commun, https://doi.org/10.5281/zenodo.8327162. [66]

Revillard, A. (ed.) (2023), *Mixed methods*, Quebec City: éditions science et bien commun, https://zenodo.org/records/8327162. [85]

Revillard, A. (ed.) (2023), *The Regression Discontinuity Design*, Editions science et bien commun, https://doi.org/10.5281/zenodo.8327162. [92]

Rouhi, M. (2005), "Thalidomide", *Chemical & Engineering News Archive*, Vol. 83/25, p. 122, https://doi.org/10.1021/cen-v083n025.p122. [75]

Sehrin, F. et al. (2024), "The effect on income of providing near vision correction to workers in Bangladesh: the THRIVE (Tradespeople and Hand-workers Rural Initiative for a vision-enhanced Economy) randomized controlled trial", *PLOS ONE*, Vol. 4, p. 19, https://doi.org/10.1371/journal.pone.0296115. [6]

The Hon Dr Andrew Leigh MP (2025), "Address to OECD International Workshop on Rigorous Impact Evaluation Approaches including Randomised Controlled Trial: The Seventh Phase of Good Government", https://ministers.treasury.gov.au/ministers/andrew-leigh-2022/speeches/address-oecd-international-workshop-rigorous-impact-evaluation (accessed on 10 June 2025). [24]

U.S. Department of Education (2005), "When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program", https://files.eric.ed.gov/fulltext/ED485162.pdf? (accessed on 10 September 2025). [25]

UCL (2018), "UK What Works Centres: Aims, methods and contexts", https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3731 (accessed on 23 January 2024). [28]

Van der Put, C. et al. (2020), "Effects of Awareness Programs on Juvenile Delinquency: A Three-Level Meta-Analysis", *International Journal of Offender Therapy and Comparative Criminology*, Vol. 65/1, pp. 68-91, https://doi.org/10.1177/0306624x20909239. [10]

Venkataramani, A., J. Bor and A. Jena (2016), "Regression discontinuity designs in healthcare research", *BMJ*, p. i1216, https://doi.org/10.1136/bmj.i1216. [64]

Wilton Park (2025), *AI and Knowledge Management in Evaluation and Evidence Synthesis*, https://www.wiltonpark.org.uk/reports/ai-and-knowledge-management-in-evaluation-and-evidence-synthesis/. [45]

World Bank (2020), "Administrative Data in Research at the World Bank: The Case of Development Impact Evaluation (DIME)", https://admindatahandbook.mit.edu/book/v1.0-rc4/dime.html (accessed on 10 September 2025). [17]