Global Evidence Report

A blueprint for better international collaboration on evidence

September 2024

David Halpern & Deelan Maru

THE BEHAVIOURAL INSIGHTS TEAM

nesta

UKRI Economic and Social Research Council

Foreword

As for all countries involved in the development of the blueprint for better international collaboration on evidence, Australia recognises the compelling benefits of building a strong evidence base to inform policy. Designing, building support for, and implementing good public policy is hard enough – at least we can facilitate its development by providing access to the best available data and analysis. Evidence-informed policy fosters accountability, transparency, and learning, as policies are monitored and evaluated for their outcomes and impacts.

Australia has been an engaged participant in the global movement for evidence-based policy and has made significant investments and innovations in this area. The challenge of generating, translating and using high quality evidence for policy is common to all countries involved in the blueprint project and we are keen to contribute to joint solutions.

Australia has closely followed developments in our partner countries like the maturing What Works Network and Evaluation Taskforce in the UK, the introduction of the Foundations for Evidence-Based Policymaking Act 2018 in the USA, and the Results Division and Policy on Results in Canada. We have recently established the Australian Centre for Evaluation (ACE) to drive our own efforts to increase the quality, volume and use of evidence for policy. I am also proud to champion the Australian public service's data profession, soon to be joined by an evaluation profession, echoing models in the UK and elsewhere.

These investments represent our countries' efforts to make better use of the available primary and secondary evidence to understand the detail of policy outcomes and impacts, and to better design and test policy interventions and scenarios.

The blueprint makes a compelling case for the potentially big gains to be had from better international collaboration on evidence. It provides a series of practical ideas for investments to improve the quality and use of evidence in policy design, implementation, and evaluation. It also highlights the opportunities and challenges ahead, as the demand for and supply of evidence continue to grow in a world with no shortage of complex, knotty policy challenges.

We are committed to engaging and collaborating with partners and advocates for better evidence, both domestically and globally, to share our experiences and learn from each other. I commend this blueprint to all who are involved and interested in contributing to the evidence-based policy agenda and hope it will inspire and inform the next phase of our collaborative work.

David Gruen

Australian Chief Statistician

Governments and public services exist to make citizens' lives better. Governments - at national, state and local level - spend trillions seeking to achieve this.

We should always ask: 'could we spend that dollar, or the precious time of that public servant, better?' We should be ceaselessly learning and innovating, in measured ways. We should be shamelessly borrowing, adapting and adopting programs and practices that work better. Learning from other countries is a good place to start - and especially from countries and states that have similar systems, characteristics and challenges.

This report picks up on 'unfinished business' from the decade I spent as the UK Prime Minister's National What Works Adviser. In this role, I and colleagues worked to champion and create a string of 'What Works' institutions, dedicated to generating, translating, and fostering the adoption of evidence-based approaches to public service and practice. It became clear that other countries, regions and cities were asking similar questions, and often pursuing closely related activities in parallel.

Teachers and parents in Manchester (UK), Melbourne (Australia), Montreal (Canada) and Memphis (U.S.) are all trying to figure out the best way to help their kids learn maths, stay out of trouble, and get on a path to a healthy and fulfilling life. There are differences between our systems, but there's a lot of similarities too. Kids find maths hard work, wherever they are.

It makes no sense for every school, state, or country to answer these questions in isolation. At the same time that there are not enough good quality 'systematic' reviews that pull together the evidence for practitioners, parents and policy makers, there is also a great deal of 'research waste' in the form of low quality, or partial, reviews. There are many gaps in our knowledge, wherever you look.

Why not collaborate to jointly support and fund such cross-national evidence 'public goods'? That is what this report examines, along with practical paths to deliver such goods.

One of the wonderful shared characteristics of liberal market democracies is a restless openness to new ideas and approaches. If someone comes along with a better way of doing something, we tend to be open to it, even if it upsets the current provider or interest group. It is a process of creative destruction that has powered our economies and public service innovation for at least a century. It is a tradition to be proud of, and embrace.

That same restless innovation, global exchange of ideas, and evidence-based practice has given you and me a very good chance of living longer than our grandparents. We need to spread those institutional habits outside of health and into other areas. And we should make it a cross-national endeavour.

David Halpern CBE

President & Founding Director, Behavioural Insights Team, and former What Works National Adviser (2013-2023)

This report expresses the views of the authors, and should not be taken as a statement of government policy from any of the four partner countries, nor of the Economic and Social Research Council.

Table of contents

Executive Summary: Strengthening the Global Evidence Ecosystem

Introduction

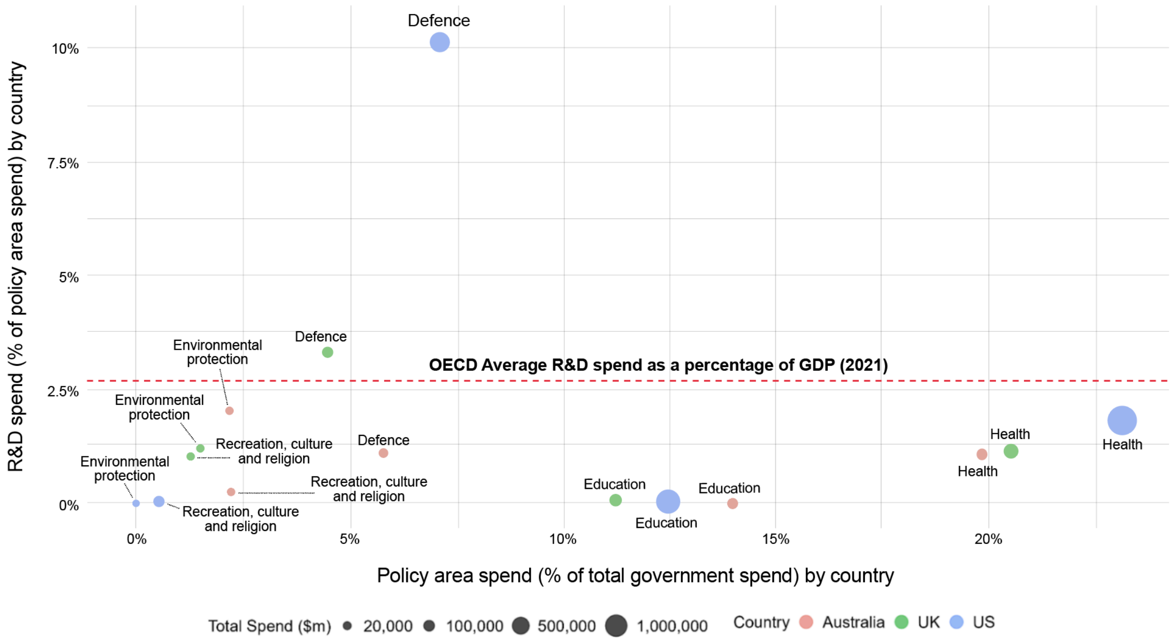
Governments across the world spend trillions annually on public services. Ideally, such expenditure is built on evidence. This report focuses on three aspects of the evidence ecosystem: primary evidence (specific studies); secondary evidence (syntheses, or robust reviews); and evidence adoption.

We use a combination of desk research, a survey of senior policy makers, and 1-to-1 interviews, to explore gaps and make recommendations. The report is focused on four countries: the U.S., the UK, Australia, and Canada. While not included within this review, other countries such as Germany and France have also expressed interest in collaboration.

The supply of evidence

There is a significant 'investment gap' in the production of evidence globally. With the notable exceptions of healthcare and defence, Research and Development (R&D) spending remains low in most areas of government, such as education, social protection, and public order. In such areas, R&D expenditure averages less than 0.25% of total spending, such as education where R&D is less than a tenth of the proportion in health or defence.

R&D spend compared to total expenditure by policy area (Australia, UK, U.S.) 2021-22 (based on OECD data: [public finance by function](#) & [government budget allocations for R&D](#))

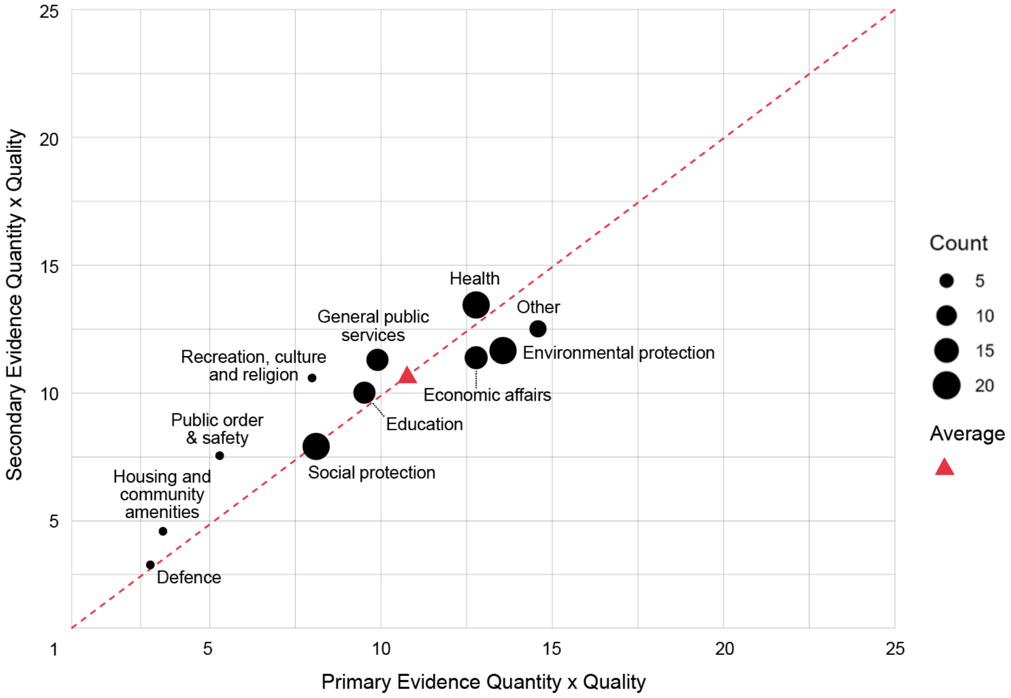


The implied 'R&D in public services gap' is around $85-115bn pa for the U.S., UK, Australia and Canada to reach comparable spending on R&D to healthcare and defence.[1]

Policy makers reported that the quality of primary evidence is high in better-funded areas like health, but lower in more under-resourced areas like housing and community amenities, and public order and safety. Policy makers face similar quality issues with secondary evidence. We found only a small minority of systematic reviews and meta analyses that can be described as high quality, such as those conducted by Cochrane or the Campbell Collaboration.

Improving the supply of evidence could improve public service productivity by identifying more cost effective practices, and improving outcomes for citizens. International collaboration would enable better leverage of domestically produced evidence and reduce duplication or 'research waste'.

Summary of policy maker ratings of primary and secondary evidence quality and quantity across policy areas, broken down by COFOG Level 1 areas of public expenditure. The count refers to the number of ratings received. [Note: defence is unusual for having high quality ratings but low quantity; also note small N respondents]



Barriers to adoption

Policy makers identified multiple barriers to the adoption of evidence in policy making, namely, access, understanding and interpretation. The following key themes emerged:

[1] This figure was calculated assuming an estimated gap of between 1.5%-2% in areas outside of healthcare and defence.

- **Relevance:** Relevant evidence is not always available in priority areas. There is significant variation in the quality and quantity of evidence that can be difficult for policy makers to interpret

- **Timeliness:** Currently, the production of high-quality evidence and evidence synthesis is too slow to match the pace of policy demand

- **Clarity and Context:** Evidence is too often 'buried' under academic language, and research questions are too narrowly focused to be of meaningful use in policy making

- **Rigour:** The inclusion of poor quality studies in secondary syntheses can undermine the value of the secondary evidence. Additionally, policy makers want research to be more tailored to policy applications.

- **Capability and resource constraints:** Policy makers highlighted the need to build evidence skills among public service professions; reduce the cost of finding and using evidence, including removing paywalls; and increase open access to research.

The case for collaboration and recommendations

This report sets out a blueprint to address the issues in our evidence ecosystem. Central to this is leveraging international cooperation to avoid duplication and maximise collective resources. Our recommendations focus on six strategic initiatives to improve evidence-based policy and practice. Collaborating to commission cross-national 'public goods' should bring direct financial savings through cost-sharing, and broader benefits and savings through improvements in the efficacy and efficiency of public services.

A. Increasing the generation and sharing of high-quality primary evidence (estimated costs in brackets):

1. **Establish a Shared Evaluation Fund** across partner countries to ensure evaluation of promising interventions ($10-50 million one-off)

2. **Promote standardised reporting and publication protocols** to facilitate inter-governmental sharing of evaluated interventions ($0.5-1 million one-off)

B. Advancing the quality and relevance of secondary evidence:

3. **Conduct evidence gap maps across priority policy areas** to obtain an overarching view of the state of the evidence ($10-30 million one-off)

4. Prioritise the synthesis of this evidence into high-quality, comprehensive reviews, or **meta-Living Evidence Reviews (meta-LERs)** ($50-100 million one-off, plus a further $10-40 million per annum)

C. Boosting evidence adoption:

5. **Strengthen international public service professional networks** to accelerate the transfer and adoption of best practices across countries ($5-20 million per annum)

6. **Conduct research into effective translation and adoption**, or ‘metascience’, to accelerate the transfer of evidence into policy and practice ($1-5 million one-off)

Options for delivery

A range of alternative implementation options are considered, with corresponding institutional forms and costs (see table).

Global Evidence Fund	**Programme Intensity & Costs ($m)**		
Recommendation	**Full / High**	**Medium**	**Low (MVP)**
1. Shared Evaluation Fund	10-50	5 - 10	-
2. Common reporting	0.5-1	-	-
3. Evidence gap maps	10-30	7 - 15	2 - 4
4. Living Evidence Reviews	50-100 + 10-40 pa	20 - 40 + 4-12 pa	7.5 - 15 + 1.5-3 pa
5. Policy and professional networks	5-20 pa	1 - 5 pa	-
6. Research on applied research, translation and adoption	1-5	-	-
(+MVP) Build ‘kickstarter’ platform			1.5 - 3
Total cost for Year 1	**76.5 - 206**	**33 - 70**	**11 - 22**
Total cost for Years 1 - 3	**106.5 - 326**	**43 - 104**	**12.5 - 25**

The full or **high intensity** option establishes a Global Evidence Fund with total costs of circa $200 million, with its own institutional form, to commission and drive the building of global evidence goods across policy areas. The **medium intensity** option at circa $60 million would fund 6-8 meta-Living Evidence Reviews (meta-LERs), underpinning evidence maps, and establish a smaller, prototype Evaluation Fund. The **low intensity (or MPV)** option at circa $20 million would work as a pilot, testing the approach to evidence maps, funding 3 meta-LERs, and establishing a ‘kickstarter’ style platform. This platform would enable countries, regions and funders to coordinate joint interests and co-funding of specific reviews without committing to a broader pooling of resources. Once a threshold number of requests for evidence in an area were made, this would trigger a funding call to independent funders and from those who expressed interest in the question. We also recommend partners consider match-funding arrangements to help crowd in third sector and specialist sectoral funding.

The coordinated efforts proposed throughout this paper aim to fill critical evidence gaps and, more fundamentally, to foster a more efficient, effective public service landscape across the countries, building a strong culture of empiricism and impact.

It is worth noting that surveys and interviews were based on a small number of senior participants. While this review provides suggestions for areas to prioritise based on these survey results, we would expect a procurement process and negotiations between funders to draw on a wide pool of views.

We are grateful to partner countries and funders for their engagement in this review, and to the Economic and Social Research Council for supporting it.

Acknowledgements

This report was funded by the UK Economic and Social Research Council (ESRC), and was commissioned following discussions between the governments of the United Kingdom, United States, Australia and Canada in July 2023, including:

- UK Cabinet Office Minister Alex Burghart,
- Australian Assistant Minister for Competition, Charities and Treasury Andrew Leigh,
- Officials from the U.S. Office of Management and Budget, Canadian Privy Council and UK Cabinet Office and His Majesty's Treasury
- Former United Kingdom National What Works Adviser David Halpern

The authors would like to thank the Steering Group for their work, challenge and support, including for their efforts on the survey of senior policy makers. Special thanks to:

- **David Gruen -** Chief Statistician, Australian Bureau of Statistics (Australia)
- **Loren DeJonge Schulman -** Program Associate Director, Office of Management and Budget (United States)
- **Conrad Smewing -** Director General, Public Spending, His Majesty's Treasury (United Kingdom)
- **Rodney Ghali -** Assistant Secretary to the Cabinet of the Impact and Innovation Unit at the Canadian Privy Council Office (Canada)
- **Stian Westlake -** Executive Chair, ESRC (United Kingdom)

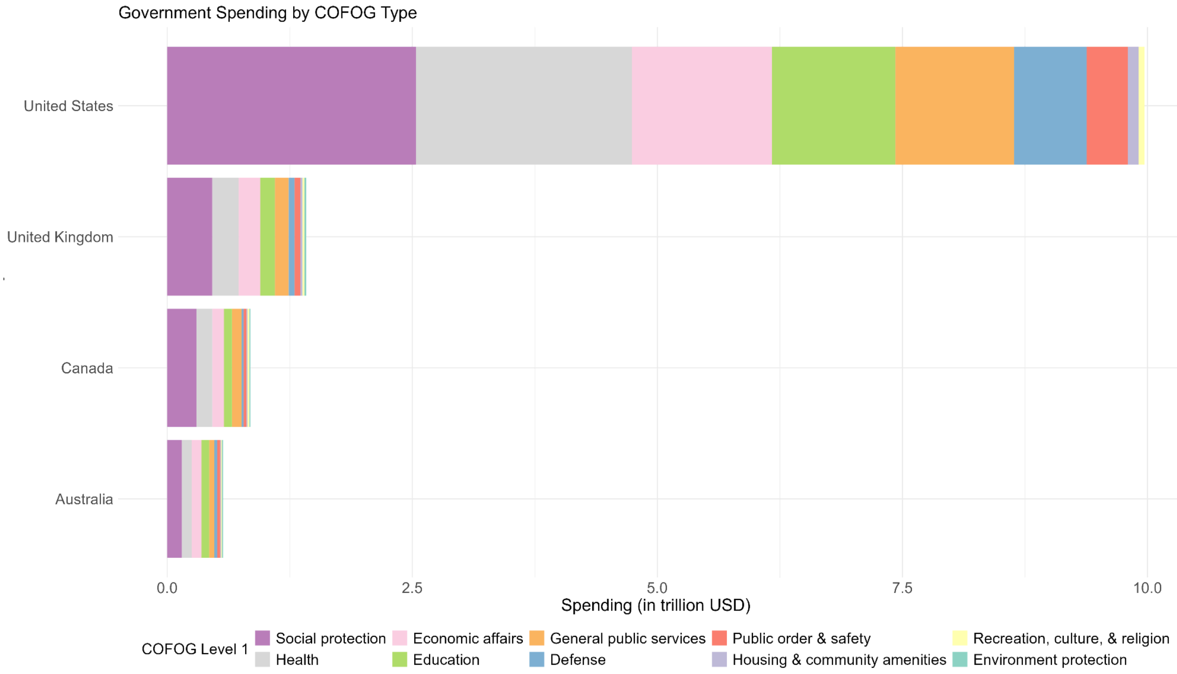
Special thanks also to:

- **Diana Epstein -** Office of Management and Budget (United States)
- **Catherine Hutchinson, Miriam Light & Fay Sadro -** Evaluation Task Force (United Kingdom)
- **Eleanor Williams -** Australian Centre for Evaluation (Australia)
- **Chad Hartnell & Justin Savoie -** Impact Canada (Canada)
- **Jen Gold & James Canton -** ESRC (United Kingdom)

Introduction

Governments around the world spend trillions of dollars every year on public services with surprisingly little evidence on the relative efficacy of expenditure, or 'what works'.

Figure 1: Government spending in 2020, by Classifications of the Function of Government (COFOG) type (based on [International Monetary Fund (IMF) data](#))



Even where evidence does exist, it can be difficult for policy makers to find, understand, contextualise or determine the appropriate action to take. The quality of evidence can be variable, skewed by anecdote, and lacking robustness. This can be aggravated by the pressure on governments to act at pace, relative to the time required to conduct research.

At the same time, the costs and demands on public services are increasing. Our populations are ageing, dependency ratios are increasing, and the expectations of citizens remain or are increasing. This raises a common challenge across countries: how can we do more with less? Against this background, many governments have been upgrading their evaluation capacity and evidence ecosystems to improve the targeting and impact of spending.

Domestic action to improve evidence architecture will be critical. Yet many current challenges - and evidence gaps - are shared internationally. A collaborative approach to addressing these challenges could increase efficiency, improve policy making, and increase the impact and reach of good evidence.

Accordingly, between November 2023 and April 2024 the governments of the U.S., UK, Australia and Canada provided guidance for a sprint review, commissioned by the ESRC, to **identify**:

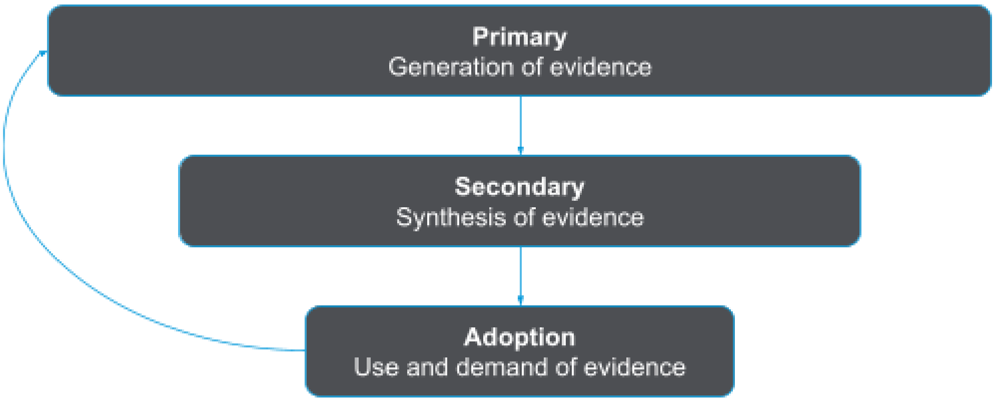
- **Policy areas** where there is greatest demand for, but gaps in, evidence across the four countries;
- **Shared product[s]** to enable policy makers to find and use evidence more easily from across countries;
- **Candidate institutions** to deliver these products and sustain collaboration between countries going forward; and
- **Funders** that could support fledgling activity in this space, and **indicative costs** for the activity to address key evidence and institutional gaps.

This report summarises the findings of the sprint review, providing a blueprint for countries and funders to collaborate on evidence going forward.

A Framework for Evidence

The 'evidence ecosystem' refers to the generation of evidence, its translation, and its use. We used the following basic categorisation in our sprint review:

Figure 2: Framework for evidence



Primary evidence refers to the original data and documentation collected directly from research, observation, or experimentation. It may be generated through quantitative and qualitative research methods. Primary evidence is produced by many sources: academics, providers, interest groups, not for profit organisations, companies, professional associations, citizens, and by governments (local, state, and national). It can also draw on the knowledge, wisdom and agency of communities themselves. The quality of this evidence varies, influenced by factors such as methodology, scale, and potential biases from vested interests. Various scales exist to rate the relative quality of primary research (e.g. the Jadad scale, GRADE, the ESSA tiers of evidence for education, the Nesta Standards of Evidence, and Maryland Scientific Methods Scale).

Secondary evidence refers to the collation and synthesis of primary evidence to summarise and clarify existing research. Secondary evidence, including systematic reviews and meta-analyses, rely on having a robust body of primary evidence on which to build. These reviews use rigorous methodological approaches to bring together the available evidence in a field of research, weighing the evidence based on the scale and reliability of the methods used. High-quality secondary evidence also clarifies the geographies and populations studied, allowing readers to better assess the applicability or transferability of the findings to different contexts.

Evidence adoption refers to the integration of evidence into policy advice and decision making. Policy and decision makers rely on both primary and secondary evidence in addition to other considerations such as: politics, public attitudes and acceptability; affordability; and public service capability.

Policy makers play a key role in influencing the production of research itself, through setting the research and regulatory environment, providing funding, and identifying research priorities, such as expressed through the UK's Areas of Research Interest and the U.S.' Learning Agendas. These research priorities indicate policy makers' demand for evidence in particular sectors, thereby providing steers to research activity. However, such steers are just one among many drivers of the broader research agenda, alongside a mixture of academic, public and funder interests. Hence, evidence generation does not always mean that policy makers and the public needs are being met.

Our review focused particularly on the **secondary** and **adoption** layers of the framework. Our stakeholders considered these areas to be the most fruitful for cross-national collaboration.

Methodology

Guided by a **Steering group** and a **Working group**, our review comprised:

- **Desk research** to understand current demand for and expenditure on primary and secondary research across the four countries;
- **A survey** of 36 senior policy makers across the four countries to identify common areas of research interest;
- **Semi-structured interviews** with 10 policy makers and researchers on key evidence gaps, in addition to informal conversations with stakeholders including potential funders.

Appendix 2 contains further details on the methodology. Appendix 5 contains a glossary of terms.

Our review identified significant data gaps on research activity and expenditure itself. We relied heavily on the OECD Classifications of the Functions of Government (COFOG) categorisations of government activities and spending, but found that even in our four partner countries, comparable data was not always available. For example, data for Canada's Research and Development (R&D) spending by COFOG is not available within the OECD datasets. Further, certain policy areas are not included within the OECD R&D expenditure dataset (e.g. public order and safety spending).

Limitations

We supplemented the data with additional research. Supported by our Steering Group, we undertook a survey of senior policy makers across the four countries to draw out experiences of and insights on the quality and stock of primary and secondary evidence across sectors, including its accessibility and impact.

This was a modest size survey of 36 senior policy makers, meaning we had limited coverage of all spending areas. This limitation hinders our ability to present a comprehensive view of the overall evidence landscape and use in these countries. It should be noted that we had significantly more respondents from Australia and the UK than the U.S. and Canada. This should be considered when interpreting results, especially for readers from those countries. Because of these sample sizes, we focus mainly on analyses by COFOG areas instead of showing breakdowns of the survey by country. While this review provides suggestions for areas to prioritise based on the survey results, we would anticipate this being refined through procurement and negotiations between funders.

There are significant data gaps. Many countries have only recently started or are still yet to use the OECD Classifications of the Functions of Government (COFOG). This means that for certain areas, governments may hold R&D data domestically but it will be under a different heading to the ten categories within the COFOG. As such, we were only able to identify five of the ten areas in Figure 4 below where there was sufficient data to allow for cross-national comparison. Data for Canada was not available via the OECD. Our own analysis of other COFOG areas using domestic data sources, indicates that the R&D expenditure in these remaining areas is low, typically less than 0.25% and for some major areas of public service activity, less than 0.1%.

Interviewees and survey respondents represent the perspectives of a small, though senior, group of individuals. Given the scope of the review, no single individual will have the depth of experience to cover all policy areas, and in some cases even to cover a single area at the COFOG Level 1. As such, caution should be taken to not overinterpret the views expressed, particularly for survey ratings that were based on the views of one or two respondents.

Structure of this Report

This report is structured in four sections.

➔ Section 1 examines **primary and secondary evidence supply gaps**, presenting administrative data supplemented by policy makers' assessments.

➔ Section 2 focuses on **adoption challenges**, bringing together survey results and expert interview insights.

- ➔ Section 3 presents the **case for collaboration and our recommendations**, outlining the candidate products and services.
- ➔ Section 4 presents **options for delivery,** outlining the options for institutional delivery and program costs.

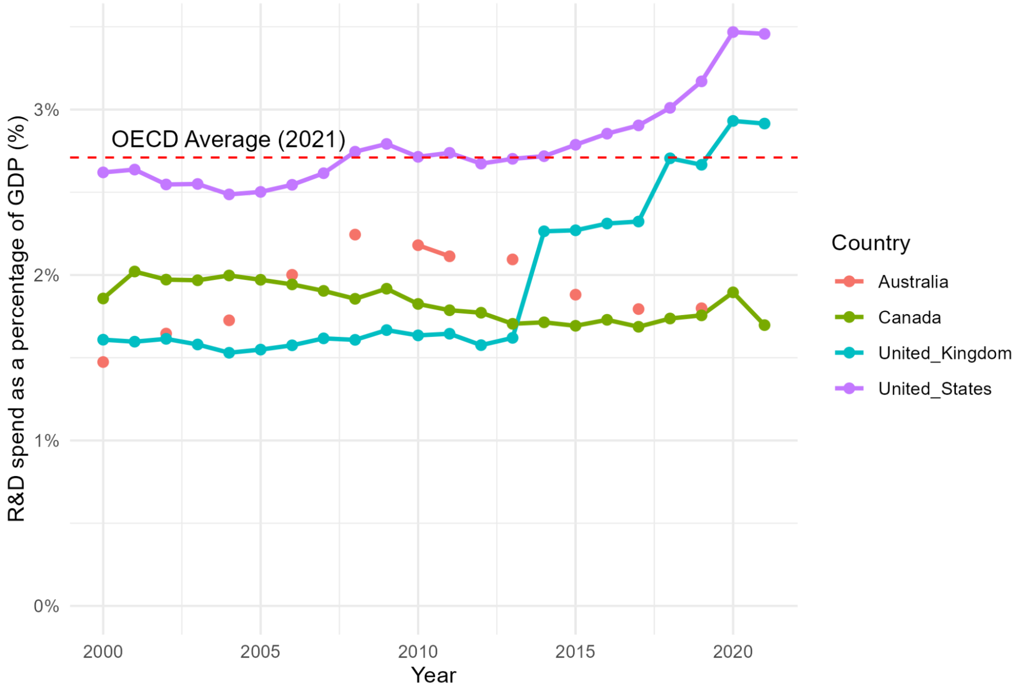
Section 1: The Supply of Evidence

> *“In most policy domains there is limited quantity and variable quality”*
>
> *“They are not bad studies but there is just a lot to filter - we need intermediaries and brokers”*

Global Spending on Evidence Production

On average, total R&D expenditure in the four countries has been rising. Current headline expenditure on R&D provides an indication of the current supply of evidence, noting that includes R&D directed and generated by private sector activity. It is far from a perfect measure - annual expenditure data indicates a level of investment in evidence production but not the value in terms of quantity or quality. These figures also encompass *development* activity (innovation) in addition to primary and secondary research, and there have been changes in definitions (such as in the UK data). However, it does provide an indicative time series and sense of overall magnitude.

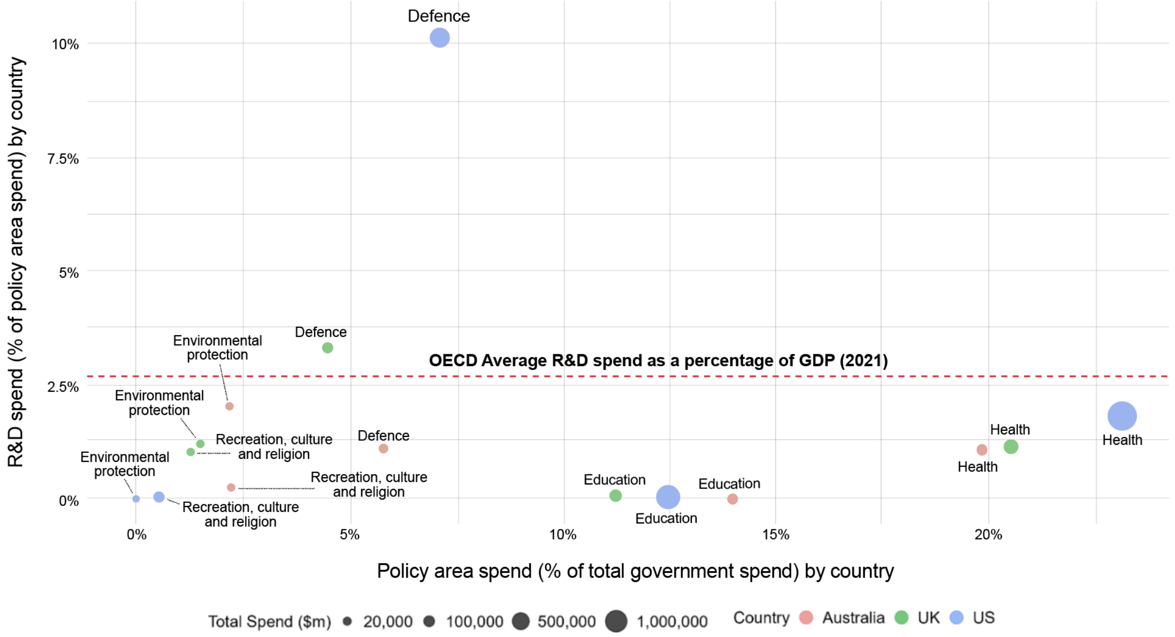
Figure 3: R&D spend as a percentage of Gross Domestic Product (GDP) by country (based on OECD data)



With respect to public sector R&D, research spending is extremely uneven across policy areas, and to some extent across countries. Breaking down investment by COFOG indicates that across all four countries, R&D spend is dominated by Defence and Health (see Figure 4). The U.S. is a particular outlier on defence spending, both in absolute expenditure and the proportion that is directed to R&D (more than 10%). Australia spends a relatively large percent on environmental protection. Health stands out across countries as an area of large expenditure, in absolute and percentage terms as a function of health expenditure: 1.84% (U.S.); 1.15% (UK), 1.10% (Australia).

There are major policy areas where relatively little is spent on R&D. Education is a large activity and focus of governments, typically 10-15% of total government expenditure or just over half of that spent on health. Yet the ratio of percent expenditure on R&D in health compared with education is x46 (U.S.), x18 (UK), and x142 (Australia). This is despite evidence of extremely high returns on investment achieved by good quality educational research (see box 3, in recommendations). These large differences remain when factoring in non-government research funding (see Appendix 3). In the UK, for example, the Vallance review (2019) highlights how “science activity and expenditure is variable across government” and “spend on R&D in some cases is a fraction of one percent of total spend”. The thinness of research in such areas makes it even more important that available evidence is shared across countries.

Figure 4: R&D spend compared to total expenditure by policy area (Australia, UK, U.S.) 2021-22 (based on OECD data: public finance by function & government budget allocations for R&D)



The implied ‘research gap’ is substantial. Many western governments are seeking to lift overall R&D levels to boost growth and productivity. If the same logic applied to the circa 25% of GDP currently spent on public services outside of health and defence, we can calculate an implied public service ‘R&D investment gap’. The implied ‘R&D in public services gap' is around $85-115bn pa for the U.S., UK, Australia and Canada to reach comparable spending on R&D to healthcare and defence. Plugging this gap is a matter for domestic governments to consider. Against this huge gap between expenditure and R&D, the little research evidence that does exist is even more valuable as a source of learning between and within countries.

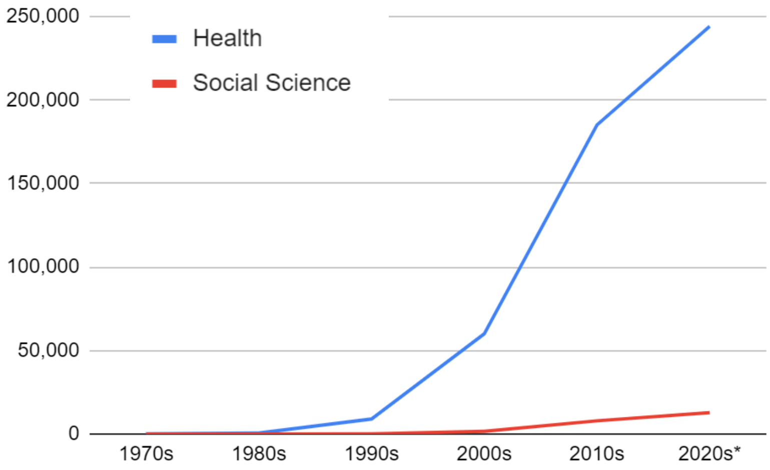
Low historic levels of policy evaluation is an additional indicator of a limited supply of evidence, and perhaps low demand. Alongside low spend on R&D, there is very limited robust evaluation directed at that spend. One UK study found that just 8% of government spend on major projects - £35 billion of £432 billion - had robust evaluation plans in place in 2019. Results for America found an average budget spend of 1-2% on evaluation in a selection of U.S. agencies. Increasing the volume of robust evaluations of new and existing policy practices and programs is an important, and likely high return, element of R&D.

The Quantity and Quality of Evidence

Primary evidence

Primary evidence has increased in all fields over time, particularly high quality evidence, albeit from a low base. High quality evidence means evidence that is based on robust methods, ie that is: valid (both internally and externally); reliable (it can be replicated), relevant (it answers the question at hand), credible (e.g. it has been peer-reviewed), and bounded error (e.g. has been significance tested). Randomised Controlled Trials (RCTs) often are considered the gold standard of such evidence as they can be used to establish causality, and meet many of the above criteria. Although the increase of high-quality evidence has generally increased across fields, it has occurred at differing rates across sectors. Health remains many leagues ahead of other policy areas and social sciences.

Figure 5: Number of RCTs in Health vs Social Science (produced by the Campbell Collaboration)



Across partner countries, the education sector is an example of an area that has improved the production of high quality evidence in recent years, now accounting for ~20% of the RCTs within social science presented above. However, even within education, interviewees remarked on the variance between different areas, such as the continuing scarcity of evidence in higher and tertiary education compared to early years education.

Policy makers provide mixed ratings of primary evidence quality and quantity across sectors. We asked policy makers to provide their ratings of the quantity and quality of primary and secondary evidence in the areas they worked, using the COFOG classification.

- Policy makers speak positively about the quality and quantity of some areas of research (despite their low volume and expenditure) such as economic affairs and environmental protection.
- Policy makers were more negative about the quantity and quality of primary research in areas such as housing and community amenities, and public order and safety.

Figure 6: Primary evidence quantity and quality from survey of policy makers, per COFOG Level 1 (data as of 30 March 2024)

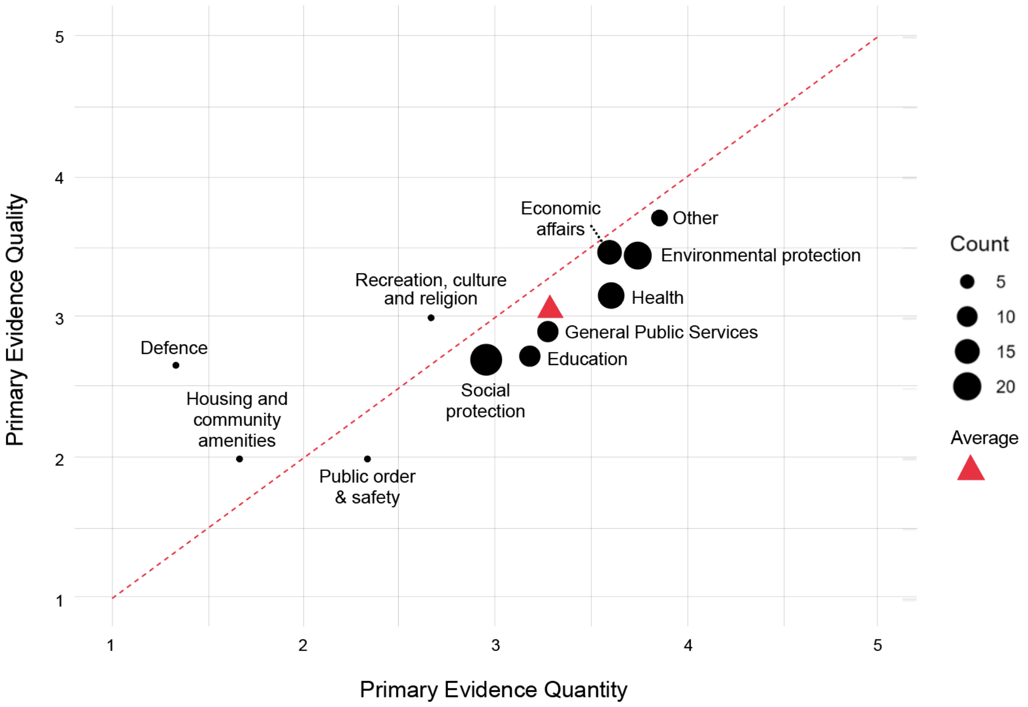
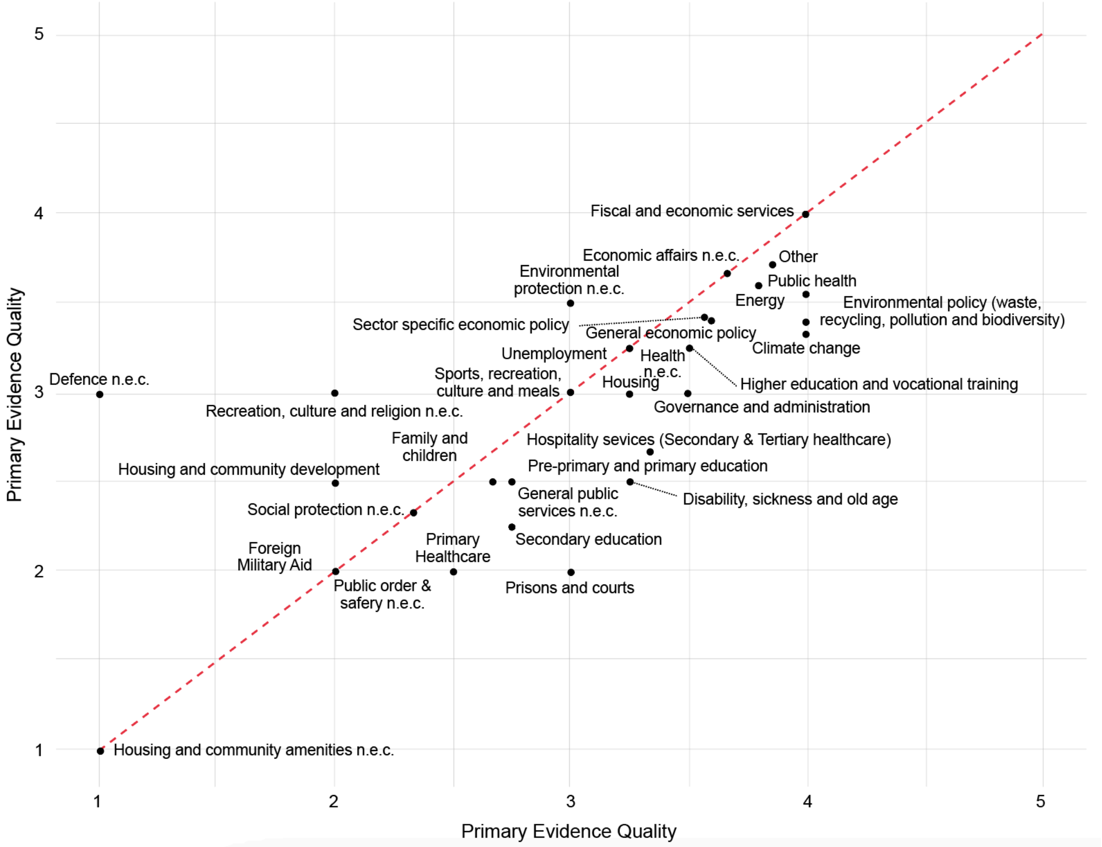


Figure 7: Correlation between primary evidence quantity and quality from survey of policy makers, per COFOG Level 2 (data as of 30 March 2024). [Correlation coefficient r = 0.74, adjusted R^2 = 54%, p = 0.001]

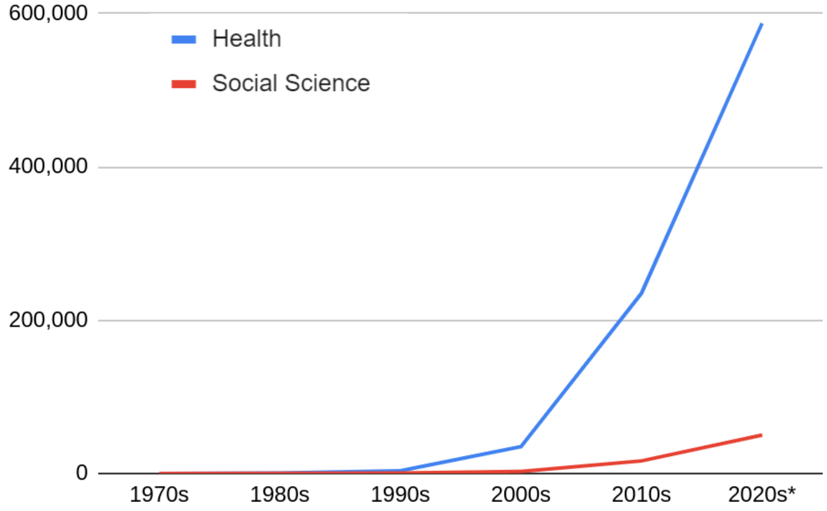


Secondary evidence

Secondary evidence plays a crucial role in helping policy makers to sift through large volumes of primary evidence - providing a synthesis of the evidence in an area of research and a summary of its quality.

As with primary evidence, improvements in the quality and quantity of secondary evidence have occurred at differing rates across sectors. The graph below depicts the change over time. The plot of secondary syntheses over time (Figure 8) echoes even more dramatically that of primary evidence (Figure 5), with rising numbers of syntheses between the 1990s and 2020s, but a large and growing gap between health and social sciences.

Figure 8: Number of syntheses in health vs social science (produced by the Campbell Collaboration)



While the quantity of secondary syntheses is increasing, quality remains variable. For example, Cochrane and Campbell are among the most renowned organisations conducting evidence synthesis. Just 2% of the reviews (above) in health were conducted by Cochrane (11,548 reviews) and around 1.3% of the reviews in social science were conducted by the Campbell Collaboration (259 reviews). As the two largest evidence synthesis organisations, this suggests many smaller organisations are producing ad-hoc syntheses which may lack the same frameworks and codes that exist for the larger organisations. Interviewees also commented on the variable quality of secondary evidence.

There appear to be more secondary syntheses (Figure 8) than high quality RCTs (Figure 5) in both health and social sciences. Our experts remarked that many of the secondary reviews currently available contain a lot of 'noise' (see next section). This is likely due to a combination of:

- A significant duplication of syntheses;
- A number of syntheses including low quality primary work (and/or practitioner judgments);
- Systematic reviews that focus on less robust types of primary research; and
- RCTs (and other experimental designs) not being properly identified and differentiated in the literature.

As with primary evidence, policy makers provide mixed ratings of secondary evidence quality and quantity across sectors:

- Health was once again at the top of policy makers' ratings. Housing and community amenities, public order and safety, and social protection were rated much lower. Defence drops to the bottom of this combined ranking, driven primarily by the low volumes of reported research - though it is important to note the small number of ratings provided.

- Public order and safety, and recreation, culture and religion were rated higher for secondary evidence than primary evidence. In contrast, environmental protection and economic affairs scored somewhat better for primary than secondary research. The latter perhaps hints at prioritising such areas as 'quick wins' for secondary reviews - especially if there is a reasonable volume of good primary underserved by translational summaries.

Figure 9: Correlation between secondary evidence quantity and quality from survey of policy makers, per COFOG Level 1 (data as of 30 March 2024)

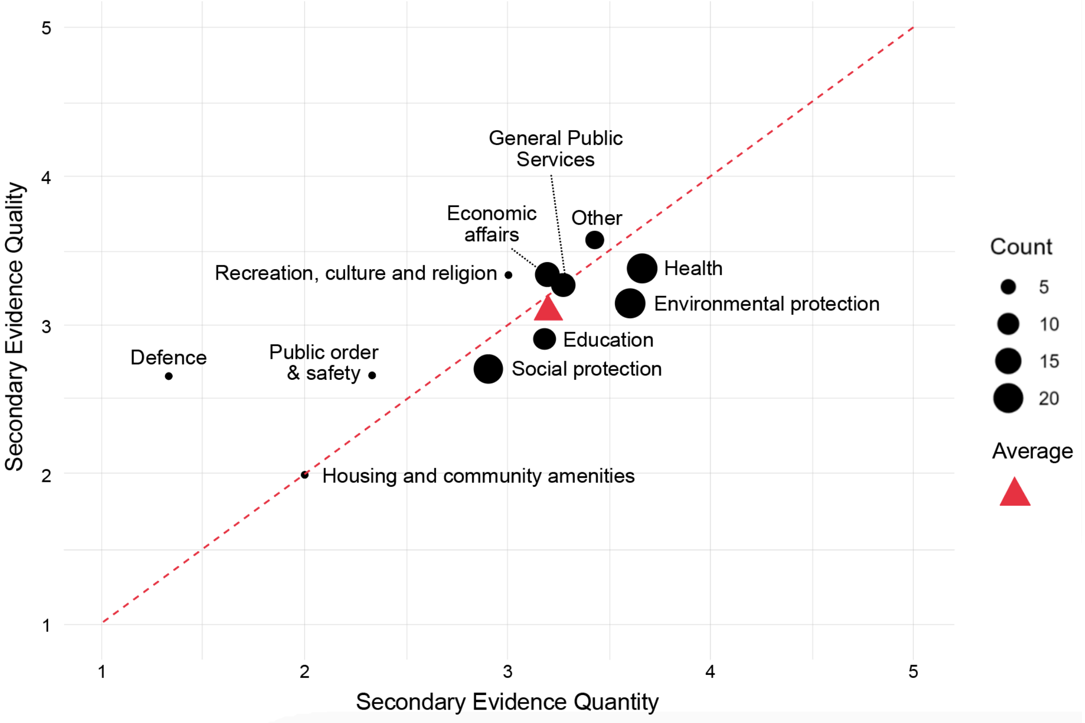
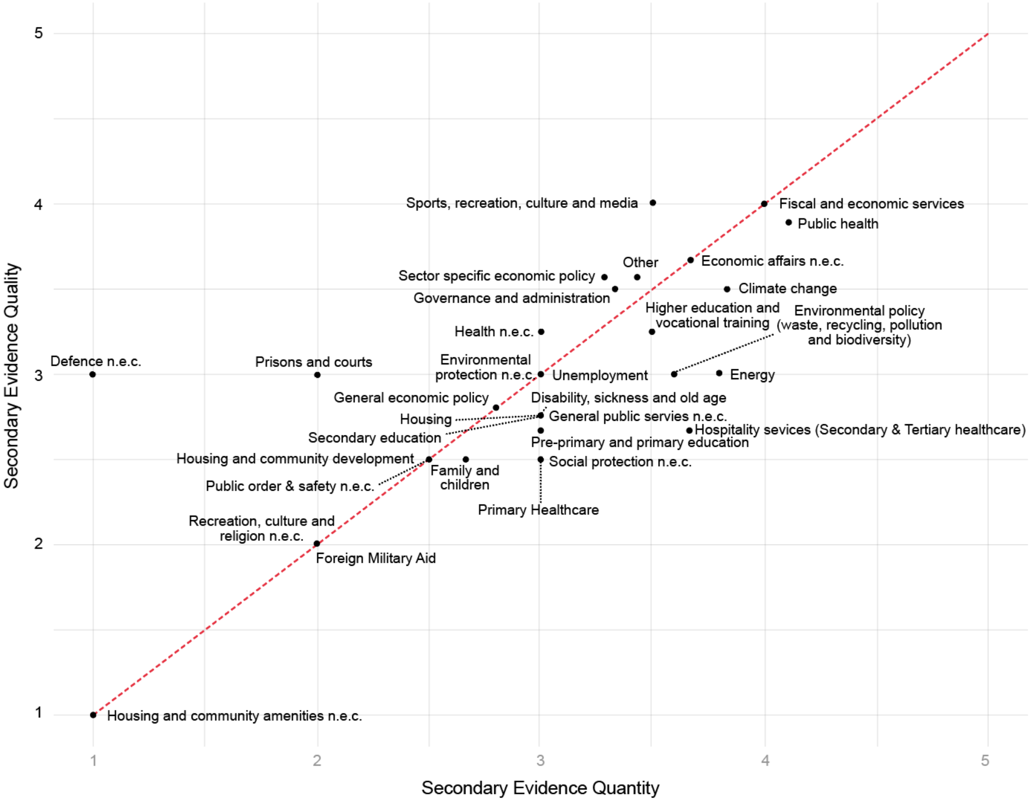


Figure 10: Correlation between secondary evidence quantity and quality ratings from the survey of policy makers, at COFOG Level 2 (data as of 30 March 2024) [Correlation coefficient r = 0.71, adjusted R^2 = 50%, p = 0.001]



An overall assessment of 'evidence strength'

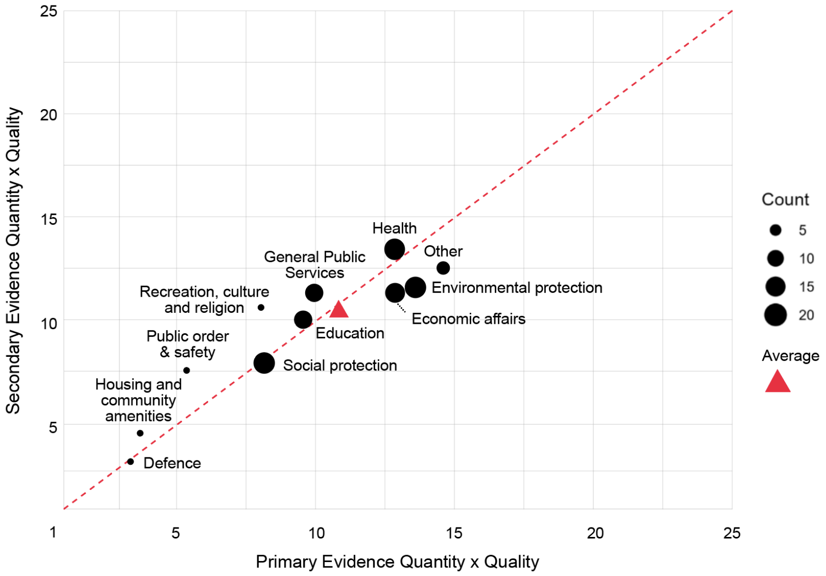
We found that there was a strong correlation between respondents' ratings of quantity and quality of evidence at both primary and secondary levels (0.74 and 0.71). This may reflect the differential maturity and quality of evidence generation across policy areas: as a field matures, it will produce both more volume and quality of evidence.

We used this observation to produce a summary 'evidence strength' measure by multiplying ratings of quality by quantity, for primary and secondary evidence. We then plotted these against each other to examine the relationship. This is shown in Figure 11.

As can be seen, there is again a strong correlation. Policy makers generally rated areas that have better evidence strength at primary level, also as having better evidence ratings at secondary level. Areas such as health, environmental protection and economic affairs were generally rated as having better evidence strength. In contrast, areas such as social protection were rated as having poorer overall evidence strength.

Note that defence is unusual for having high *quality* ratings but low *quantity* ratings. This leads to lower overall 'evidence strength' ratings on our combined measure. As in previous figures, it should also be noted that a number of areas received relatively few ratings, such as defence, housing and community, and recreation, culture and religion.

Figure 11. Summary of policy maker ratings of primary and secondary evidence quality and quantity across policy areas, broken down by COFOG level 1 areas of public expenditure. The count refers to the number of ratings received.



This plot arguably gives a good 'view from 10,000ft' of policy makers' overall assessments of the 'evidence strength' of different policy areas.

Summary

The data from R&D expenditure and the survey of senior policy makers gives a broadly consistent picture:

- R&D expenditure has been rising, but at an extremely uneven rate across policy areas. There is, however, still an implied research gap of c.$85-115 billion to raise public sector R&D in most policy areas to levels comparable with healthcare and defence.

- Health stands out as the area of domestic policy with highest research expenditure relative to total service expenditure across countries, and most favourable overall policy maker ratings on quantity and quality (noting significant unevenness within aspects of health research);

- A number of areas of domestic policy and practice - notably housing and community, public order, and social protection - stand out as having both low expenditure and low quality and quantity ratings. Such areas attract extremely low levels of research expenditure relative to the billions spent on direct delivery, with R&D typically less than 0.25% of total expenditure, and often less than 0.1%.

- Other areas, such as economic policy, achieve good ratings relative to modest expenditure (R&D as a % of domain expenditure).

- The plot of primary (quality x quantity) x secondary (quality x quantity) provides a reasonable “view from 10,000ft” of the current supply of evidence in our four countries. **It should be noted that these ratings - particularly at COFOG Level 2 - are based on a modest number of policy makers. Nonetheless, the Steering Group felt this plot gave a fair and plausible assessment of the overall position.**

Section 2: Barriers to Adoption

> *“[I’d love to see] my inbox full of emails from senior policy makers asking for the latest evidence on X ”*
>
> *“[In terms of adoption] we do better on the practice side than the policy side…we are getting better but still putting evidence in front of policy makers rather than them understanding and asking for evidence.”*

Policy makers and practitioners express their research interests in a number of ways. Most obviously, government departments and public bodies across countries can directly commission research to answer particular questions. This research is sometimes referred to as ‘commissioned’, ‘applied’ or ‘directed’ research to distinguish it from basic or ‘pure’ research. For example, applied research might ask ‘how best to reduce the use of illicit drugs?’ while pure research might ask ‘what is the structure of this chemical, and how does it interact with the brain?’

Policy makers can also direct research more strategically, through setting up institutions, research commissioning bodies, or even publishing lists of research priorities. The U.S. ‘Learning Agendas’ and the UK ‘Areas of Research Interest’ - both relatively recent and novel developments - are examples of policy makers publicly indicating their demand for research in particular areas. These priorities are intended to signal to research communities areas to target, and can be used by researchers to demonstrate to funders that their work is likely to be of interest and have impact. However, there is an open question as to how complete these lists of interests are. More than one senior policy maker remarked that government bodies do not always want to signal gaps in their knowledge, especially in areas of political sensitivity.

We asked policy makers to rate how much evidence was used in policy decision-making, and how easily they could access, understand and interpret evidence. Policy makers in most fields were optimistic about the impact of evidence in policy decision-making. Nonetheless, access, understanding and interpretation of evidence was identified as an issue across almost all policy areas.

Figure 12 and 13 show that there is significant variability in perceptions of the impact of and access to evidence in different policy domains. There is a positive correlation between ratings of impact and access to evidence. While we cannot attribute hard causality to these findings, it is reasonable that improving access might improve impact.

Figure 12: Correlation between impact of evidence on policy decision-making and access to evidence, per COFOG Level 1 (r=0.40; rsq=15%, n.s., data as of 30 March 2024)

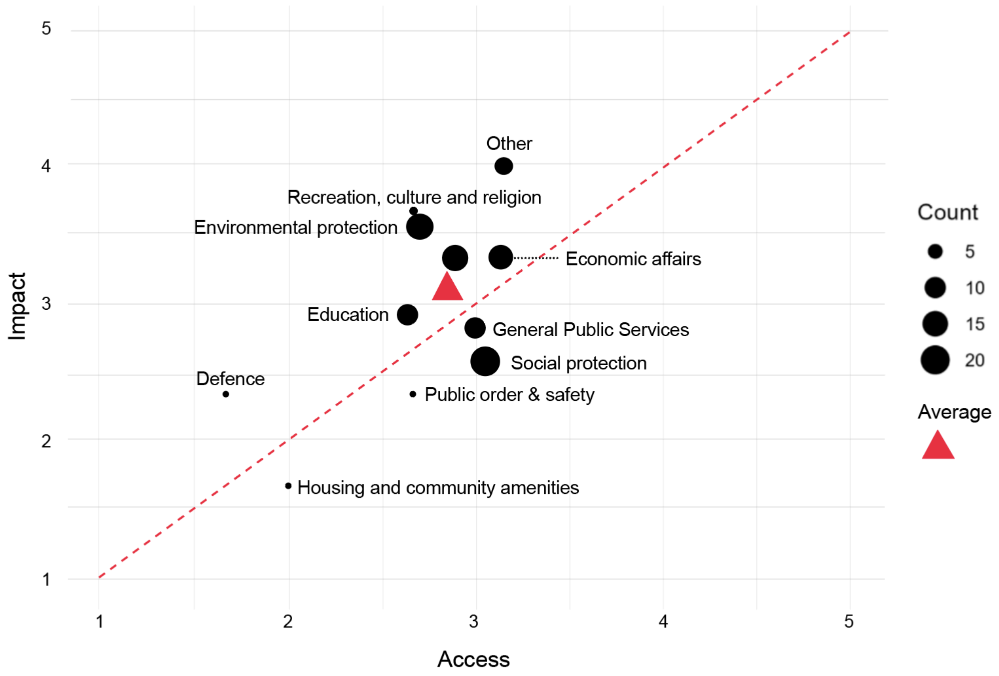
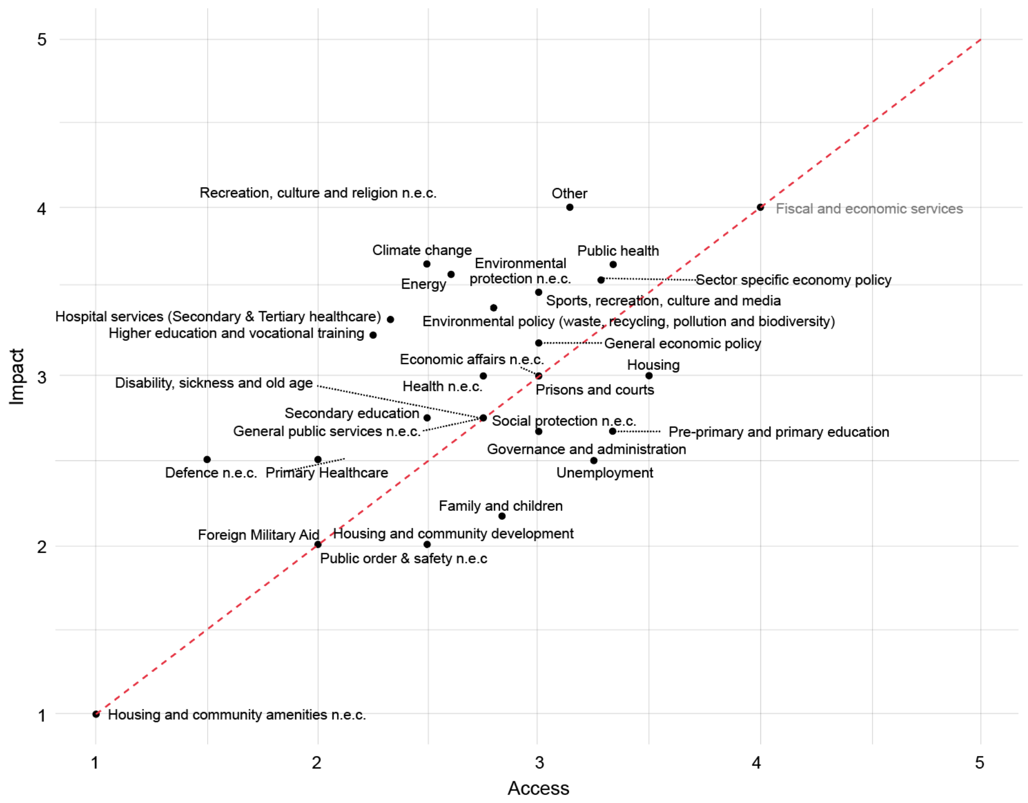


Figure 13: Correlation between impact of evidence on policy decision-making and access to evidence, per COFOG Level 2 (data as of 30 March 2024)



We asked policy makers to provide free text responses detailing the barriers they face to the adoption of evidence to supplement the interviews. **Qualitative text analysis on responses revealed the following themes:**

- Relevance
- Timeliness
- Clarity and Context
- Rigour
- Resource constraints

Each of these themes are explored below.

Relevance

> *"If we create a study, even an RCT, and it has a clear impact and we know the population and it's scalable, we assume we've done our job. In our experience, that's not the case at all. Even a quality research study is just another bottle floating in a sea of other research studies."*

Relevant evidence is not always available in the areas policy makers need, despite high profile statements of policy research interests (e.g. the U.S. Learning Agendas and UK Areas of Research Interest). Interviewees stressed that evidence users and evidence producers are not always well aligned.

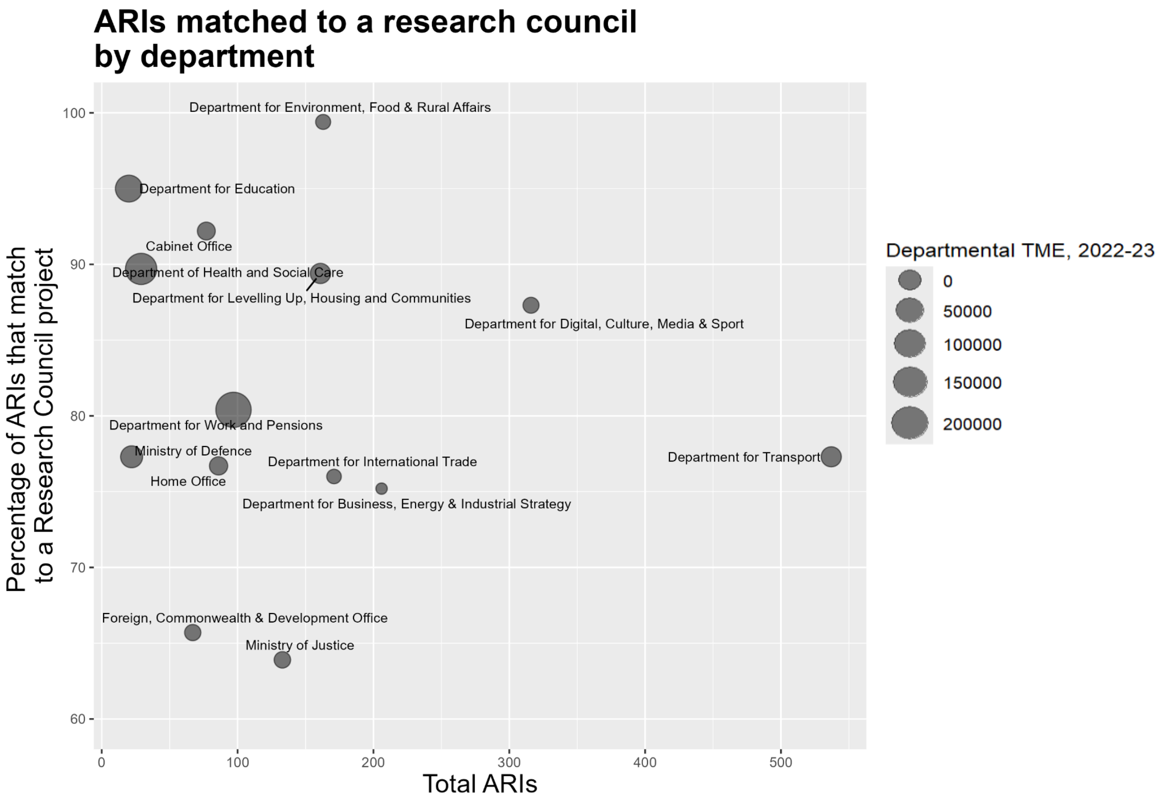
A matching exercise between the UK Areas of Research Interest (ARIs) and research activity by UK Research Councils reveals priority research gaps (see Figure 14).

- Analysis, conducted on behalf of the funding body UK Research and Innovation (UKRI), shows that most policy questions have at least some research conducted in that domain. However, a relatively high proportion of questions (i.e. around one third of questions) in some departments such as the Ministry of Justice and Foreign Commonwealth and Development Office have no identifiable research in that topic.

- Departments also differ greatly in the number and specificity of research questions they would like answered, ranging from a small number of high level questions (e.g. Department of Health and Social Care or Ministry of Defence), through to a large number of relatively specific questions (e.g. Department for Transport).

- Even where there are high matches between ARI questions and UKRI activity, there remain access challenges (as presented in Figures 12 and 13). For example, education has a 95% match, yet the survey of policymakers rated access to education evidence at less than 60%. This aligns with some policy makers' views that even if there is research in a given area, it does not always answer their particular questions, or research stays in the academic sphere without reaching policy makers or practitioners.

Figure 14: UK Areas of Research Interest matched to UK Research Council activity, by Government department [data provided by the ESRC & Overton]



Researchers themselves may not have a clear sense of where gaps are in the evidence base. One interviewee commented on how research tends to follow areas where there is already existing evidence, leaving fields without any existing evidence under-studied. This supports the case for more systematic 'evidence maps', to help both policy makers and researchers target research efforts [see recommendation 3].

The evidence landscape is 'noisy', meaning that high quality evidence can be lost in a sea of low quality evidence. This creates friction and confusion for policy makers. A lack of high quality syntheses means that primary evidence often sits in isolation, making it difficult for policy makers to identify. For areas with sufficient high quality primary evidence, we need high quality syntheses to make robust key findings easier to find [see recommendation 4].

*Box 1: **Machine-learning to reveal demand by policy makers***

Researchers could increase their understanding of the needs of policy makers and align research to these needs. A machine learning exercise to uncover overlapping 'revealed preferences' among policy makers could add significant value to areas urgently needing more evidence. This could draw from press releases, parliamentary or Congress debates and a range of other publicly available information to mine for common areas of interest across countries.

This might be further supplemented by machine learning powered analysis of the quantity and quality of research referenced in government documents, building on the coding frame and pilot developed by Sense about Science and the UK Institute for Government.

Timeliness

> *"Emerging topics like climate change have huge gaps - the science is known but human activity and behavioural change [have] rapid production gaps"*

Respondents to the survey frequently mentioned the lack of timely evidence for policy makers working to short timelines. Evidence synthesis can be time-consuming. The median time to publication of a traditional systematic review by Cochrane is 2.4 years.

To provide answers at speed, some evidence institutes have developed rapid response teams, such as The Australian James Martin Institute for Public Policy Rapid Response Team, or the U.S. Societal Experts Action Network (SEAN). However, these teams are limited in the volume of evidence they can draw on within a short period. They may, therefore, fail to draw from all the available evidence, or may struggle to accurately determine which evidence meets an appropriate quality threshold for a systematic review.

There is a need for faster, reliable synthesis [see recommendation 4]. Artificial intelligence is already playing a role in reducing timelines for evidence production (see Box 2: The Future of Evidence). Ideally, the main areas of policy would have reviews already done in anticipation of likely policy interest.

Clarity and Context

> *"As a community we have put in very little time into evidence application"*

The adoption of evidence into policy making can be hindered by how the evidence is presented. Traditional academic text and language can make it difficult for policy makers to extract the most pertinent information from a review.

Further, policy makers are generally interested in answers to a broader set of questions, rather than narrower questions on the efficacy of specific interventions. For example, traditional systematic reviews tend to focus on a question such as 'does Scared Straight work?' but policy makers and practitioners tend to ask questions from the 'other end of the telescope' such as, 'what is the most effective way to stop young people drifting into crime?'. This is where clearinghouses play a critical "translation" role.

Reviews should both answer broader questions and be presented in a way that enables policy makers to act upon the evidence [see recommendation 4].

Rigour

> *"Most things don't work….the amount where there is enough there to justify public investment is very low ("really tiny")... [and it is] lost in the noise from the clearing houses."*

Policy makers emphasised the high standards that should be imposed for accepting primary evidence in a secondary review. This should include triangulating across a range of different types of evidence, including quasi-experimental studies, mining of variance, and qualitative approaches that help unpack mechanisms and the experiential and empowerment aspects of interventions. Poor quality primary inputs can undermine the quality of secondary outputs.

Policy makers also highlighted methodological problems that are constraining the adoption of evidence. Policy makers pointed towards the inflexibility in the design of studies, the overreliance on qualitative data, and the lack of longitudinal studies.

Much more can be done to ensure that international evidence is adopted. Variation in practices across (and within) countries can provide powerful insights and inspiration, though caution has to be taken around other differences that may limit the transferability of that evidence. In our interviews, most policy makers admitted they do not draw from evidence outside their own country - even among a sample that was drawn from four closely allied countries, with common language and broadly similar governments. Reviewing c.95,000 citations in 28,000 policy documents, one paper found that most studies are not referenced significantly in the policies of other countries:

Table 1: The frequency that studies are referenced in policies outside of the home country / institution (source: Mahfouz, B, Mulgan, G & Capra, L, 2022)

	% references in Intergovernmental Organisations (IGO) policies	% references in U.S. policies	% references in UK policies	% references in EU policies
Intergovernmental Organisations (IGO) studies	N/a	10.6%	12.57%	13.5%
U.S. studies	4.1%	N/a	5.4%	4.2%

UK studies	3.9%	4.3%	N/a	5.4%
EU studies	4%	3.16%	5.13%	N/a
Common across all entities	**0.62%**			

Risks of translation should be balanced with advantages of drawing from cross-national variation. Where there is little national evidence in a policy area, it is greatly preferable to review the evidence elsewhere rather than start from scratch. It is important to maintain methodological rigour, including on when and how to generalise results from one setting to another. This methodological caution should be balanced with the huge potential of learning from other approaches. 'Inspiration', rather than 'copy-and-paste' is a good way forward. This should be combined with local consultation, testing and replication.

Capability and Resource Constraints

> *"[We need] a strong and active research community - people aren't being well trained. There is quite a job to do to build the skills."*
>
> *"Broader challenge is coordination and dissemination. Understanding and utilisation of evidence is incredibly nascent"*
>
> *"Political capital, money - [these are] the currencies that drive real behaviours"*

The survey and interviews highlighted the skill, time and cost constraints faced by policy makers in adopting evidence:

Skills

- **Toolkits and guidance notes are not sufficient.** Governments have developed these to increase the use of evidence by policy makers (e.g. this toolkit by the U.S. Department of Labor). However, one interviewee suggested that these have been ineffective to-date and that there has been an over-reliance on toolkits.

- **Capability and awareness are required to maximise the value of existing investments in the evidence ecosystem.** Creating well-presented evidence is not enough. For example, one interviewee discussed Results for America's Economic Mobility Catalog, which uses a well-presented interface to allow users to quickly assess the evidence across a range of areas for improving economic mobility. However, according to our interviewee, the Catalog is not being used.

- **Nurturing the skills of policy makers is imperative.** Evidence-curious professionals will demand more evidence, making tools such as the Catalog much more likely to be used. Evaluation skills training can nurture an ability to understand evidence and determine the most appropriate interventions to adopt. Greater scrutiny on the evaluation of new projects will build the evidence base over time, and bend the spend towards well-evidenced projects.

- **Countries could valuably scale-up training that supports better evidence use,** including on methods, practices and use of evidence within governments. This could include establishing ‘evaluation schools' to train policy makers in core methodological skills. These can be drawn from and build upon early efforts within partner countries, such as the Evidence-based Decision-Making Leadership Academy in the U.S., and the UK's Evaluation Task Force's Evaluation Academy. Wider efforts to scale-up training could draw from the work of the Australia and New Zealand School of Government (ANZSOG) work to upskill senior public servants as well as the Australian Public Service (APS) Academy. Such evidence-generation and use programs can be created within countries. Even if they do not require cross-national collaboration, there is room to share materials, use-cases, and learn what types of programs are most effective [see also recommendation 6].

- **Increasing demand for value-for-money in public policy is likely to increase the profile of evaluation for policy makers**. This may shift robust evaluation towards a more central role within policy decision-making, along with demand for the relevant analytic and trial design skills.

- **Cross-national collaboration could be used to create international fora for policy makers to convene and share ideas** [see recommendation 5].

Cost

- **The cost of finding and using research can be a significant barrier to its adoption.** Primary evidence is frequently behind a paywall (for example, in expensive academic databases). Government evaluations, even within a domestic setting, can be similarly difficult to access. Gaining access to evidence can therefore take significant time and cost.

- **One way to optimise the allocation of resources is to pool them across countries** [see recommendation 1].

Sources of evidence

- **Policy makers noted they most commonly use government evaluations,** often from their own department, and evidence from other government institutions. This includes using data collected directly by government departments. The lower use of external sources of evidence was reinforced by barriers to accessing wider sources of evidence. For example, the costs of research platforms and journal access, even if these are often ultimately funded by the public sector.

- **Researchers should be continually encouraged to publish papers with ‘open access’.** UKRI and the U.S. National Institutes of Health have made this a requirement for their funding. The European Commission has created Open Research Europe, allowing Horizon Europe beneficiaries to access research for free.

Box 2: ***The future of evidence***

> *“[We’re on the] precipice of a new way that evidence is going to be synthesised and distilled into make it practical for leaders - wave of LLMs and AI that will wash over this - real opportunity to build upon that”*
>
> *“[There’s] a lot of talk about innovative things but [we’re not seeing] as much that’s groundbreaking [in practice] - info [is] not coded or prepared well enough to be useful for AI”*

It would be remiss not to mention the role of AI when thinking about the future of the evidence ecosystem. Almost every interviewee mentioned the potentially “transformative” role it could play, both in the synthesis of evidence, and in driving up adoption of evidence through how evidence is presented.

We see 3 main areas where AI can have an outsized impact on the evidence landscape:

1. The search for evidence
2. The synthesis of evidence
3. The adoption of evidence

Search

Searching for evidence appears the most well developed out of the three areas. Tools are already in place to support the automation of the search and screening process. Elicit, Scite, Perplexity and Consensus can quickly identify relevant publications given a specific research question.

However, further work may still be needed before these tools can replace manual searchers. One study finds that while ChatGPT was able to generate Boolean research terms, these weren’t appropriate for article extraction.

Synthesis

Synthesis appears the largest area of opportunity of the three areas outlined. Currently, AI tools can conduct abstract screening, but still struggle with data extraction.

AI aside, there is still more basic technology that can be adopted within synthesis. Exploiting APIs, structured databases and standardised data formats would allow searches to more readily extract relevant details from abstracts. This requires consistency in terminology and formatting that is still lacking in primary evidence production.

Work is underway to redress this. For example, the Evidence for Education Network have funded work through the Evidence for Policy and Practice Information (EPPI) Centre on the use of the OpenAlex library for updating existing reviews, and have funded the development of education classifiers to make the machine process more effective at identifying relevant studies.

The Future Evidence Foundation has been working to reduce the timelines for evidence production, particularly for Living Evidence Reviews. Part of this work draws on advancements in AI to reduce the time taken for synthesis.

Large language models (LLMs) could also help unlock a significant body of research currently neglected by researchers in the Anglosphere where non-English studies are often ignored. A review of all Campbell Collaboration systematic reviews (as at July 2016) found that of 123 reviews, only 17 included non-English language studies.

One article suggests that AI tools could be housed *"within institutions that have clear mechanisms in place to ensure robust governance, broad participation, public accountability and transparency. For example, national governments could build on current efforts, such as the U.S. What Works Clearinghouse and the UK What Works Network."*

Adoption

Some AI tools already offer to help write policy briefs (e.g. Taskade). While at present these tools appear relatively crude, they have the potential to significantly expand the adoption of evidence by policy makers.

AI could help translate meta-Living Evidence Reviews [see recommendation 4] into the toolkits, making them much more accessible for policy makers. This should be less subject to bias than drawing on general content, as the AI tool would only be using the input given to it from the meta-LER.

Section 3: The Case for Collaboration - our recommendations

“Countries are often over-reliant on their own research, using media oriented research”

“There is ample opportunity for collaboration [in education synthesis across countries]”

“EEF and IES are natural collaborators but we have not gotten over technical process hurdle - seems like terrible waste”

The case for collaboration on the production, synthesis and adoption of evidence is strong. Cultural and contextual differences between countries are important, but there are many common interests and questions: How best to screen for cancer? What is the best way to teach a child to read and write? How best to reduce recidivism?

Collaboration on evidence can accelerate diffusion of best practice and increase the effectiveness of (constrained) investment. Better applied evidence can ‘bend the spend’ towards higher-value interventions and practices. In this sense, variance in approaches to public services is an asset, enabling policy makers and practitioners to hone in on variations that are achieving better outcomes and systematically testing whether they can work in other places (see box 3).

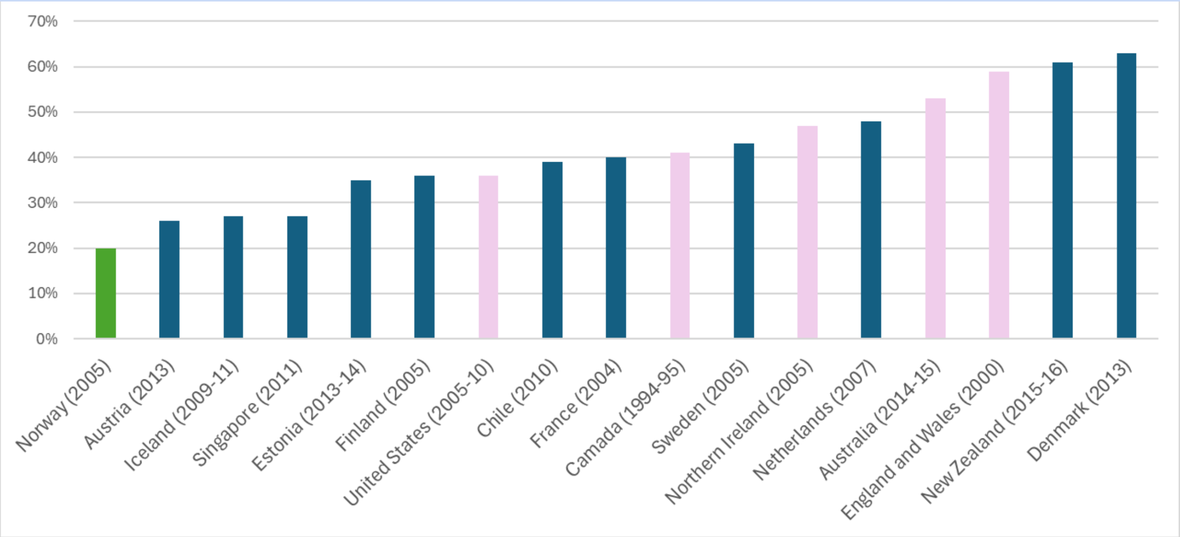
Direct savings are possible. Reducing duplication of effort, such as by sharing the cost of syntheses, can provide immediate cost savings, The savings will be greater as more countries collaborate. This approach could prevent redundancies such as the experience of one interviewee who spent years on an education research synthesis, only to see a similar review published a month later.

Collaboration can add value for evidence producers and consumers. Most evidence is currently slow to generate and hard to absorb. Collaboration between evidence consumers and producers, alongside leveraging technological developments, could accelerate evidence synthesis without sacrificing quality.

Box 3: ***Case Study: How to Reduce Recidivism?***

Billions are spent on the criminal justice system internationally, including on courts, prisons and rehabilitation services. Despite significant investment, recidivism is a significant (and shared) challenge, with recidivism rates at 2 years typically approaching 50%.

Figure 15: Recidivism rates by country (2-year reconviction rate, various years, data from the World Population Review) (Pink = partner countries for this report, green = lowest reconviction rate).



The four partner countries to this report have strong departmental, and within-country state interests, in reducing recidivism, with many linked research budgets:

- The U.S. has the Department of Justice [budget $35bn], comprising the National Institute of Justice [budget $63m], and the Federal Bureau of Prisons who commission research. U.S. states also have some of their own research capacity, including the Washington State Institute for Public Policy, famous for an early influential review of ways of reducing youth crime.
- The UK has the Ministry of Justice (but no corresponding 'What Works' Centre).
- Canada has the Correctional Service Canada, which includes its own Research Branch, with additional work and responsibilities at the provincial level.
- In Australia, equivalent work is led more at the state level, such as by the New South Wales Department of Communities and Justice, with its own list of research priorities including system quality and effectiveness.

In expenditure classification terms, all of this activity falls within 'Public Order and Safety' at COFOG Level 1. The four partner countries spent c.$537 billion on 'Public Order and Safety' in 2020 (based on IMF data). At COFOG Level 2 most of the expenditure falls within 'prisons', 'law courts', and 'R&D related to public order'.

Even across just the U.S., UK, Canada, and Australia, between 50 and 100 government institutions can be identified as having a strong interest in understanding what works to reduce reoffending, and many of these are independently commissioning research and reviews on the topic. At the same time, it is an area identified in our survey of policy makers as having a relatively low quality of both primary and secondary research.

This is an illustration of what the authors of the Global Evidence Commission call 'research waste'. Multiple funders are commissioning low to intermediate quality reviews on strongly overlapping areas. At the same time, all have a common interest in learning 'what works' from across the world. These interests include understanding how countries such as Norway appear to achieve recidivism rates of roughly half of that of Anglosphere countries (and even of adjacent Nordic nations such as Sweden) - possibly due to the distinctive highly engaged role played by its prison officers, and their much more extended training than seen in most other countries (though recording differences may also be a factor).

A more efficient and effective way forward might include: collaborating to co-fund a single high quality review of all the existing evidence [see recommendation 4]; co-funding evaluations and replications of promising interventions in other countries [see recommendation 1]; and collaborating to share 'grey literature' evaluations and results from within governments [see recommendation 2]. There would be a high value for independent funders to fund and support such activity.

The section that follows sets out our recommendations for better collaboration between governments, funders and evidence providers, alongside indicative costs and delivery options (institutional form and funding arrangements). The recommendations are organised by the primary, secondary, adoption structure used elsewhere in the report, and include costing assumptions at "ideal" and minimum viable level. All costings presented within this section are in USD, unless otherwise specified.

Primary Evidence: Increasing Production and Dissemination

There is a broadly similar investment pattern across countries: substantial funding for evidence production in health and defence, with significantly lower levels of funding for evidence production in most other areas of public sector activity.

Increasing funding for under-explored areas of research is largely a domestic fiscal challenge, and would require sustained focus over coming decades. Raising R&D in social and economic policy areas outside of healthcare and defence to comparable levels (circa 2.7%) would involve domestic expenditure of circa $85-115 billion pa across the U.S., UK, Australia and Canada.

Strategic collaboration between countries could reduce the cost, and maximise the value-added, of evaluations and primary research that are conducted.

Recommendation 1: Establish a Shared Evaluation Fund to evaluate important and novel interventions ($10-50 million)

Why?

Learning if, and how, a new intervention works is of significant value. But the costs of an evaluation - financial and political - frequently fall on the originator, presenting a hurdle for both volume and quality of evaluations. Policy makers sometimes seek out interventions and programs in other countries, but too often there is no robust evaluation to know whether the intervention actually worked.

A number of countries have begun to upgrade the quality and volume of their domestic evaluation efforts. The U.S. was an early leader, with Obama-era programs funding robust evaluations and replications of education interventions. Recent legislation (The Foundations for Evidence-Based Policymaking Act of 2018) established a statutory system for program evaluation, and the substantial Arnold Ventures program evaluating promising programs shows the role Foundations can play. In the UK and Australia, the Evaluation Task Force (ETF) and more recently the Australian Centre for Evaluation (ACE) are dialling up evaluation activity across their respective governments. These are extremely valuable efforts. At the same time, key evaluation gaps remain, such as due to domestic prioritisation and capacity challenges.

What?

Establishing a shared evaluation fund ('SEF') would address these barriers by addressing the 'public good' problems that lead to under-funding and under-conducting of evaluations. The fund should aim to increase the volume of evaluations, and increase mutual learning on what does and doesn't work.

The SEF would be available for funding evaluations of interventions in partner countries, and elsewhere, which would not otherwise be evaluated (noting domestic funding of evaluation is still best practice).

The SEF could fund responsive and proactive evaluations. In responsive mode, policy makers and practitioners in any country could apply to the fund to receive financial and technical support for a robust experimental or quasi-experimental evaluation. In proactive mode, fund managers would actively identify innovative programs or interventions that lack evaluations, and offer funding and support to get them in place. The SEF should seek to establish the return on investment from evaluations undertaken, as well as the size of the untested pool of promising interventions (i.e. promising interventions that lack evaluations), and use this as the basis for further iterations of the SEF.

In the short term, the SEF should prioritise evaluations within areas of overlapping interest from the four countries. For example, this might currently include effective ways to reduce labour market inactivity, address climate change, or topical areas of interest such as mechanisms to protect the integrity of elections. If funding an evaluation outside a partner country, the SEF should focus its efforts on countries that are sufficiently similar in characteristics to make transferability of the result more likely.

In the longer term, the SEF should seek to develop partnerships with other funders including organisations such as Arnold Ventures, the Bill and Melinda Gates Foundation, Bloomberg Philanthropies (from a city perspective), Ramsey Foundation, and, for more specialist areas, Wellcome and the Jacobs Foundation.

Costs

A reasonable initial funding target is $10 - 50 million. Domestic evaluation teams, such as ETF and ACE, have been launched with funds in this range. ETF funding has also included an Evaluation Fund, that public bodies can bid for.

Recommendation 2: Promote standardised reporting and publication protocols for policy-relevant research and trials (~$0.5 - 1 million)

Why?

Governments are increasingly active producers, as well as consumers, of evidence. Much of this currently exists as 'grey' literature (not published in academic journals). For example, the UK's ETF has so far identified around 1,000 evaluations conducted within or by the UK government to date. Other governments have informally reported similar volumes of grey research. Ensuring this grey literature is publicly available will allow governments to draw from each other's work. The ETF's plans to publish evaluations on a new public research register, and to include the open source code on its Evaluation Registry for UK government officials. This approach should be celebrated.

Research is often published in very different formats, and hosted in many places. This presents a significant barrier to identification and assembly of evidence for researchers, policy makers and practitioners. For example, key details on the costs of the intervention or of the characteristics of the target population or geography are often not included in the reporting of the result. Moving towards more standardised reporting or coding formats and publication methods would make it easier for researchers and policy makers to find and compare studies, and for secondary reviews to be conducted.

What?

Partner countries could establish standardised evaluation reporting and formats to enable better comparison of interventions globally**.** This should include:

- Summary descriptions of the population target and measures;
- Methods used, including randomisation method (if applicable);
- Whether a research protocol was published;
- Sample sizes and characteristics;
- Cost of intervention per unit; and
- Key results.

Partner countries could also publish registers of protocols and completed evaluations, which could be accessible via a joint evaluation portal. Standardised reporting would support integration in a common evaluation registry (or portal). This would enable studies to be published - either publicly or if necessary with security classification - to allow cross-national learning and replication. In the short term, we consider there would be value in Australia (ACE) and the UK (ETF) leading the way to agree shared reporting formats for their respective registers of evaluations, and develop a platform through which both registries can be accessed, and evaluations compared.

Costs

Costs would be primarily administrative within countries, with some monitoring and maintenance activity of agreed changes required. By pursuing an 'evolutionary' convergence of reporting formats, rather than a 'big-bang' change, costs could be reduced. It is likely that if several partner countries agreed on a common format, other countries - including non-English speaking countries - would align over time. We have included a small sum ($0.5 - 1m) to cover liaison. A more ambitious approach would be to create a common register, which would cost significantly more. Our judgement is that common APIs, formatting, and searchable nationally held registers would work - and could be revisited in the future if appetite grows for a common register.

Box 4: ***Collaboration at the local level***

Thus far, we have considered evidence and collaboration at a national level, with the national governments of the four countries pooling resources and agreeing on activity.

There is great promise in increased collaboration at a regional and city level. For example, at the city-level, city leaders are often constrained in their resources and only have a small team at their disposal. One UK Local Government leader described how he typically received around 200 prompts or documents a week directing him to take up one idea or another, with no guidance or capability to tell which ones were based on robust evidence. Similarly, work by Bloomberg Philanthropy with Mayors has found that successful innovations are often very slow to diffuse between U.S. cities, let alone between city leaders in different countries.

These local and regional leaders should be enabled to:

- Identify interventions that have been shown to work elsewhere;
- what the 'active ingredients' of effective interventions are thought to be; and
- which interventions worked in places that are most comparable to their own.

The Shared Evaluation Fund might support city level evaluations. This is also something individual countries could move towards: within-country pooled evaluation funds twinned with statements of areas of interest at city, regional and other local level.

Secondary Evidence: Advancing Quality and Relevance

Policy makers and researchers do not have good access to high quality evidence overviews in many areas. **We consider the strongest case for global collaboration exists at this secondary review level.**

By its very nature, high quality secondary reviews draw from evidence of studies from across the world. They are also much more efficient as a 'go to' for a busy policy maker or practitioner than ploughing through primary studies.

At the same time, the volume of current syntheses, and the evidence from our interviews, suggest there is **considerable duplication and research waste occurring at the secondary level.**

Recommendation 3: Conduct evidence gap maps across priority policy areas (~$10-30 million)

Why?

Evidence Gap Maps (EGMs) can be used to expose areas of priority research and direct further funding and research activity, such as through the UK Areas of Research Interest and U.S. Learning Agendas**.** EGMs also provide a foundation and direction for systematic evidence reviews. They highlight areas where a reasonable volume of evidence exists that could be usefully summarised for policy makers and practitioners.

There is a stronger tradition of using EGMs in international development, but a dearth of EGMs for shared domestic policy interests. Collective investment in EGMs for policy, along with unified cross-national calls for evidence could mobilise collaborative research efforts from academics and Foundations to address critical gaps.

What?

An EGM provides an overarching view of the state of the evidence within a field / topic. They are used to show where there is an abundance - and conversely, absence - of evidence within a policy area. Typically, an EGM will include impact evaluations and systematic reviews of intervention effectiveness. They may also include qualitative studies.

Our medium term target should be to deliver comprehensive evidence gap maps across all key policy areas. However, in as far as prioritisation is required, this work should target areas where policy makers believe there is a reasonable quantity of primary evidence, but limited or uncertain access to that evidence. From our survey and interviews, this would include areas such as:

- Climate change and mitigation
- Disability, sickness and old age
- Environmental policy (waste, recycling, pollution and biodiversity)
- Family and children
- Higher education and vocational training
- Housing and community development

Defence and healthcare are considered lower priorities, given there is already high investment and a more mature research landscape in these areas.

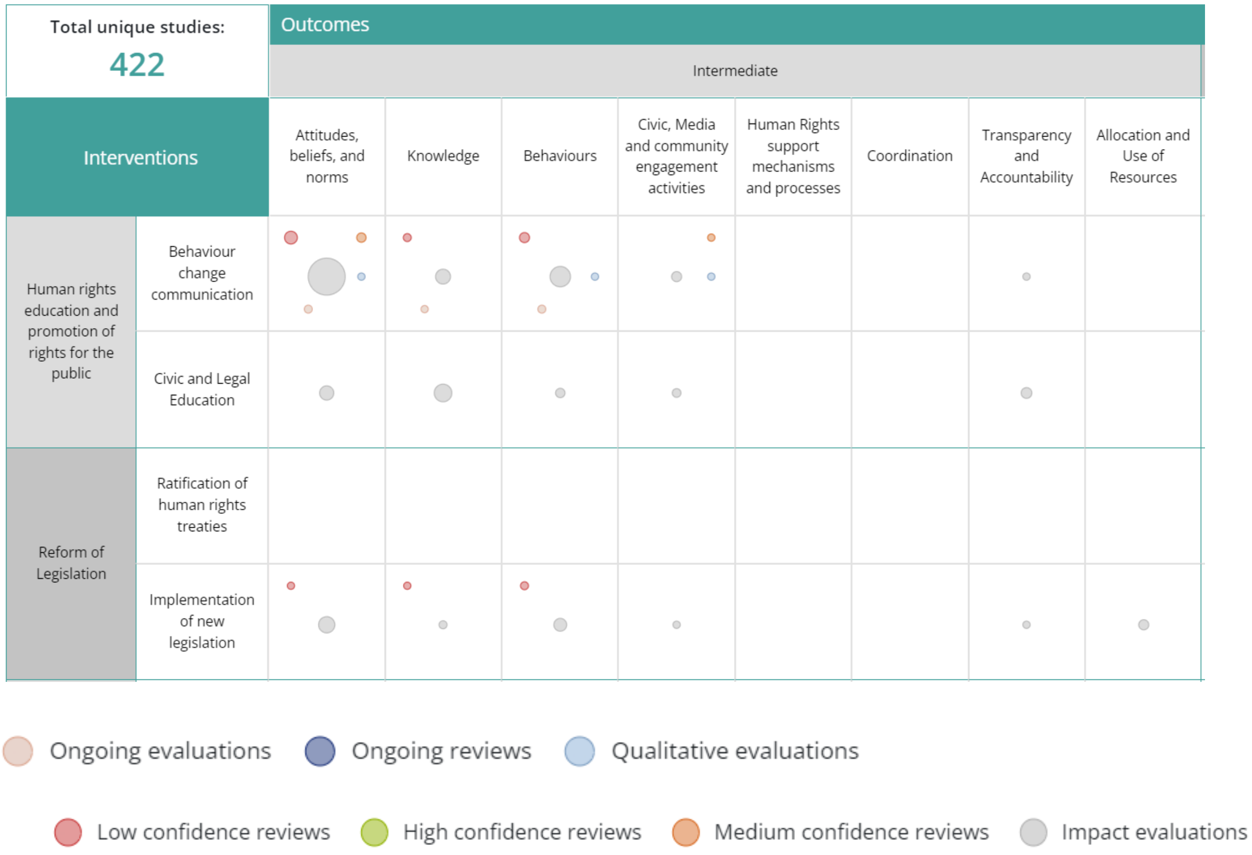
EGMs provide a more systematic springboard for setting out 'Areas of Research Interest' or Learning Agendas, by identifying gaps in the research landscape in areas of policy concern. These calls for evidence are separable and significant exercises in their own right. They could include assembling urgent calls for evidence (see 'kickstarter' proposal below); identifying 'low hanging fruit' of areas with good volumes of primary evidence but lacking good summaries; to more formal, comprehensive processes. It should also be possible to scrape together domestic calls for evidence across countries linked to forthcoming legislation or possible government action, and combine these with nascent independent attempts to assemble policy demand for evidence.

Costs

Costs to produce EGMs are relatively modest. Evidence gap maps typically cost $250,000 - $750,000. Producing EGMs for ~40 COFOG Level 2 areas would therefore have a total cost of $10-30 million. However, upfront investment in automation could reduce these costs significantly. This is discussed in further detail under recommendation 6 below.

Costs for systematic 'calls for evidence' or 'Areas of Research Interest' are less clear. We consider there could be quick ways of scraping and assembling existing calls for evidence across sources at a relatively modest cost (~0.5 - 1 million) noting that only a few countries have set these out in a formal way.

Figure 17: Example evidence gap map, by 3iE (the size of the circle denotes the number of studies within that classification.)



Recommendation 4: Conduct evidence syntheses, in the form of meta Living Evidence Reviews (meta-LERs), for high priority policy areas (~$50-100 million)

Why?

Policy makers and practitioners benefit from easily accessible summaries of what does, and doesn't work, for who, when, where and why. Such summaries should be foundational to any policy work, and for shifting public expenditure towards more effective practice and higher public sector productivity.

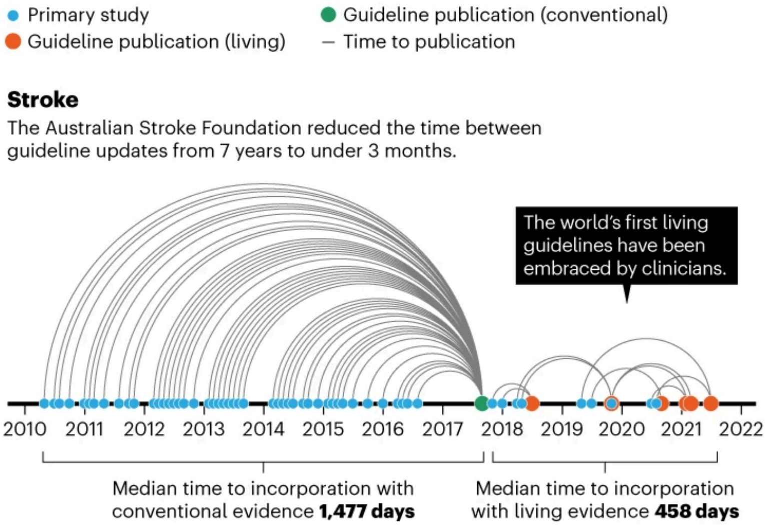
Despite the value in such summaries, countries currently have no mechanism to pool the commissioning or cost of such reviews. This results in considerable 'research waste' associated with conducting large volumes of low quality, or duplicative reviews (see Figure 8). At the same time, there is relatively little funding for, or academic prestige in, the conducting and maintenance of high-quality meta-living evidence reviews (meta-LERs).

Collaboration between countries can increase access to higher quality, more comprehensive summaries of the evidence - and at a lower cost than if commissioned and procured separately.

What?

Both evidence producers and evidence users pointed us towards LERs as a tool to make reviews more useful to policy makers, rather than 'traditional' or static reviews. A Living Evidence Review (also known as a Living Systematic Review or Living Evidence Synthesis) is a systematic review that is continually updated, allowing for the incorporation of new evidence on an ongoing basis. Whilst LERs are relatively nascent, recent interest in them is growing, with grants being awarded within the last couple of years to expand their number (e.g. by Wellcome). Once a LER is established it can facilitate rapid advice to policy makers, as high quality evidence is already collated, summarised, and up to date.

Figure 18: Time to publication using living evidence (source: Elliott et.al, 2021)



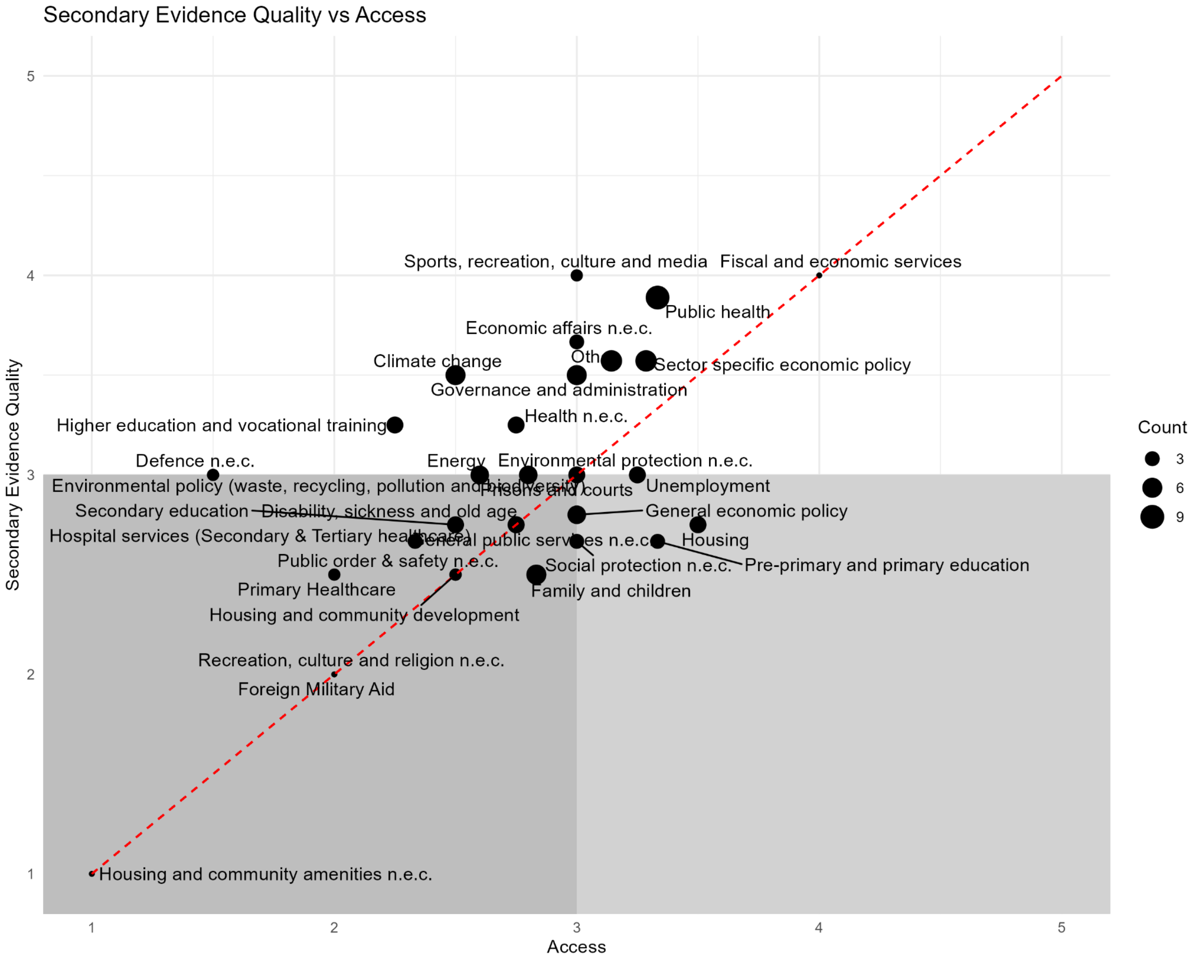
This report coins a new term: 'meta'-LERs, recognising feedback we received from policy makers on evidence being presented too narrowly. Meta-LERs would work at a higher policy or outcome level, than at the 'traditional' intervention-level used in many current systematic reviews. For example, a meta-LER might cover effective ways to reduce types of crime, as opposed to a 'conventional' systematic review that might look at the efficacy of one particular approach to reducing crime (such as a review of 'Scared Straight' style programs).

Partner countries could collaborate on, and co-fund a set of meta-LERs covering a range of COFOG Level 2 areas noting that in some areas these may already be covered by existing institutions (see Appendix 4). The meta-LERs should have a common reporting standard, agreed between the four countries. For example, reviews should include details on the interventions, target populations, implementation and effectiveness comments, and cost benefit analyses where possible (see also recommendation 2).

The COFOG Level 2 categories are broad, and the scale and focus of these reviews was much discussed by the working group linked to this review. Some felt that COFOG Level 2 areas are so huge that it may be unrealistic to commission reviews at this level, even as meta-LERs. Others felt that the broad questions were exactly what they wanted an answer to. Some respected examples of the latter do exist, such as the Education Endowment Foundation (EEF) toolkit (see Figure 20) or the Results for America Economic Mobility Catalogue, both of which allow users to drill down to more specific questions.

With this in mind, we recommend 'families' of LERs should be conducted under the banner of COFOG level 2 areas. Evidence gap maps, as recommended above, can help guide and structure the LERs. Early priority should be given to topics within the COFOG Level 2 category with strong primary evidence but that have not yet been covered by strong secondary reviews. To provide an indication of areas that might be priority targets for LERs, we present a further cut of the data presented in Section 1. This should be taken as indicative only, given the relatively small size of our survey of policy makers. We would anticipate that prioritisation would be tested more systematically through a procurement process and negotiations between funders. A summary plot of policy makers ratings of secondary evidence quality and access is shown below.

Figure 19: *Access to evidence and secondary evidence quality, as reported by the survey of policy makers, per COFOG Level 2 (data as of 30 March 2024). [Correlation coefficient r = 0.54, adjusted R^2 = 29%, p = 0.001]*



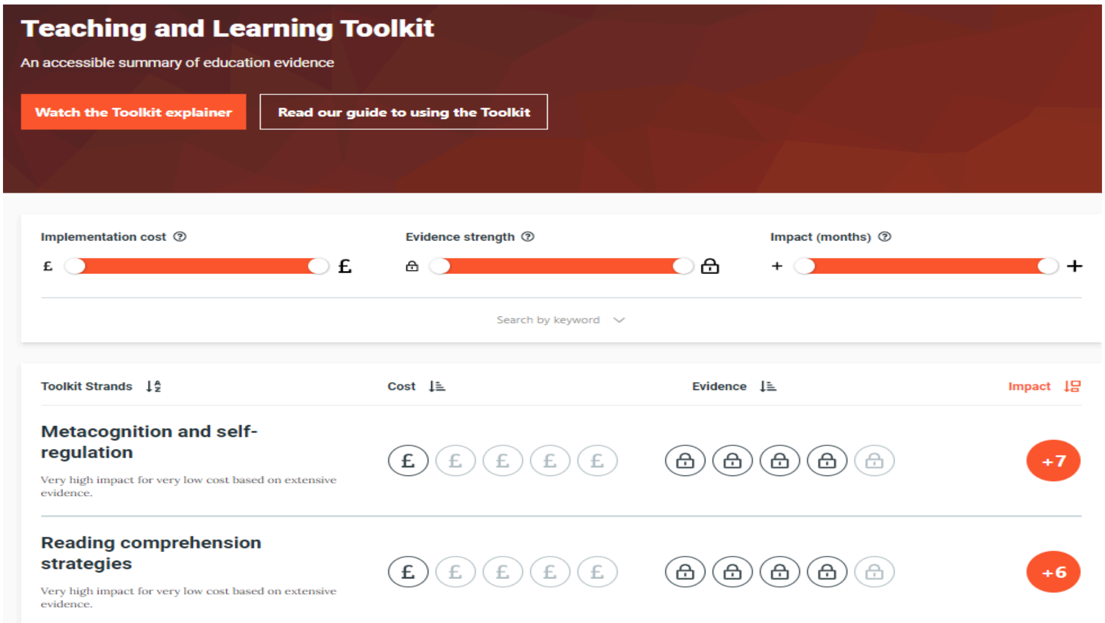
The bottom left segment captures *prima facie* priority targets for meta-LERs: policy makers reported difficulty getting access to the evidence and that the secondary reviews they were aware of are of low quality (scores less than 3 for each, as shaded in dark grey). These included:

- Housing and community development
- Foreign Military Aid
- Recreation, culture and religion
- Public order and safety
- Family and children
- Disability, sickness and old age
- Social protection
- Secondary education
- General economic policy

Each of these areas would be broken down into subcategories. For example, 'disability, sickness and old age' might be disassembled into areas such as: reducing work inactivity and illness in the working age population; improving outcomes for those with long-standing disabilities; and promoting healthy ageing in seniors.

Alongside the evidence covering more comprehensive topics, a key focus should be on making the evidence actionable. Secondary evidence needs to be presented in a manner that makes it easy for policy makers and practitioners to understand and interpret. The EEF, for example, provides a toolkit for practitioners that makes it easy for the user to identify the strength of the evidence, the cost involved and the overall impact. The EEF 'toolkit' is, in all but name, a meta-LER given its range and regular updating. Users can also drill down to see the detailed studies and impact sizes behind each of the summary conclusions. It marries comprehensive coverage (a COFOG Level 2 category is covered) with a high degree of useability. The usability is reflected in the use of the tool by schools: around 3 in 4 UK schools now report using the toolkit to help guide their decision making.

Figure 20: Education Endowment Foundation Teaching and Learning Toolkit



Teaching and Learning Toolkit

An accessible summary of education evidence

Watch the Toolkit explainer

Read our guide to using the Toolkit

Implementation cost ⓘ £ £

Evidence strength ⓘ 🔒 🔒

Impact (months) ⓘ + +

Search by keyword

Toolkit Strands	Cost	Evidence	Impact
Metacognition and self-regulation Very high impact for very low cost based on extensive evidence.	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+7
Reading comprehension strategies Very high impact for very low cost based on extensive evidence.	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+6

National or regional bodies should be able to use the meta-LERs to 'power' their own toolkits or 'shop-windows' to the evidence, re-cutting or re-weighting them as necessary to match the administrative context. These tool-kits would then be updated in real-time, whenever the meta-LER is updated. These 'toolkits' should be coupled with complementary actions on adoption (below).

Costs

Given that each Level 2 category covers a broad topic area, we expect that each would typically need around 10 individual LERs to make a meta-LER (see discussion above). From discussions with providers, we estimate costs would be around $250,000-$500,000 per LER, and around $2.5-$5million per meta-LER.

After deducting areas that are already well covered by existing reviews (such as medical health), and anticipating that there will be policy areas will be lacking in sufficient primary evidence for meta-LERs, we have estimated that around 20 meta-LERs would be a reasonable target for coverage (or 200 individual LERs). This would imply a total cost of around $50-100 million. Costs could be dialled up or down, depending on ambition (see minimum viable product (MVP) discussion below).

Once established, the ongoing cost of updating the LER database would be ~$0.5-2 million per meta-LER. Therefore, the annual cost for updating 20 reviews would be ~$10-40 million. We include a substantial cost range because (a) it depends heavily on the rate at which new research is produced, and (b) the extent to which automation may be possible.

Box 5: ***Case study: Secondary evidence for 'Prisons and courts'***

Building on the example of recidivism earlier, 'prisons and courts' is a COFOG Level 2 area (see Appendix 2) within the Level 1 category 'Public order and safety'.

Prisons and courts comprise many different topics, including the important issue of reducing 'recidivism' as set out in Box 3. However, it will also include other topics, such as violence within prisons, treatment of victims and witnesses (including how to reduce 'cracked trials'), and public perceptions that 'justice has been done'.

Under our recommendations, an evidence gap map would be conducted across the public order and safety domain (COFOG 1). This would provide a breakdown of the sub-group topics, including within prisons and courts. This gap-map would determine the evidence availability within each of these topics.

Subsequently, those topics which are identified through the mapping to have sufficient primary evidence (e.g. recidivism) would have a LER conducted. Those areas without sufficient evidence (e.g. sentencing) would be directed towards further primary research, and could be an area picked up using the Shared Evaluation Fund (SEF) as outlined in recommendation 1.

Finally, ongoing complementary work would be needed to ensure that the findings were translated and utilised by policy makers and practitioners, both domestically and where relevant between countries (see recommendation 5).

Boosting Evidence Adoption

For policy makers and practitioners, primary and secondary research is a means to an end: to improve services, policies and outcomes. This hinges on effective translation, adoption, and implementation.

Research for this report has identified good quality syntheses that appear rarely used or accessed by policy makers. Similarly, there are striking examples of robust results that have been slow to diffuse, or 'fade out' of use even in sites where they were first adopted. For example, it is often reported that it takes around 17 years for research evidence to reach clinical practice.

Adoption and implementation need to be a focus in their own right. A number of experts and policy makers told us there is a need for a 'human element' in the evidence chain. This means a person who can be called up to provide a rapid summary of recommendations, tailored, and adapted to the needs of the policy maker. In public service professions, spread and adoption of best practice often hinges on professional networks and on respected individuals - or 'seeds' - who act to spread new practice among peers.

Much of the work to enable adoption needs to be done domestically, often at a regional or local level. The 'last mile' of translation or adoption requires an understanding of the local or professional context.

There are some crucial ways in which international collaboration can propel domestic action further and faster. The recommendations below seek to accelerate currently sluggish adoption processes to improve the timeliness of evidence-based policy-making.

Recommendation 5: Strengthen international professional networks and institutions focused on accelerating the transfer of knowledge between countries (~$5-20 million)

Why?

Focus on local practices and institutions overlooks the potential for learning from other countries and systems. Broader perspectives could enable policy makers and professionals to 'break out' of the assumptions and constraints of the systems within which they operate.

Our interviews, and wider work, underscored the importance of the 'human touch' to enable evidence transfer. This could involve hearing directly from someone who is personally familiar with the intervention and context, twinned with curation to guide attention towards proven and effective interventions.

There are good examples of such networks. Medicine has evolved into a deeply evidence-based profession, with most specialisms boasting their own learned societies, journals, and mechanisms to accelerate the diffusion and implementation of evidence-based practices, both domestically and internationally. Another example is the Five Eyes network, which operates between the four partner countries to this review (and New Zealand), focusing on military and intelligence. The network includes an element of best practice and evidence sharing, albeit within a classified context.

Most public service professions, despite millions of people within them, have yet to develop the depth, breadth and empiricism of 'evidence-spreading' and adopting networks that have been established in medicine.

What?

There is a strong case for accelerating the learning between countries, including from the LERs proposed in recommendation 4. These networks can also be used to agree strategic priority areas and avoid overlapping work by institutions.

Targeted funding and support to build evidence networks could enable public service professionals (such as headteachers, senior police officers, and senior social

workers) to learn from best practice from other countries in a more structured way. This initiative seeks to build robust networks across all COFOG Level 2 areas. The purpose of these networks, and associated events and institutional capacity, would be to expose policy makers and public service leaders to alternative but effective approaches. This would bring to life the evidence assembled by Living Evidence Reviews, evaluations and transferable innovations elsewhere (recommendations 1, 2 and 4) . The aim is to increase the absorptive capacity of public service leaders - time, outward-looking curiosity and capacity to distinguish evidence-based claims.

These networks and activities would be similar to those already existing in fields like medicine (e.g. the International Council of Nurses). They could also build on and support nascent networks such as the Evidence for Education Network or the Society for Evidence Based Policing and potentially existing institutions such as the OECD's Observatory of Public Service Innovation. The funding would target the lateral diffusion and adoption of evidence-based practices across countries, including use of the LERs.

As a minimum viable pathway, partner countries could fund a demonstration project or network, and assess its value-added, for example, to strengthen the existing Evidence for Education Network (EEN). Outside of health, education is arguably the area to have seen the most extensive build-up of robust applied research over the last decade. At the same time, teaching careers last decades, so the majority of teachers today will have been trained before such evidence was generated. The EEN secretariat is housed within the UK's EEF and has partnerships with organisations in Australia, New Zealand, Spain, Belgium, Netherlands, Chile, and Jordan. As a first port of call, expanding the network to the U.S. and Canada would greatly increase the reach of the network and mutual learning between comparable systems.

Policing is considered another good early target. There are nascent but irregular visits and events that connect senior police (and mayors) between the Anglosphere nations, alongside promising networks within countries of officers seeking to make policing more evidence-based. These are currently poorly funded, but could provide high value learning and exchange - especially since many crime types have an international dimension (such as fraud, drugs, and trafficking).

Costs

We have estimated that establishing effective networks in a range of priority COFOG areas would be in the range of $5 - 20 million per annum. This is based on costs from the World Medical Association, which costs approximately EUR 2 million per annum to run, including events, advocacy and developing codes of practice.

An intermediate option would be to fund a more modest drive focused on a selection of priority public service professions, such as policing and education (circa $1-5m).

Recommendation 6: Research on applied research, translation and adoption ($1-5 million)

Why?

There are significant gaps in 'research on research', sometimes known as 'matascience'. These range from gaps in basic statistics, such as levels of research expenditure broken down beyond COFOG Level 1 or by city or regional level, through to estimates of the returns on investment (ROI) delivered by different areas of research.

The variance between countries in research, translation and adoption is itself an important source of learning 'what works' in applied research and adoption. Different countries have pursued different strategies in their R&D and translation activities, including examples of approaches moving from one country to another. The U.S.' Washington State Institute for Public Policy, founded in 1983, was a powerful influence on the UK's What Works Centres model - though it took nearly 30 years for that transfer to occur. Similarly, the U.S. Defense Advanced Research Projects Agency (DARPA) model has been (at least partially) copied elsewhere, such as the UK's Advanced Research and Invention Agency (ARIA), and DARPA itself borrowed heavily from the Israeli Directorate of Defense Research & Development model.

There are also significant innovations and variations in research funding practices, including those aimed at increasing the translation, evaluation and adoption of research. Examples include Canada's decision to increase their expenditure on overt translation activity of health research to increase adoption; the U.S. government's programs to encourage the secondary replication and adoption of evidence-based approaches; and the Australian Government's recently created the Australian Centre for Evaluation (ACE).

What?

Gaining empirical insights on the relative merits and impacts of different approaches should be a priority. This should include (a) estimating the marginal returns to changing levels of research expenditure across different areas of applied research (see below for returns on investment from the EEF); (b) establishing variations in return within areas (e.g. relative return from different types of translational or implementation spending; and (c) finding the best ways of harnessing AI (see box 2 above).

Box 6: ***Return on investment from the Education Endowment Foundation (EEF)***

Preliminary analysis by Boston Consulting Group (BCG) for EEF suggested that:

1. Every pound spent establishing an evidence base for a top-10 intervention generates a 20–fold return when translated into lifetime earnings.
2. Every pound spent delivering a top-10 intervention by a school generates a 74-fold return when translated into lifetime earnings.

Identifying more efficient ways to support the diffusion and successful implementation of evidence-based interventions in policy is an important research question in its own right. For example, what is the right level of 'human touch' that needs to be combined with the work of clearing houses to enable effective translation and adoption? What is the best way of making larger reviews of evidence digestible by policy makers and practitioners? What are the skills that policy makers and practitioners themselves need to be able to generate and apply evidence? What are the limitations on transferability of evidence from one country to another? To what extent should evidence guides be designed for wider public audiences? Is mobilising the public a neglected channel to drive evidence adoption?

There are opportunities for significant methodological innovations around the generation, translation and adoption of evidence. These include unlocking new data assets to map and understand variations in outcomes and practices that could lead to major breakthroughs in identifying better practice and policy. Similarly, it is widely thought that AI is on the cusp of being able to reduce the cost of research synthesis, and is already being employed by some governments to help draft summary submissions within governments. Experts in the area report we are still some way off full automation (see box 2).

Costs

A basic review on the efficacy of variations in approaches to applied research, translation and adoption could be undertaken with ~$1 million. A key issue that it may need to address is the quality of underlying data, including on ROIs, as well as the viability of using 'natural' variations in approach. A more extensive program, including working with funders and clearing houses to test variations in approach, could be achievable at ~$5 million.

It has become a truism that any research paper or report ends with a call for further research. Nonetheless, this type of metascience is an issue that research councils, and government Departments that are responsible for research budgets, should have a particular interest in pursuing.

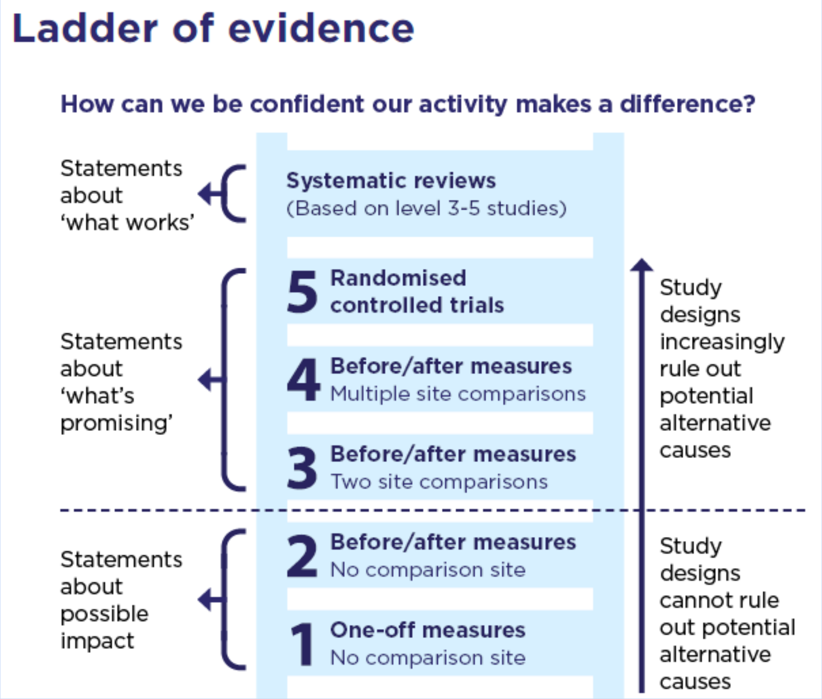
Box 7: ***Methods and controversies***

"Still part of this problem [is that] they are old school thinking around purist nature on high quality data - fall into trap of 'we have to ask the question and we have to ask in this way'"

"Hierarchy of evidence - like experimental design as a block rather than just RCTs"

A few interviewees noted controversies and dilemmas around how evidence reviews combine and weigh the quality of evidence. Many secondary reviews conducted by What Works Centres and Clearing Houses have adopted the Maryland Scientific Methods Scale when assessing the robustness of the evidence from a primary study to include within a review (e.g. the What Works Centre for Local Economic Growth, the Center for Evidence-Based Crime Policy, and The College of Policing). The Maryland scale provides a five-point scale against which evidence is reviewed. Below is a depiction of the scale in practice, in the context of policing:

Figure 21: College of Policing's use of the Maryland Scale.


Ladder of evidence
How can we be confident our activity makes a difference?
Statements about 'what works'
Systematic reviews (Based on level 3-5 studies)
5 Randomised controlled trials
Statements about 'what's promising'
4 Before/after measures Multiple site comparisons
3 Before/after measures Two site comparisons
Study designs increasingly rule out potential alternative causes
2 Before/after measures No comparison site
Statements about possible impact
1 One-off measures No comparison site
Study designs cannot rule out potential alternative causes

A couple of interviewees felt that there was a risk of over-reliance on RCTs in evaluations and reviews, and that qualitative and other sources of evidence should not be overlooked. This could include experiential or culturally specific aspects of interventions, including whether an intervention is experienced as 'done to' a people or group, as opposed to one that is 'co-owned' or chosen.

Qualitative research tends to be harder, and more controversial, to rate within a conventional methods 'hierarchy'. Some frameworks exist for assessing the quality of qualitative evidence:

- The Cabinet Office framework, developed by the National Centre for Social Research, which has 18 questions to assess quality of qualitative evaluations.
- The Critical Appraisal Skills Programme (CASP), which has developed a number of checklists to perform critical appraisal across a number of different study types.
- The Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) which is commonly used in systematic reviews to assess the quality of evidence and the strength of recommendations.
- The Cochrane Risk of Bias Tool, which is focused on assessing the risk of biases in RCTs.
- The JBI (Joanna Briggs Institute) Critical Appraisal Tools, used to assess the quality of a wide range of study types, including qualitative research, case reports,

and economic evaluations.

- The Newcastle-Ottawa Scale (NOS), which is focused on assessing the quality of non-randomized studies, such as cohort and case-control studies.

Despite the wide range of frameworks available, including and assessing all types of evidence may be too cumbersome to embed within LERs. Considering all of the above, we propose:

1. Mixed-methods studies should be included within LERs, with the quantitative element assessed based on the Maryland scale or another appropriate scale
2. Where possible, linked qualitative research should be used to supplement the findings within the LER, including to unpack the underlying mechanisms which are driving the success of, or variation within, interventions
3. The exact calibration and combination of methods, context and conclusions should be considered an important research question in its own right.

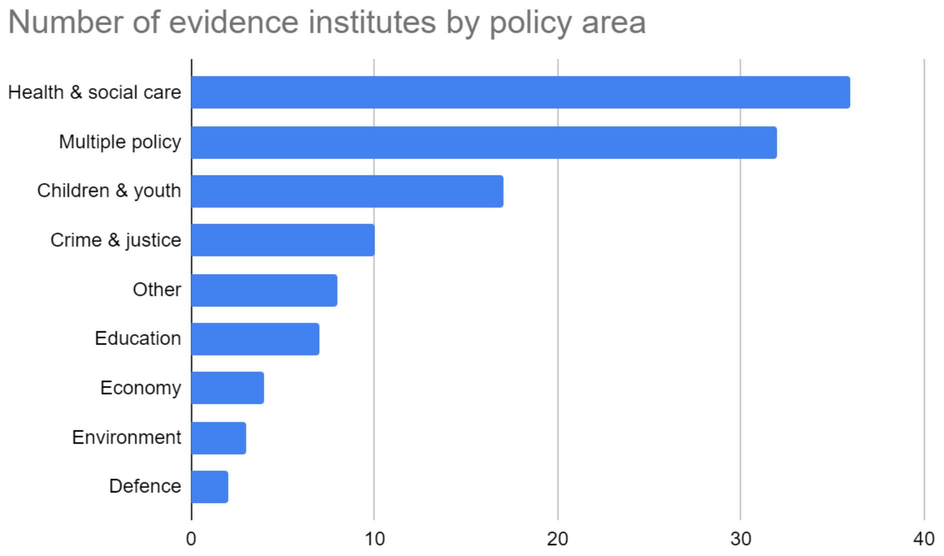
Section 4: Options for Delivery - Institutions and Program Cost

A range of implementation options exist, ranging from comprehensive to minimum viable product (MVP). The section below identifies and explains the suite of options available, including institutional roles and responsibilities, and implications for impact and timing.

Institutions

We undertook a mapping exercise to identify institutions that form part of the evidence ecosystem across the four countries to policy areas (see Figure 22). We compiled this list using the Paul Ramsay Foundation's Evidence Institute report, alongside our own research. This list is indicative, rather than exhaustive, as there are likely other institutions not captured, including some university-based units.

Figure 22: Number of evidence institutes by policy area in the UK, U.S., Australia and Canada (data compiled by BIT)



It is clear that there are already many institutes that could and should be leveraged wherever possible, rather than creating new institutions. Most institutions are domestically focused, though some seek to draw on evidence from other countries. For example, the proposed evidence gap maps and meta-LERs could be housed and organised within these institutions, through an initial competition.

There are a number of international institutions operating in the secondary evidence space. These include providers (or coordinators) of evidence reviews, such as the Campbell

Collaboration (for social sciences), the Cochrane Collaboration (for health), and the new Alliance for Living Evidence (ALIVE) collaboration.

There are also large and well developed institutions that are at least partly in the business of collating and sharing evidence and best practice. These include the OECD (mainly for economic and education policy); the World Bank (mainly developing world); and the UN (noting a recent surge of interest around building more evidence around the 'how' of SDGs).

The Global Commission on Evidence also merits mention. A Canadian-led initiative in the wake of covid, it has brought together a broad international coalition of experts to identify the methodological capabilities that contemporary governments and public services should draw on. Though primarily academy-led, it has had input and interest from the policy community through its commissioners, and has provided an important springboard for the current review.

Against this institutional landscape, we see three broad pathways, or approaches, through which the recommendations might be delivered. Our recommendations focus on plugging gaps in the global evidence ecosystem, specifically around underinvestment in evidence public goods that span countries. Existing institutions should be utilised wherever possible, but there may be a role for novel governance layers or institutions to plug gaps.

1. Loose collaboration

Partner countries and funders could agree between themselves which of the recommendations they wish to contribute to, and pursue these as a loose collaboration. For example, specific funders might wish to support different evidence gap maps and meta-LERs (recommendations 3 and 4), forming a patchwork that would eventually cover most of the COFOG areas (see Appendix 4).

Similarly, specific funders might be prepared to support particular professional networks, and within domain evaluations (recommendations 5 and 1).

This is likely the most pragmatic route forward. It will be time consuming, and risks a diffusion of responsibility with no single organisation taking ownership. We should expect it to leave some gaps. It would probably not cover the grey literature or cross-domain standardisation of reporting. It would also probably miss the opportunity to drive evaluations and reviews to cover multiple outcomes at once, such as interventions that might simultaneously improve educational attainment, reduce offending, reduce welfare costs and improve economic outcomes - and yet be of marginal cost-effectiveness if only one outcome was considered.

2. Task an existing institution

If there is sufficient interest between partner countries and funders, a common pool of funding could be created and commissioned out. We note that at least some national funding agencies have restrictions on the extent to which they can fund outside of their country, but if several partners were involved, this could be resolved. The organisation commissioning the research reviews should ideally not itself be a provider, to avoid conflicts of interest.

There are existing institutions that may be interested and capable of overseeing and administering such a commissioning fund, such as the OECD or UN. The advantage would be that a new institution would not be needed. For example, the existing institution could run a competition to deliver the Evidence Gap Maps and LERs, which might then be won and supplied by providers ranging from Campbell to existing specialist clearing houses.

Funders may need convincing that the commissioning organisation would administer the fund to deliver the priority areas of the funders. At the same time, some existing institutions would not wish to operate the fund in a way that was focused on a particular group of countries, and would want to expand it to cover a much broader range of countries (e.g. the UN).

This model could be combined with a kickstarter approach (see box 8).

*Box 8: **The 'kickstarter' model***

A 'kickstarter' model could be used to construct a common platform to enable different countries, regions and funders to coordinate joint interests, without committing to a broader pooling of resources.

This would revolve around a pooled set of 'areas of research interest', collating common areas of interest such as LERs, to facilitate joint funding. Countries, state governments, cities, or departments would submit their questions and research priorities to the platform. If other countries submit similar questions or priorities, and an agreed-upon threshold number of requests is made, this would trigger a funding call from interested independent funders and from those who expressed interest. This would lower the cost for each of the funders, given the common pooling and commissioned mechanism. The funding could be directed towards further primary research, evidence gap maps or meta-LERs, depending on the specific question.

The 'kickstarter' model would also provide a natural flexibility, enabling other countries and funders to join, expanding the collaboration into other areas, while reducing the risks of 'free-riding'.

3. Create a new institution

Old policy hands generally seek to avoid creating new institutions. However, there are times when a bespoke and unconflicted structure is needed to ensure that the funder goals are delivered on. One reason for pursuing this review with a small group of similar countries, rather than via the G7 or G20, was the commonality of interest, including similar institutional structures, pre-existing connections, and broadly shared methodological approaches. Collaboration is easier to achieve in a smaller, like-minded group.

A relevant parallel may be the construction of the Global Innovation Fund (GIF) in 2013. The fund was initiated as a partnership between the U.S. and UK, to support innovative, low cost interventions in low income countries to improve the lives of those living on less than $5 a day. By the time of GIF's launch in 2014 with a $200m budget, the founding partners had expanded to include Australia, Sweden and the Omidyar Network. GIF has since crowded-in further funding and partners, has administered close to $1bn, and appears to have achieved a significant ROI.

A stand-alone Global Evidence Fund (GEF) could enable comprehensive coverage of the recommendations and COFOG areas. Specialist interests from particular funders could still be crowded in, not least through matched funding and signalling of interest from governments. Some funders are swayed by clear statements of interest from governments, and the prospect that the work could 'bend the spend' of wider government expenditure towards higher impact. Having a purpose built institution with clear governance would also ensure that the interests of government partners are not neglected, such as creating an interoperable registry of government evaluations and protocols.

This option would come with the costs of establishing and operating a new institution. As such this might depend on the scale and ambition of the initial response to this report (see below).

Costs options

This section pulls together summary costs of a full, intermediate, and MVP version of a 'Global Evidence Fund'.

Full / High intensity model

We estimate that to deliver the full model as outlined above would cost in the range $100m to $300m over three years. This range largely reflects the ambition of coverage, such as the scale of the Evaluation Fund (recommendation 1), but also reflects some genuine uncertainty over the cost of delivery of these evidence goods (notably recommendations 3 and 4). If the full model was pursued, we strongly recommend setting up a dedicated institution to administer the commissioning of this overall Global Evidence Fund.

Recommendation	Cost ($m)	Comment
1. Shared Evaluation Fund	10 - 50+	
2. Standardised reporting protocols	0.5 - 1	Costs mainly born domestically.
3. Evidence gap maps	7 - 20	Factors in $2m prior investment in AI automation
4. Meta - Living Evidence Reviews	50 - 100	Further 10 - 40 pa for updates.
5. Policy and professional networks	5 - 20	Further 5-20 pa.

6. Research on research	1 - 5	
Total cost in Year 1	**73.5 - 196**	
Total cost in Year 1- 3	**103.5 - 316**	$20-80m for Meta-LERs updates and $10-40m for policy and professional networks in Years 2 & 3

These are substantial costs, even in the context of government budgets, but they should pay off due to savings through reducing research waste, such as repeated low quality systematic reviews. More importantly, the central purpose of this work is to improve the effectiveness of circa 25% of GDP (i.e. public expenditure excluding health and defence). This is around $8 trillion per annum for the four countries. In other words, a $200 million expenditure would have to achieve 0.0025% improvement in efficiency to pay for itself (or a 1/40000th improvement). In practice, this could be achieved by shutting down even a handful of programs identified as ineffective.

Medium intensity model

Various intermediate options of a global evidence program exist. The version outlined below would cost around $40m - $100m over three years, or around ½ to ⅓ of the 'full model'. This illustrative version would seek to deliver 6-12 meta-LERs, or enough to cover 2-4 COFOG level 1 areas; drops funding for common reporting standards (or leaves this to individual governments); drops 'research on research' (recommendations 2 and 6); and trims elsewhere.

At this level, there would be merit in running a competitive process to host the Global Evidence Fund, or to identify a lead funding partner to administer the fund on behalf of the initial partners.

Recommendation	Cost ($m)	Comment
1. Shared Evaluation Fund	5 - 10	Pilot version only
2. Common reporting	-	Domestic led
3. Evidence gap maps	4 - 10	Factors in $2m prior investment in AI automation
4. Living Evidence Reviews	20 - 40	Further 4 - 12 pa for updates
5. Policy and professional networks	1 - 5	Further 1-5 est pa.

6. Research on research	-	
Total cost in Year 1	**30 - 65**	
Total cost in Year 1- 3	**40 - 99**	$8-24m for Meta-LERs and $2-10m for policy and professional networks in Years 2 & 3

Low Intensity model (Minimum Viable Product)

Partners have asked what might constitute a minimum viable product (MVP). We estimate that an MVP could be delivered at around $12.5m - $25m. Our highest priority recommendation is to collaborate around the commissioning and funding of LERs and the evidence mapping to underpin them. Accordingly a MVP would:

- Identify the top three to four areas of common interest that lack LERs. As identified in recommendation 4, these might include: housing and community development; disability, sickness and old age; public order and safety; family and children; and/or perhaps secondary education. Shortlisting would be agreed by the partners, noting that some non-government funders may have specific interests.
- Fund an initial evidence gap map in these areas (total circa $1-4m) including basic exploratory testing of AI to accelerate these and future evidence mapping (trimming this element to $1m).
- Proceed to conduct meta-LERs on three of these areas, ideally using a match-funding mechanism to crowd in additional support (total circa $7.5-15m)
- Use the process to build a version of the evidence 'kickstarter' (see box 8) (circa $1.5-3m)

Recommendation	**Cost ($m)**	**Comment**
1. Shared Evaluation Fund	-	Cut
2. Common reporting	-	Domestic led
3. Evidence gap maps	2 - 4	Factors in $1m prior investment in AI automation
4. Living Evidence Reviews	7.5 - 15	Seek to update at least once at 1.5-3
5. Policy and professional networks	-	Cut

Recommendation	Cost ($m)	Comment
6. Research on research	-	Cut
Build ‘kickstarter’ platform	1.5 - 3	Platform to collate demand and build stronger fund
Total cost in Year 1	**11 - 22**	
Total cost in Year 1- 3	**12.5 - 25**	$1.5-3m for Meta-LERs in Year 2 only

The matched funding mechanism is recommended as a way for partner governments to draw in the resources and expertise of independent funders. It would further increase the leverage of relatively small sums from partner governments. For example, if just two of the four partner governments to this report (or associated non-governmental organisations) put in $5m each, and this triggered matching funds of a similar amount from partners, this would likely establish an MVP and collaboration that could be further built on. It would also enable the partners to test the ROI and utility of the results. In fact, at this level of MVP, we think there is a strong prospect that an existing single partner could proceed to establish a basic GEF - though it would arguably be better for the nature of the project if the initial pilot or MPV involved at least two countries or funders from the outset.

At the MVP level, we would not recommend creating a stand-alone institution or separate Global Evidence Fund. Rather it would likely be better for the funding partners to commission and supervise the Fund directly, seconding in a Director and team as necessary..

Summary: Institutional form and funding

We have set out three delivery options, ranging from comprehensive to minimum viable:

- The **full / high intensity version** would involve the establishment of a Global Evidence Fund with total costs in the range of $73.5m-196m in Year 1 and its own institutional form. The Fund would commission and drive the building of global evidence goods, with enough resources ‘in the tank’ to cover more than half of all domestic policy areas, helping to improve the effectiveness of trillions of public sector spending.

- The **medium intensity version** costs in the range of $30m-65m in Year 1. This would cover around 6-8 LERs at COFOG 2 level, with underpinning evidence maps, and prototype a shared evaluation fund. It could be housed within a bespoke Global Evidence Fund institution (especially at the upper end), or could be competed for housing within an existing institution.

- The **low intensity/MPV option** is estimated at $11m-22m in Year 1. It would in effect act as a pilot, testing the approach to evidence maps, conducting around 3 meta-LERs, and establishing a ‘kickstarter’ style platform to enable better collaboration between countries, regions and independent funders to fill out global shared evidence gaps.

Even the MVP would be a valuable step forward. Our discussions with funders have indicated that they would respond positively to a clear signal from several governments that plugging these evidence gaps would be welcomed and have impacts on policy.

Further detail is provided in Appendix 4, outlining our assessment of the current state of evidence by COFOG areas, the institutions that could be leveraged, and funders that could support these activities.

Conclusion

This report has identified large gaps in the evidence ecosystem, seen in almost all policy areas outside of healthcare (bio-medical) and defence. These gaps are seen in research expenditure and in the views expressed by senior policy makers. Public service areas with low R&D together account for around a quarter of GDP across partner countries, pointing to a massive opportunity to improve public service productivity by identifying more cost effective practices.

The ‘investment gap’ in primary research and evaluation within these areas runs to many $billions, but should mainly be for domestic governments and their agencies to address. This report has focused on areas where collaboration across the four partner countries - U.S., UK, Australia, and Canada - would be particularly valuable and cost-effective.

At the top of this list is the strong recommendation to collaborate around ‘secondary’ research or synthesis. Policy makers across countries are asking essentially similar questions in parallel. Collaborating to commission such secondary or summary research reviews offers three linked benefits. First, costs can be pooled and reduced for each participating country. Second, a broader range of literature and interventions could be covered (i.e. learning from what other countries are doing). Third, the research is more likely to be used - and useful - if partner countries have been involved in commissioning and crafting it.

We recommend collaboration on a number of other linked activities under the umbrella of the proposed ‘Global Evidence Fund’. These include: creating a pooled fund for evaluation activity; collaborating to create and share evidence within-government evaluations through interoperable registries of protocols and evaluations; and supporting evidence-based learning between public service professionals and policy specialists across countries.

Next steps

We are grateful to partner governments and funders for considering this report. Determining the level of collaboration and funding that may be realistic and valuable would be a useful next step. This could be done in a ‘big-bang’ agreement, or by two or three first movers advancing elements of the recommendations, such as the minimum viable product.

There are a number of international gatherings with a focus on evidence in 2024-25. These gatherings could be used to bring together funders, providers and governments.

Appendices

Appendix 1: Terms of reference

Objectives as agreed with the four partner countries:

The U.S., UK, Australian and Canadian governments are jointly participating in a sprint to identify:

- Policy areas where there is greatest demand for, but gaps in, evidence across the four countries, as demonstrated through a combination of desk research and interviews
- Shared product[s] to potentially fill the gaps, including in translation and adoption - evidence presented in easily digestible and actionable formats for policymakers or practitioners
- Candidate institutions that could produce the proposed product[s] and/or the institutional form that should be adopted to elicit collaboration between the four countries going forward
- Funders that may be able to kickstart funding in this space, including non-government, and indicative costs for addressing the key evidence and institutional gaps

The output for this project will be an outline blueprint for international collaboration on evidence generation, translation and adoption across the four countries.

Role of the steering group

To:

- Challenge and check core assumptions
- Ensure the review is addressing questions of interest across the four countries and stakeholders
- Provide support and follow-through for conclusions

Appendix 2: Methodology

Semi-structured interviews

We conducted 10 semi-structured interviews between 24 January 2024 and 11 March 2024 with evidence producers and evidence users across the UK, U.S. and Australia. No interviews were conducted in Canada.

Interviewees were asked a series of questions across three key areas:

1. Current use / production of evidence

2. Tools and methods to assist with evidence production and synthesis
3. Institutional arrangements that can support collaboration between countries

Direct quotes from interviews included within the report are anonymised.

Survey with policy makers

The survey was available from 23 February 2024. The data presented below was extracted on 30 March 2024. The survey remains open at the time of writing (April 2024) and we may seek to update any of the graphs presented in this report if further responses are received that materially alter the results presented or proposed recommendations.

As at 30 March 2024, we received 37 responses from senior policy makers currently working within government, comprised of:

➔ 19 from the UK
➔ 14 from Australia
➔ 3 from the U.S.
➔ 1 from Canada

Respondents were from the following COFOG categories (noting that respondents could select more than one category):

COFOG 1 category	**Count of responses**
Social protection	21
Environmental protection	20
Health	18
Economic affairs	15
Education	11
General Public Services	11
Other	7
Defence	3
Housing and community amenities	3
Public order & safety	3
Recreation, culture and religion	3

Survey respondents were asked the following questions:

1. Personal details: Name, email, job title, department they work in, country of work, what activity best describes their role (policy, research & evaluation, etc.).

2. What areas they work in, using a re-mapped version of the COFOG Level 1 & Level 2 categories (see below)
3. For the selected COFOG Level 2 categories, they were then asked the following questions:
    a. How they rate the quantity of primary evidence [1-5 scale]
    b. How they rate the quality of primary evidence [1-5 scale]
    c. How they rate the quantity of secondary evidence [1-5 scale]
    d. How they rate the quality of secondary evidence [1-5 scale]
    e. The impact evidence has on the money and time spent, and practices and policies adopted [1-5 scale]
    f. How easy it is to access, understand and interpret evidence [1-5 scale]
    g. What sources they use to find evidence? [free text]
    h. Barriers they face in the adoption of quality evidence [free text]
    i. Their confidence in the responses they have provided [0-100 sliding scale]
4. They would then repeat these questions for any other areas they selected, if they selected multiple COFOG Level 2 categories
5. Finally, they were asked if there are any specific areas they feel there is a pressing need for more robust evidence for policy making [free text]

Light-touch qualitative text analysis was conducted on free text responses to the barriers faced to the adoption of evidence which were presented in Section 2. This involved a qualitative researcher manually categorising responses and ordering them by number of mentions to provide a sense of consensus in responses.

COFOG Level 2 re-categorisation

To make it easier for respondents of the survey to select an appropriate response, we re-mapped the COFOG Level 2 categories as follows:

COFOG Level 1	COFOG Level 2	Re-categorisation
Economic Affairs	General economic, commercial, and labour affairs	General economic policy
	Agriculture, forestry, fishing, and hunting	Sector specific economic policy (please specify)
	Fuel and energy	
	Mining, manufacturing, and construction	
	Transport	
	Communication	
	Other industries	
	R&D economic affairs	-
	Economic affairs not elsewhere classified	Economic affairs not elsewhere classified
Environmental	Waste management	Environmental policy (waste,

Protection	Waste water management	recycling, pollution and biodiversity)
	Pollution abatement	
	Protection of biodiversity and landscape	
	R&D environmental protection	-
	Environmental protection not elsewhere classified	Environmental protection not elsewhere classified
		Climate change
		Energy
Housing and Community Amenities	Housing development	Housing and Community development
	Community development	
	Water supply	Public utilities (water, lighting...)
	Street lighting	
	R&D housing and community amenities	-
	Housing and community amenities not elsewhere classified	Housing and community amenities not elsewhere classified
Health	Medical products, appliances, and equipment	Primary healthcare
	Outpatient services	
	Hospital services	Hospital services (Secondary & Tertiary healthcare)
	Public health services	Public health
	R&D health	-
	Health not elsewhere classified	Health not elsewhere classified
Recreation, Culture, and Religion	Recreational and sporting services	Sports, recreation, culture and media
	Cultural services	
	Broadcasting and publishing services	
	Religious and other community services	Religion and other community services
	R&D recreation, culture, and religion	-
	Recreation, culture, and religion not elsewhere classified	Recreation, culture, and religion not elsewhere classified
Education	Pre-primary and primary education	Pre-primary and primary education
	Secondary education	Secondary education

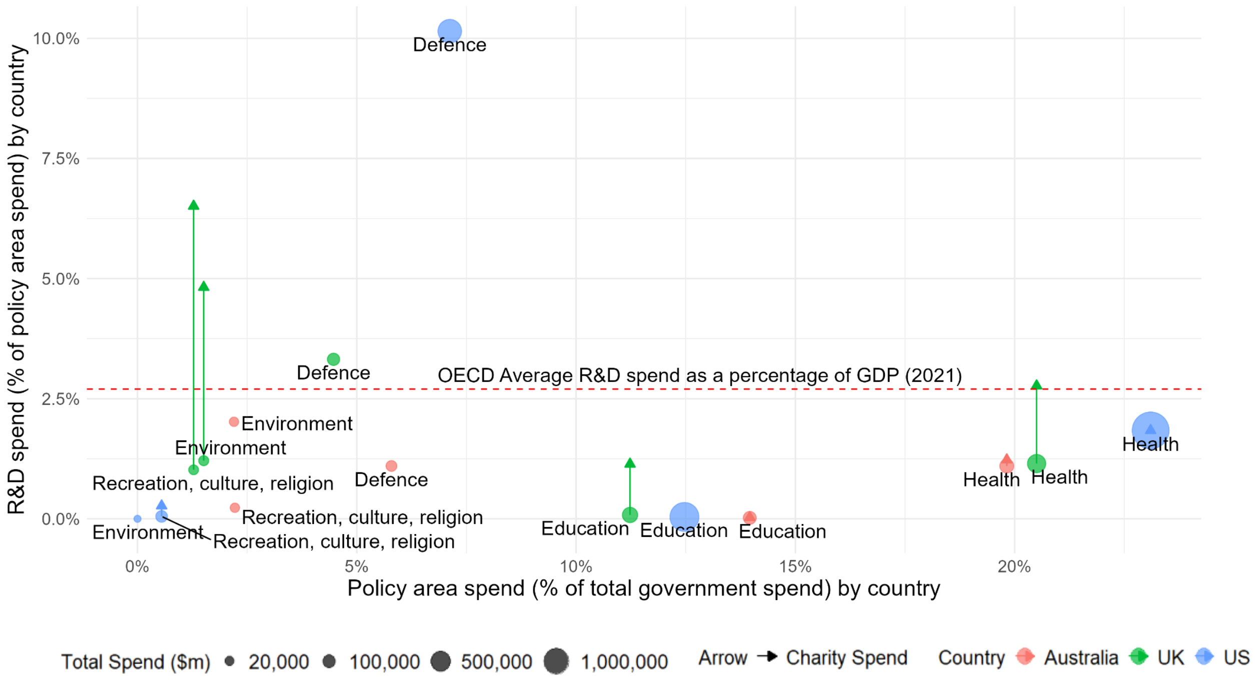
	Post-secondary non-tertiary education	Higher education and vocational training
	Tertiary education	
	R&D education	-
	Subsidiary services to education	Education not elsewhere classified
	Education not definable by level	
	Education not elsewhere classified	
Social Protection	Sickness and disability	Disability, sickness and old age
	Old age	
	Survivors	Family & children
	Family and children	
	Unemployment	Unemployment
	Housing	Housing
	R&D social protection	-
	Social exclusion not elsewhere classified	Social protection not elsewhere classified
	Social protection not elsewhere classified	
General Public Services	Executive and legislative organs, financial and fiscal affairs, external affairs	Governance and administration
	General services	Fiscal and economic services
	Public debt transactions	
	Transfers of a general character between different levels of government	
	R&D general public services	-
	Basic research	
	Foreign economic aid	Foreign economic aid
	General public services not elsewhere classified	General public services not elsewhere classified
Defence	Military defence	Military defence
	Civil defence	Civil defence
	Foreign military aid	Foreign military aid
	R&D defence	-
	Defence not elsewhere classified	Defence not elsewhere classified
Public Order and Safety	Police services	Police services
	Law courts	Prisons and courts
	Prisons	

	R&D public order and safety	-
	Fire-protection services	Public order and safety not elsewhere classified
	Public order and safety not elsewhere classified	

We removed R&D from the categories given the respondents of the survey were policy professionals.

Appendix 3: R&D compared to total expenditure, including the top 10 charities

R&D spend compared to total expenditure by policy area (Australia, UK, U.S.) including the top 10 charities within each country 2021-22



The largest change when introducing charity spend is in the UK, where there is significant expenditure by charities on Health, Education, Environment and Recreation, Culture and Religion. Charity does not appear to demonstrably change the figures for Australia or the U.S., hence the arrows are not discernible on the graph above. Though, it is worth noting that this graph is simply presenting the additional spend by charity as data is not readily available to apportion charity spending into research activity. It also rests on a partial list of funding sources that provide an indication only of spend, rather than a definitive depiction.

Data sources:

- OECD data
  - Public finance by function
  - Government budget allocations for R&D

- Charity data:
    - UK: Top 10 charities
    - U.S.:
        - Feeding America.
        - Salvation Army
        - Good 360
        - Direct Relief
        - St Jude’s Children’s Hospital
        - Americares
        - Samaritan's Purse
        - Goodwill
        - Habitat for Humanity | FY2023 Annual Report
        - Boys and Girls Club of America
    - Australia: Australian Charities and Not-for-profits Commission - Charity Register

Appendix 4: Diagnostic of policy areas

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
Economic Affairs	General economic policy	★★★☆☆	★★☆☆☆	★★★☆☆	Clearinghouse for Labor Evaluation and Research MAPS Pathways to Work Evidence Clearinghouse WWC Local Economic Growth	Possible candidate for international professional networks within specific sectors of economic policy. May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions the work	The Bill and Melinda Gates Foundation The Pew Charitable Trusts
	Sector specific economic policy	★★★☆☆	★★★☆☆	★★★☆☆			

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
Environmental Protection	Environmental policy (waste, recycling, pollution and biodiversity)	★★★☆☆	★★★☆☆	★★★☆☆	What Works in Conservation Collaboration for Environmental Evidence	Pilot evidence gap maps.Subject to the results of the gap map, pilot meta-LER (subject to evidence gap mapping). A competition can be run through the pooled evaluation fund to determine which institution commissions the evidence gap maps and meta-LER.	The Pew Charitable Trusts The Ian Potter Foundation
	Climate change	★★★☆☆	★★★☆☆	★★☆☆☆	Collaboration for Environmental Evidence	Pilot evidence gap maps. The Collaboration for Environmental Evidence may wish to commission this work.	Wellcome The Pew Charitable Trusts The Ian Potter Foundation
	Energy	★★★☆☆	★★★☆☆	★★☆☆☆		Pilot evidence gap maps. The Collaboration for Environmental Evidence may wish to commission this work.	

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
Housing and Community Amenities	Housing and Community development	★☆☆☆☆	★★☆☆☆	★★☆☆☆	None identified	Pilot evidence gap maps.Subject to the results of the gap map, pilot meta-LER (subject to evidence gap mapping). A competition can be run through the pooled evaluation fund to determine which institution commissions the evidence gap maps and meta-LER.	Paul Ramsay Foundation The Ian Potter Foundation
	Public utilities (water, lighting...)	*No survey responses*	*No survey responses*	*No survey responses*		May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions the work	

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
Health	Primary healthcare	★☆☆☆☆	★★☆☆☆	★★☆☆☆	Multiple, including NICE, WWFH, McMaster, Institute of Evidence-Based Healthcare	Existing research and evidence institutions already focus strongly on primary, secondary and tertiary healthcare. We therefore do not see a need for immediate collaboration in these areas.	Wellcome
	Hospital services (Secondary & Tertiary healthcare)	★★☆☆☆	★★☆☆☆	★★☆☆☆			
	Public health	★★★☆☆	★★★★☆	★★★☆☆		May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions the work	
Recreation, Culture, and Religion	Sports, recreation, culture and media	★★☆☆☆	★★★☆☆	★★★☆☆	None identified	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted.	

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						A competition can be run through the pooled evaluation fund to determine which institution commissions the work	
	Religion and other community services	*No survey responses*	*No survey responses*	*No survey responses*		May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions the work	

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
Education	Pre-primary and primary education	★★☆☆☆	★★☆☆☆	★★★☆☆	EEF WWC IES E4L	Synthesis of evidence is already seeing improvement through existing evidence institutes. Scale up the existing Evidence for Education Network, bringing in the U.S. and Canada and conduct an in-person event to bring together education evidence professionals.	Jacobs Foundation
	Secondary education	★★☆☆☆	★★☆☆☆	★★☆☆☆		Pilot evidence gap maps.Subject to the results of the gap map, pilot meta-LER (subject to evidence gap mapping). A competition can be run through the pooled evaluation fund to determine which institution commissions	

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						the evidence gap maps and meta-LER.	
	Higher education and vocational training	★★★☆☆	★★★☆☆	★★☆☆☆	TASO WWC IES	Pilot evidence gap maps. A competition can be run through the pooled evaluation fund to determine which institution commissions the evidence gap maps.	Arnold Ventures Alfred P. Sloan Foundation
Social Protection	Disability, sickness and old age	★★☆☆☆	★★☆☆☆	★★★☆☆	Centre for Ageing Better	Pilot evidence gap maps.Subject to the results of the gap map, pilot meta-LER (subject to evidence gap mapping). The Centre for Ageing Better could house this work, though it has a greater focus on old age, than disability or sickness. As such, it may be suitable to run a competition through the pooled evaluation fund to	UKRI / ESRC Wellcome

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						determine which institution commissions the work.	
	Family and children	★★☆☆☆	★★☆☆☆	★★★☆☆	Multiple, including Australian Research Alliance for Children and Youth, CEBC, Foundations, YEF and YFF	Pilot evidence gap maps.Subject to the results of the gap map, pilot meta-LER (subject to evidence gap mapping). A competition can be run through the pooled evaluation fund to determine which institution commissions the evidence gap maps and meta-LER.	Paul Ramsay Foundation William T. Grant Foundation The Ian Potter Foundation
	Unemployment	★★★☆☆	★★☆☆☆	★★★☆☆	Clearinghouse for Labor Evaluation and Research Pathways to Work Evidence Clearinghouse	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to	The Bill and Melinda Gates Foundation The Pew Charitable Trusts Paul Ramsay Foundation

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						determine which institution commissions the work	
	Housing	★★☆☆☆	★★☆☆☆	★★★★☆	Canadian Observatory on Homelessness Centre for Homelessness Impact	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions the work	The Bill and Melinda Gates Foundation The Melville Charitable Trust Paul Ramsay Foundation The Oak Foundation
General Public Services	Governance and administration	★★★☆☆	★★★☆☆	★★★☆☆	None identified	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition can be run through the pooled evaluation fund to determine which institution commissions	Pew Charitable Trust Susan McKinnon Foundation

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						the work	
	Fiscal and economic services	★★★★☆	★★★★☆	★★★★☆	None identified	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. Possible candidate for international professional networks within specific sectors of economic policy.	Arnold Ventures
	Foreign economic aid	*No survey responses*	*No survey responses*	*No survey responses*	Evidence aid Development Experience Clearinghouse (USAID)	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. A competition could be run through the pooled evaluation fund to determine which institution commissions this work.	
Defence	Military defence	*No survey*	*No survey*	*No survey*	RAND	Defence already sees	*N/a - given no*

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
		responses	*responses*	*responses*	Corporation	significant investment and the five eyes network exists to collaborate on evidence. Therefore, we do not propose any activity in this area.	*activity proposed*
	Civil defence	*No survey responses*	*No survey responses*	*No survey responses*	None identified		
	Foreign military aid	★☆☆☆☆	★☆☆☆☆	★★☆☆☆			
Public Order and Safety	Police services	*No survey responses*	*No survey responses*	*No survey responses*	Australian Institute of Criminology Center for Evidence-Based Crime Policy College of Policing Crime Solutions	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. Possible candidate for creating an international professional network.	UKRI / ESRC
	Prisons and courts	★★☆☆☆	★★☆☆☆	★★★☆☆	What Works in Reentry Clearinghouse	May be suitable for evidence gap mapping and meta-LERs, once pilots have been conducted. The What Works in Reentry Clearinghouse may	Paul Ramsay Foundation Arnold Ventures

COFOG Level 1	COFOG Level 2 (recategorised)	Primary evidence (per survey of policy makers)	Secondary evidence (per survey of policy makers)	Ease of accessing evidence (per survey of policy makers)	Relevant institution	Proposed action	Potential funder
						wish to commission this work.	

Diagnostic of COFOG Level 2 policy areas. Star rating is based on scores from the survey with policy makers. A combined score was derived from the survey results for quantity and quality of evidence, ranging from 1-25 for both primary and secondary evidence. 1 star = 1-5, 2 star = 6-10, 3 star = 11-15, 4 star = 16-20, 5 star = 21-25. Ease of accessing evidence was derived from the survey results for access to evidence, ranging from 1-5. The 1-5 ratings were mapped onto star ratings.

Appendix 5: Glossary of terms

- **Classifications of the Functions of Government (COFOG)**: COFOG is a classification defined by the United Nations Statistics Division. It classifies government expenditure data from the System of National Accounts based on the intended purpose of the funds.
    - **Levels in COFOG**: Level 1 COFOG (Classification of the Functions of Government) categorises expenditure data into ten "functional" groups or sub-sectors (such as defence, education, and social protection). Level 2 further divides each of these first-level groups into up to nine sub-groups.
- **Evidence adoption**: Integration of evidence into policy advice and decision making. Policy and decision makers rely on both primary and secondary evidence in addition to other considerations such as: politics, public attitudes and acceptability; affordability; and public service capability and capacity.
- **Evidence gap map**: A tool used to represent, normally visually, the existing evidence on a particular topic, highlighting areas where there is substantial research (evidence) and areas where information is lacking (gaps).
- **Living evidence review (LER)**: A systematic review that is continuously updated to incorporate new research. This approach ensures that the review remains up to date and reflects the most recent evidence.
- **Match funding:** Financial arrangement by which funds provided by one source (often a grant-making body or government) are matched by funds from a second source, such as a private organisation, a non-profit, or the beneficiaries themselves. The objective is to share the cost of an initiative between multiple parties, thereby leveraging additional resources and promoting collaboration.
- **N:** In statistics, sample size (i.e. total number of observations or participants in a study).
- **Order of magnitude:** the scale or size of a value in terms of powers of ten.
- **Primary evidence:** Original data and documentation collected directly from research, observation, or experimentation. It may be generated through both quantitative and qualitative research methods.
- **Randomised control trial (RCT)**: Form of scientific experiment used to control factors not under direct experimental control by randomly allocating participants to the treatment or control group. It is considered the gold standard for testing the efficacy of interventions due to its ability to minimise bias and establish causality.
- **Secondary evidence:** Systematic collation and synthesis of primary evidence to enhance the accessibility and clarity of existing research in a specific field. Secondary evidence, including systematic reviews and meta-analyses, can be produced when there is a robust body of primary evidence available.