OECD Public Governance Reviews

Mobilising Evidence for Good Governance

TAKING STOCK OF PRINCIPLES AND STANDARDS FOR POLICY DESIGN, IMPLEMENTATION AND EVALUATION



OECD

OECD Public Governance Reviews

Mobilising Evidence for Good Governance

TAKING STOCK OF PRINCIPLES AND STANDARDS FOR POLICY DESIGN, IMPLEMENTATION AND EVALUATION



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

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

ISBN 978-92-64-75266-5 (print)
ISBN 978-92-64-85138-2 (pdf)

OECD Public Governance Reviews
ISSN 2219-0406 (print)
ISSN 2219-0414 (online)

Photo credits: Cover © Panchenko Vladimir/Shutterstock.com.

Corrigenda to publications may be found on line at: *www.oecd.org/about/publishing/corrigenda.htm*.

© OECD 2020

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

Foreword

Governments are increasingly facing complex, multidimensional challenges requiring them to make policy trade-offs and assess the implications of decisions on many different fronts. Some of the challenges are 'wicked' in nature, involving social fragmentation, institutional complexity, and scientific uncertainty. The COVID-19 pandemic has further increased the number of "unknown unknowns" that governments face. The need for evidence to navigate these issues as well as for global collaboration in this area is clearer than ever. Moreover, in an age where trust has been severely eroded in some countries, governments must demonstrate the quality and reliability of the evidence on which their policies and decisions are based.

The traditional hierarchical approach to knowledge dissemination in government relies upon recommendations from scientists and academia, as well as from clearinghouses, knowledge brokers and government bodies specialised in gathering and translating evidence. Yet, the standards for such evidence differ among organisations.

This report offers, for the first time, a mapping of existing *principles for the good governance of evidence,* which are focused on the process through which evidence is obtained, followed by a comprehensive stocktaking of *standards of evidence*, which apply to the content and quality of the evidence produced.

It provides concrete solutions and approaches that can be used by knowledge brokers and agencies to ensure that their evidence can be safely used by policy makers. It seeks to promote greater coherence among the standards applied, and thus increase confidence in them among users. It also offers a set of tools that can help analysts and practitioners mobilise quality evidence for decision making, thereby promoting sound public governance. In the longer term it can serve as a first step towards developing guidance on strengthening evidence-informed policy making.

This report is based on research and engagement with a wide variety of experts. It is the third in a series of OECD reports in this area, together with *Building capacity for evidence-informed policy-making* and *Improving Governance with Policy Evaluation*. The report also complements OECD work on a data-driven public sector, which emphasises the importance of data governance.

Acknowledgements

The report was prepared by the OECD Directorate for Public Governance (GOV) under the leadership of Elsa Pilichowski, as part of the work programme of the Governance Reviews and Partnerships Division, headed by Martin Forst.

The report was drafted by Daniel Acquah, policy analyst at the OECD at the time of drafting, under the strategic direction of Stephane Jacobzone, Head of Unit for evidence, monitoring and policy evaluation. Lizeth Fúquene also contributed significantly to the report drafting. Danielle Mulima and Katarzina Lisek contributed in the data collection and provided highly valuable research assistance. The report was prepared for publication by Ciara Muller and Amelia Godber helped with the cover. Roxana Glavanov and Meghan Hennessy provided secretarial assistance. Andrea Uhrhammer provided editorial guidance.

This report benefitted from insightful comments and feedback from many experts in government and academia. In particular, the authors would like to thank to all the panel of invited experts who participated in the expert meetings that took place in October 2018 at the inception of the project: Jonathan Breckon (UK, The Alliance for Useful Evidence), Diana Epstein (US, Office for Management and Budget); Peter Gluckman (President Elect, International Science Council and Chair International Network for Government Science Advice), Justin O. Parkhurst, (UK, and London School of Economics), Paolo Paruolo (Head of the Competence Centre on Microeconomic Evaluation, European Commission), Ulla Rosenström (Prime Minister's office, Finland) Knut Sundell (Sweden, Swedish Agency for Health Technology Assessment and Assessment of Social Services), Howard White (CEO of Campbell Collaboration). The authors would also like to thank the participants in an expert meeting in June 2019 to discuss a first version of the report and in subsequent exchanges. The report received detailed and comprehensive feedback from government officials. The authors are grateful in particular for the feedback and materials received from the following experts: Gabor Arany SGG, France; Adam Baiz and Marine de Montaignac (France Stratégie); Francis P. Crawley (Good Clinical Practice Alliance - Europe & EWG European Academy of Paediatrics); Charlotta Eriksson and Helen Forslind (Sweden); Marica Ferri (European Monitoring Centre for Drugs and Drug Addiction - EMCDDA); Takeshi Fukaya (Advisor, Ministry of Internal Affairs and Communications - MIC), Japan; David Gough (UCL London); Daniel Gruss (Digitization of Public Administration Office, Germany); Marta Hugas (European Food Safety Authority), Julianna Kisné (Prime Minister's Office, Hungary); Marie Lintzer (HAVP France); Susannah Luck (Australian Public Service Commission), Majella McCloskey Nuala Doherty (Centre for Effective Services, Ireland); David Mair (EC Joint Research Centre); Robyn Mildon, (Centre for Evidence and Implementation, Melbourne); Aline Pennisi and Alessandra de Lellis (Ministry of Finance Italy); Barbara Schmidt (Eurofound); Kasey Treadwell (Department of Children and Youth Affairs, Ireland); Ruth Puttick (NESTA, UK); Nynke De Witte (Ministry of Finance, the Netherlands).

The authors are grateful for the feedback received from Monica Brezzi, Barbara Ubaldi, Benjamin Welby, Arturo Rivera Perez and Cecilia Emilsson, of the OECD Public Governance Directorate, as well as from Andrew Wyckoff, Dirk Pilat and Suguru Iwaya of the OECD Directorate for Science, Technology and Innovation.

This document was approved by the OECD Public Governance Committee on 8 December 2020 and declassified on 8 December 2020.

Table of contents

Foreword 3

Acknowledgements 4

Executive summary 8

1 Introduction 10
- The role of quality evidence to support evidence-informed policy-making 11
- The role of the use of evidence in improving public sector effectiveness 14
- Goal and focus of this report 16
- Methodology 18
- Consolidated overview of the principles and standards 19
- References 23
- Notes 26

2 Principles for the good governance of evidence 27
- Appropriate evidence for the policy concern 28
- Ensuring integrity (honest brokerage) 31
- Accountability 36
- Contestability 38
- Public representation in decision making 42
- Transparency in the use of evidence 44
- Building evidence through emerging technologies and mobilising data 47
- References 54
- Notes 60

3 Standards of Evidence: Mapping the experience in OECD countries 62
- Introduction 63
- Evidence Synthesis 66
- Theory of Change and Logic Model underpinning the programme 74
- Design and Development of Programmes 81
- Efficacy of an Intervention 87
- Effectiveness of Interventions 92
- Cost (effectiveness) of Interventions 96
- Implementation and Scale up of Interventions 107
- References 111
- Notes 119

4 Conclusion 120
What has this report achieved? 121
Opportunities for future work 121
References 122
Note 123

Annex A. Overview of the principles for the use of evidence across jurisdictions 124

Annex B. Mapping of Existing Standards of Evidence across a Range of Jurisdictions 128

Tables

Table 2.1. Honest brokerage roles and their attributes in Finland 32
Table 3.1. Two Basic Approaches to Evaluate Public Sector Innovation 96

Figures

Figure 1.1. Mapping of the principles and standards for the good governance of evidence 20
Figure 2.1. Appropriate evidence for the policy context 28
Figure 3.1. Focus of the standards of evidence 63
Figure 3.2. Type of evidence 65
Figure 3.3. Methodology for conducting a review 72
Figure 3.4. Monitoring and Evaluation framework 76
Figure 3.5. Decision tree for the different types of economic evaluation 98
Figure 3.6. Sample table of CBA results 106

Boxes

Box 1.1. Background definitions of types of evidence 13
Box 1.2. An overview of related OECD work on Evidence-Informed Policy-Making 17
Box 2.1. Context Matters: Transporting Nurse Family Partnership to different contexts 30
Box 2.2. Ensuring the ethical actions of researchers providing advice on policy implementation 33
Box 2.3. OECD Guidelines for managing conflict of interest in the public service 34
Box 2.4. Creating a Governance Framework for Ministerial Advisors 37
Box 2.5. The European Food Safety Authority's Guidance on Expert Knowledge Elicitation in food and Feed Safety Risk Assessment 39
Box 2.6. HomVEE's 'Requests for Reconsideration of Evidence Determinations' 41
Box 2.7. Stakeholder engagement in NICE recommendations 41
Box 2.8. Using Mini-Publics for Deliberation 43
Box 2.9. Transparency in the Presentation of Evidence 45
Box 2.10. Transparency in Evidence Synthesis 46
Box 2.11. New Zealand Principles for the safe and effective use of data and analytics 50
Box 2.12. The OECD Principles for Responsible Stewardship of Trustworthy AI 52
Box 2.13. Good Practice Principles for Data Ethics in the Public Sector 53
Box 3.1. Evidence Syntheses as a tool to support decision making 67
Box 3.2. The Education Endowment Foundation's Padlock Rating System 69
Box 3.3. ROBIS - tool to assess risk of bias in systematic reviews 70
Box 3.4. Crime Solutions – Practice Scoring Instrument 71
Box 3.5. Grading of Recommendations Assessment, Development and Evaluation (GRADE) 73
Box 3.6. The benefits of developing an intervention theory of change and logic model 75
Box 3.7. Guidance Document on Monitoring and Evaluation – European Commission 76
Box 3.8. The Green List Prevention – Conceptual Quality Criteria 77
Box 3.9. Two approaches to grading the quality of a theory of change 78
Box 3.10. EMCDDA's European drug prevention quality standards – intervention description 79
Box 3.11. Office of Juvenile Justice and Delinquency Prevention Model Programs Guide 80

Box 3.12. Feasibility: Can an intervention work? 82
Box 3.13. Standards concerning Design and Development: Green List Prevention 83
Box 3.14. Definitions: Reliability and validity 84
Box 3.15. Standards concerning Design and Development: Social Policy Evaluation and Research Unit-Superu 85
Box 3.16. Standards concerning Efficacy: What Works Centre for Local Economic Growth 89
Box 3.17. Standards concerning Effectiveness: Blueprints 94
Box 3.18. Type of economic evaluations 97
Box 3.19. Cost rating bands for EEF Toolkit 99
Box 3.20. Methods for costing an intervention 100
Box 3.21. Type of costs 101
Box 3.22. Summary statistics on CBA 102
Box 3.23. Health Valuation 103
Box 3.24. Examples for avoiding double counting and externalities: The Treasury of New Zealand 104
Box 3.25. Mechanism to test uncertainty 105
Box 3.26. Canadian Best Practices Portal – Adaptability of an intervention 108
Box 3.27. EU-Compass for Action on Mental Health and Well-being 109
Box 3.28. SUPERU's intervention consistency and documentation criteria 110

Follow OECD Publications on:




http://twitter.com/OECD_Pubs


http://www.facebook.com/OECDPublications


http://www.linkedin.com/groups/OECD-Publications-4645871


http://www.youtube.com/oecdilibrary



http://www.oecd.org/oecddirect/

Executive summary

This report looks at how countries can mobilise evidence for good governance, taking stock of standards that OECD countries may use to support the design, implementation, and evaluation of public interventions. This is part of a broad agenda on fostering evidence-informed policy-making that is shared by many countries. Existing knowledge brokers, such as the UK What Works Centres, the US clearinghouses or other government bodies specialised in providing quantitative evidence, have made extensive progress in developing standards of evidence for many areas where the public sector can make a decisive difference in terms of inclusive growth, such as health, education, employment, labour and social policies. Yet, users report being confused at the differences among the standards used by various organisations to determine what counts as evidence and how evidence is to be employed in policy-making. Bringing coherence to existing approaches would offer governments more operational options for using evidence.

A first goal of this report is to increase coherence among existing approaches, and the thus raise confidence in them. The existence of robust standards of evidence is fundamental for convincingly communicating on the quality of evidence brought to bear in the evaluation of policies and programmes. The current review is designed to serve central government agencies and knowledge brokers aiming to improve public sector effectiveness by strengthening the evidence and decision-making interface.

Evidence in this report is not limited to data and statistics; rather, it refers to a broader range of analysis that can also include evaluations, scientific research, and other facts derived from observation or social sciences. Knowledge brokers have an important role to play in bridging the divide between producers of raw evidence and policy makers. They carry out a range of functions, including that of ensuring the appropriate evidence makes its way to policymakers in a timely, attractive, relevant, and accessible manner. This evidence has to feed into the policy making cycle for designing, implementing and evaluating policies at the right time and the right format.

The report maps a wide range of existing standards of evidence, related to the quality of content, including over 50 existing approaches. These standards are derived from selected existing and varying practices currently in use by OECD countries and international research bodies. The report synthesises these existing standards through a stocktaking exercise, articulating the results across seven key dimensions: the role of evidence synthesis; the theory of change and logic underpinning an intervention; the design and development of policies and programmes; the efficacy, effectiveness, and cost effectiveness of an intervention as well as how standards for evidence are related to the implementation and scale up of public interventions.

The consideration of standards of evidence, while necessary, is not sufficient for an evidence-informed approach that maintains trust in policy making processes for the design, implementation, and evaluation of public interventions and policies. Maintaining trust requires an investment in good governance, so that citizens and the broader public can infer that the policy process is designed as a way for the government to be able to achieve the best possible outcome taking into account existing constraints. Evidence is but one contributing element to policy making. Policy makers must also weigh up other considerations, such as ethics, economics, and political considerations. Policy makers need to think critically about what

evidence is needed in the context of a decision, as well as whose voices, interests, and values need to be taken into account.

As a result, the report also presents a set of principles for the management of evidence used across OECD countries. These principles focus on the process and the conditions for good use through which evidence is brought to bear as part of any process aimed at ensuring trust in decision making. The seven principles identified include: whether evidence is appropriate for the policy concern, accountability of the policy advisory system through which evidence is brought to bear, contestability (i.e. that evidence is open to questioning and has been subject to quality control), public representation in decision making (through public engagement), transparency in the use of evidence, and building evidence through emerging technologies and mobilising data in some ethical ways. The report maps 29 approaches that cover one or more such principles.

Altogether, the combined analysis seeks to improve the quality and use of “standards” and “principles” of evidence for government agencies through a comprehensive mapping and stocktaking approach. While the report is not prescriptive as such, for each of the 14 principles that have been highlighted, it offers a set of questions to help frame the main concerns that should be raised to the attention of government agencies and knowledge brokers dealing with evidence to inform the policy-making cycle. This analysis would aim to support better use of evidence by government agencies.

The findings of this report can strengthen the evidence and decision-making interface, bringing greater understanding and support to the role of knowledge brokerage organisations in their efforts to improve public sector effectiveness. The use of evidence should not be limited to informing the policy-making processes. Evidence is also critical to identifying positive and negative consequences of public interventions, while quality evidence can contribute to learning how to maximise benefits and minimise damages of public interventions. Evidence matters for understanding how existing programmes can be evaluated and improved. This “learning” dimension is an important part of the process of evaluating policy choices and developing systematic approaches to improving public policy and improving good governance. This report highlights the benefits of and the need for global collaboration in terms of data and evidence sharing to support both scientific and government processes. The mapping offered by the report can help further strengthen evidence informed policy making in the future and support increasing citizen trust in the decision-making process.

1 Introduction

The chapter begins by introducing the role of quality evidence for policy making, the opportunities for evidence-informed policy-making to improve public sector effectiveness, and the current challenges to using evidence in decision-making processes. In response to these challenges, the chapter explores a core set of characteristics in terms of standards and governance principles to support policy makers in the use of evidence for decision-making. The chapter then presents the goal and focus of the report, along with the methodology. Lastly, the chapter summarises the principles and standards that are examined in chapter 2 and chapter 3.

The role of quality evidence to support evidence-informed policy-making

This report reviews the principles and standards that OECD countries use to support the design, implementation and evaluation of public interventions, as part of their efforts to foster evidence-informed policy-making. The report addresses two challenges in realising the potential of evidence-informed policy-making. The first challenge is that not all evidence is equal: some evidence is more robust and trustworthy and deserves to be given more weight in decision-making. Determining when evidence is robust and sufficient, and then communicating this to decision makers is challenging. The evidence concerned in this report is the whole range of analysis that can be made using various information and data sources that are brought to be bear in the design, implementation and evaluation of policy interventions. Standards of evidence have been developed by knowledge brokers in many countries to help communicate the strength of evidence of policies and programmes to policy makers. The second challenge is that evidence is one of many inputs of policy-making and policy makers must balance various considerations such as ethics, equity, values, economics, special interests, privacy, security, and political objectives in order to maintain trust in the policy making process. Principles for the good governance of evidence help to ensure that evidence is used appropriately and with reference to the wider administrative and political system into which it is being introduced. At the same time, it is important to recognise the messy nature of the policy process where conflicting values, interests and constraints have to be addressed. While evidence can should inform policy-making, it certainly cannot "determine" it, as political decision needs to be given a discretionary space of appreciation, which reflects the political responsibility entrusted to policy makers and ministers in government. This is the reason why the approach is qualified as "Evidence-Informed Policy Making", against "Evidence-Based Policy Making", recognising that policies are also driven by values, conflicts and other non-evidential factors.

What are the challenges of contemporary policy-making?

Contemporary policy problems are increasingly 'wicked' in nature, involving social pluralism, institutional complexity, and scientific uncertainty (Head and Alford, 2015[1]). Simultaneously, public trust in government and political organisations has been eroded, following the global financial crisis in a large number of countries at a time when arguably, policy makers need it most (OECD, 2017[2]). Trust has since then recovered on average to pre-crisis levels, but not in the countries that were most affected (OECD, 2019[3]).

Countries, however, have now fallen into a much deeper crisis in public confidence caused by the COVID 19 pandemic. In particular, it has tested governments' capacities to maintain and ensure trust in decision-making processes that have an immediate and deep impact on the lives of their citizens. The impacts of eroding trust in government and institutions are wide ranging, and include consequences for the effective functioning of government, which may result in costly micromanagement between organisations, government, and citizens (Prange-Gstöhl, 2016[4]). The COVID 19 pandemic has also put science under increased public scepticism: "Trust in research and its role in political decision making and policy changes have never been more at the forefront of public discussion and scrutiny than during the current public health crisis" (The Lancet, 2020[5]). Indeed, during this Covid-19 crisis the fabric of our democracies appears threatened by an inability of government to present a convincing model of the role of science and evidence in decision-making. This report is preceded by years of research and discussion. It arrives in a timely manner, asking the question: "But what is everyone's role in strengthening this trust?" (The Lancet, 2020[5]); In particular, the role of knowledge brokers and government agencies in providing evidence to policymakers as they navigate a complex and challenging political landscape.

Moreover, the questioning of the role of traditional institutions has coincided with the digitalisation of societies, including within the public sector, and the adoption of digital communications platforms. There is growing concern about the potential for disinformation (so called fake news nowadays) through traditional and social media, where the origins and motivations of traditional and new sources of evidence are questioned. Both trends have only been exacerbated during the COVID 19 crisis. This has led to the

traditional hierarchical approach to the dissemination of knowledge being replaced by peer-to-peer recommendations and algorithms, even if the need for authoritative voices remains (D'Ancona, 2017[6]). Whilst citizens still hold science in high regard and there is a high level of trust in scientists (Yarborough, 2014[7]) what can no longer be taken for granted is the authority of science in the face of other pressures and alternative sources of 'knowledge'. The many "unknown unknowns" and uncertainty created by, for example, the COVID 19 crisis have also generated significant challenges while increasing the reliance on science in the decision making process.

Another enabler for trustworthy and evidence-informed decision making is the increasing capacity of governments to collect, process and store digital data, and to integrate these into policy processes. The OECD has had for a long time a recommendation for good statistical practice, which is directed at the legal and institutional frameworks for official statistics, the professional independence of national statistical offices, the resources, data privacy issues, the right to access administrative sources, as well as the impartiality, objectivity and transparency of official statistics, together with sound methodology and professional standards.[1] More recently, the use and application of data in crisis management efforts has aimed at enabling more transformative, open, collaborative, pinpointed and agile data driven approaches while reemphasising challenges to good data governance, including within the public sector, such as lacking health data standards and crisis-adjusted data ethics (OECD, 2019[8]). The OECD *Recommendation on Health Data Governance* (2019[9]) fosters co-ordination within government and among organisations to adopt common data elements and formats, quality assurance, and data interoperability standards; as well as, robust, objective and fair reviews and approval procedures conducted by the expertise necessary. In the case of COVID 19, several countries have turned to digital technologies and advanced analytics to collect, process and share data for effective front-line responses. However, this has also come with new data governance and privacy challenges, which have had to be addressed in a crisis context. (OECD, 2020[10]). For instance, few countries do not count with frameworks to contact-tracing and population-wide surveillance. Some have countered with laws specifying how data collection will be limited to a certain population, for what time, and purpose.

The result is that a wide range of new and alternative sources of information are now available to governments, a plethora of which could all feed into the policy-making and decision making process. It is therefore all the more important that some form of principles and standards for mobilising evidence be developed to strengthen the coherence and the trustworthiness of such decision-making processes. This of course is complementary to all the efforts that are needed upstream to address data certification or data provenance in order to avoid data corruption and therefore misleading outputs. Moreover, in the COVID 19 crisis, public authorities on the field have a key role to play in advising on proposed new legislations and providing clarity regarding the application of existing privacy and data protection frameworks (OECD, 2020[10]).

Sound public governance has a critical role to play in maintaining trust as it can promote fair processes and outcomes for citizens. Good public governance is also need to improve the quality, access and responsiveness of public services. This hinges on the good or "appropriate" use of evidence, to feed into the design, implementation, and evaluation of public programmes and interventions. Both are key features of a smart and agile state.

Defining the role of evidence and evidence-informed policy-making

Defining what 'evidence-informed policy-making' and 'evidence' mean in a policy context is a necessary step in defining how to improve it. Evidence-informed policy-making can be defined as a process whereby sources of information–including statistics, data or available published research–are consulted before making a decision to design, implement and/or alter policies, interventions, programmes. (Langer, Tripney and Gough, 2016[11]). This report adopts a correspondingly broad definition of evidence: 'a systematic investigative process employed to increase or revise current knowledge' (Langer, Tripney and Gough,

2016[11]). The use of evidence contributes to good governance as it enables actors to evaluate, design and update public interventions.

The evidence used to inform policy making processes may be derived from (i) 'scientific evidence' , ii) policy evaluation (iii) anecdotal observations, (iv) subjective opinions polls, each with varying strength of evidence (see Box 1.1). Scientific evidence on the quality and impact of public interventions, policies and programmes is based on the application of the scientific method to data coming from either:

a. Randomized control trials (RCT), i.e. data from a controlled experiments, or
b. Observational data (i.e., data collected through the observation of individuals, communities, organisations).

Box 1.1. Background definitions of types of evidence

- *Scientific evidence* consists of the results of observations and experiments that serve to support, refute, or modify a scientific hypothesis (belief or proposition) when collected and interpreted in accordance with a widely accepted scientific method. The scientific method is an empirical method of acquiring knowledge that has characterized the development of science since at least the 17th century. It involves:
  - the formulation of hypotheses as tentative responses to questions arising from what is already known in a specific field of knowledge and as a way to generate new knowledge;
  - the testing of the hypothesis through a controlled experiment or other methods;
  - the analysis and ongoing investigation of the experimental findings
- *Policy evaluation* is a structured and objective assessment of an ongoing or completed policy or reform initiative, its design, implementation and results. It determines the relevance and fulfilment of objectives, efficiency, effectiveness, impact and sustainability as well as the worth or significance of a policy.
- *Anecdotal evidence* is evidence from anecdotes, i.e. evidence collected in a casual or informal manner and relying heavily or entirely on personal testimony.
- *Opinion polls* are a human research survey of public opinion from a particular sample of a target population.

Sources: Adapted from Oxford Dictionary, Wikipedia. OECD (2016[12]), Open Government: the Global Context and the Way Forward, https://doi.org/10.1787/9789264268104-en. OECD-DAC (2009[13]), Guidelines for Project and Programme Evaluations https://www.oecd.org/development/evaluation/dcdndep/47069197.pdf.

Appropriate scientific methodology needs to be applied to data to transform them into evidence and knowledge. Observational data is often treated with methods labelled quasi-experimental, because they compensate for the departures from an experimental setting via appropriate statistical methodology; (see e.g. (Crato and Paruolo, 2019[14]). Overall strength of evidence comes from the quality and reliability of the data and data sources, from the appropriateness of the methodology applied to control for bias, and from the concurrence of different parts of the evidence.

The role of the use of evidence in improving public sector effectiveness

Achieving the promise of evidence-informed policy-making

Evidence has a critical role to play in improving the quality, responsiveness, and accessibility of public interventions, services or programmes. Evidence also plays a role throughout the key stages of the policy cycle when public interventions are designed, implemented, and evaluated. This may involve an interactive process, where citizens and users can share views, collaborate with peers or express dissatisfaction as part of a feedback loop to better understand needs and embrace innovation. (OECD, 2020[15]).

Policy design benefits from 'policy memory': an understanding of where challenges have been experienced in the past and what previous good practice could be incorporated into the current reform effort. For instance, evidence synthesis, such as systematic reviews (more information in section: Evidence Synthesis), helps policymakers to prevent one-sided policy design, avoid duplication and ensure that scarce resources are directed at areas of policy requiring further solutions. It is also used to identify policies and practices that have been found to be ineffective, suggesting caution should be exercised regarding further investment in the absence of additional refinement and testing (Gough, Oliver and Thomas, 2013[16]; Torgerson and Torgerson, 2003[17]).

The implementation of public interventions and policies requires significant planning and management. Evidence provides an understanding of how policies should be adapted to meet local needs, whilst safeguarding against changes that may affect outcomes: this can make the difference between a successful implementation of an intervention and one that is ineffective or potentially even harmful (Moore, Bumbarger and Cooper, 2013[18]). Gathering evidence on factors that help and hinder implementation can facilitate dissemination of effective interventions at scale and the delivery of outcomes at the population level (Castro, Barrera and Holleran Steiker, 2010[19]). The attention given to implementation aspects has been subject to growing interest from the economics profession over recent years. (Duflo, 2017[20])

The evaluation of public policies and interventions is also critical to understand why some complex policies work and why others do not. As one source of relevant knowledge, policy evaluation supports policy choices through an evidence-informed decision-making process. Solid evidence and its strategic use throughout the policy cycle can foster policy effectiveness, value for money, accountability, and transparency - enriching public scrutiny, facilitating shared learning and ultimately increasing public trust (OECD, 2020[21]). Ex post evaluation can also be complemented by a more agile feedback loop, so that adjustments may be made in real time.

The need to create a supportive institutional environment

An institutionalized approach to evidence gathering and evidence usage contributes to aligning isolated and unplanned evidence efforts into more formal and systematic practices, with the ability to set guidelines and incentives for evidence creation and use (Gaarder and Briceño, 2010[22]).The OECD comparative report on policy evaluation (2020[21]) identifies that legal frameworks constitute a basis for embedding the practice of evaluations across government in a systematic way. Policy frameworks also give strategic direction to a specific sector or thematic area of policy evaluation. A related OECD report (2020[23]) shows that building capacity for evidence use requires strengthening organisational tools, resources and processes, mandates, legislation and regulation. As these two related OECD report provide significant institutional information on country frameworks and good practices, the current report focuses on a core set of characteristics that are relevant to the content of evidence and the process for integrating it in policy making, in terms of standards and governance principles.

Promoting the use of evidence and improving the evidence to policy interface through knowledge brokerage

Ensuring the effective use of evidence in policy-making depends on the capacity, motivation and skill of policy makers and researchers as well as the quality of the political-administrative interface. Whilst many OECD countries have adopted mechanisms to promote a culture of evidence-informed policy-making, the systematic use of evidence and data in policy making often remains an aspiration and requires further progress to be engrained in the civil service as a whole (OECD, 2015[24]).

On the data as such, there is also a need for a coherent data governance approach not only to support the access and sharing of data within the public sector or across different sectors, but to secure the deployment of controls to ensure that data itself is trustworthy and, depending on the data accessed or shared, protected without compromising basic rights in adherence to ethical values (OECD, 2019[25]).

Addressing the challenges of promoting the use of evidence and improving the evidence to policy interface requires an acknowledgement of the fact that the worlds of policy-making and research are very different. Policy-making is intrinsically political and often reflects a mix of complex interactions and deliberative processes. This reflects different professional cultures, resources, imperatives and time frames than those of science (Olejniczak, Raimondo and Kupiec, 2016[26]). Scientific language and discourse is also distinct from the language of policy-making (Meyer, 2010[27]). As a result, policy makers may not have the time and capacity to synthesize the research literature and face a number of obstacles when accessing the latest knowledge and research (Burkhardt et al., 2015[28]; Oliver et al., 2014[29]). The process of translating knowledge and research so that it can be used in the policy-making cycle is messy and complex and requires governments and other stakeholders to create favourable contexts, incentives and a supportive culture for evidence-informed policy-making (Ellen et al., 2013[30]).

One solution for strengthening the evidence-policy interface is establishing ‘knowledge brokers’ to bridge the gap between researchers and decision makers. These are also called ‘intermediary organisations’ in the literature (Proctor et al., 2019[31]; Association of Children’s Welfare Agencies, 2018[32]). Knowledge brokers are organizations or individuals, who thanks to their intermediary position in the system, can establish and maintain links and cooperation between knowledge producers and users. Harnessing their deep knowledge about both the research process and the policy cycle as well as their extended connections with representatives of these two worlds, they can increase availability of the robust knowledge and build a culture and capabilities for evidence use.

Acting as the bridge between knowledge users and producers, broker organizations must fulfil a wide range of functions. To be able to introduce more evidence into decision-making process, they need to be sure that reliable knowledge is accessible from the field. They synthesise what researchers and end users already know, identify what kind of evidence is missing in relation to decision makers’ information needs and fill those gaps by performing new analysis. They need to ensure that gathered evidence is introduced to the right recipients in a timely and attractive manner. They can disseminate knowledge through general channels of communication and also through targeted messages. Their role is to maximise the chances that evidence will be taken on board by decision makers. They must also foster trust in their results, for example through addressing concerns for reliability and for conflict of interest. For this purpose, they have to promote an evidence – based culture among stakeholders, to translate knowledge into usable tools and to build networks of knowledge producers and users.

The need for standards of evidence to bring coherence to policy making processes

While governments recognise the importance of using evidence in the policy-making process, not all evidence is equal. Some evidence is more robust and trustworthy and deserves to be given more weight in decision-making. Figuring out which evidence is robust and communicating this to decision makers is challenging. Standards of evidence attempt to strengthen the evidence of policies and programmes

transmitted to policy makers. Existing knowledge brokers, which include government agencies, evidence-based clearinghouses and "What Works centres" have already made extensive progress in developing standards of evidence that are firmly rooted in the academic literature.

Although standards of evidence have great potential to improve the quality of evidence-informed policy-making, current approaches face several limitations that impede their use. Due to the proliferation of different approaches to standards of evidence, it is possible to reach different judgements about the strength of evidence of the same policy or programme. It is not surprising that stakeholders are confused about the difference between the standards of evidence used by different organisations (Puttick, 2018[33]; Means et al., 2015[34]). This can even happen within the same government, or across two sides of the same ministry, where different key parameters such as the implicit value of a life saved can be used to obtain incoherent calculations supporting a range of apparently equally valid public interventions. In other policy areas, standards cover only a restricted set of the issues that are important for the policymaking process. The existing stocktaking exercises have found that many of the approaches focus on how to determine whether an intervention is efficacious and effective (Gough and White, 2018[35]), which ignores key features of intervention design, implementation and potential for scale-up. Some approaches to evidence standards may also not serve the needs and realities of public policy (Parkhurst and Abeysinghe, 2016[36]).

Ensuring the good governance of evidence

Consideration of the standards of evidence is necessary but not sufficient for an evidence-informed approach to policy-making. It is necessary to also consider balancing these technical features with the inherently political nature of policymaking (Parkhurst, 2017[37]). As evidence is but one input to policy-making and policy and practice decisions actors must also weigh up broader considerations, such as ethics, equity, values and political considerations. Decision makers need to think critically about what evidence is needed in the context of a particular decision – and whose voices, interests and values need to be heard and how power shapes knowledge production and use (Oliver and Pearce, 2017[38]). This ensures that ethical values and appropriate safeguarding can be assured (Leadbeater et al., 2018[39]). The governance of evidence is therefore essential to limit undue bias in decision making, to reduce the potential impact of lobbying and to ensure that government can act in the best general public interest.

Goal and focus of this report

This report aims to improve the quality and use of evidence both for central government agencies and for knowledge brokers through greater uptake of standards and principles, with a view to improving public sector effectiveness and strengthening evidence at the decision-making interface.

The report offers a first mapping of existing principles for the good governance of evidence, followed by a stocktaking of standards of evidence. The methods are described in detail below. Each substantive part offers details of the principles and standards and provides a coherent narrative about their features and focus. For each principle or standard, the report offers detailed case study examples describing the use of principles and standards in different jurisdictions. The report treats these principles and standards as cutting across different parts of the policy cycle. Each chapter offers a self-assessment checklist for the use of agencies and organisations to help them think about key questions that they may consider including in their own approaches. A consolidated presentation of the principles and standards is offered at the end of this introduction. The principles and standards presented in this report should not be seen as a full comprehensive mandatory checklist but as a sort of a mental check of the issues that might need to be addressed. For example, having a logic model, or a theory of change is not always possible, in the case of multiple, competing or conflicting policy goals. However, when feasible, this can be useful and the report offers a range of insights and relevant sources.

At this stage, the current mapping remains descriptive. It opens the floor for future discussions on the issues that should be consider for the review and production of evidence. This report provides an initial resource for agencies and national experts thinking about the role of standards and principles in their evidence architecture. Reflecting both the nature of existing practice and the thematic interests of the project, the current review has primarily focused on public interventions and especially in the social policy area. As a result, the report makes a focused contribution to the full range of considerations involved in providing evidence for policy-making. For example, it is recognised that evidence plays a role in policy issues such as developing a financial aid system in the area of development, and in developing clinical guidelines for how physicians respond to an issue in the health area[2]. However, such issues are not directly addressed in the current report which is focused on domestic policy intervention for the public sector as a whole. Similarly, whilst the chapter of efficacy does refer to OECD's work on the evaluation of regional development policy (OECD, 2017[40]) and industrial policy (OECD, 2014[41]), these issues are not addressed in detail in this report.

This report complements two other existing OECD reports on building capacity for evidence-informed policy-making (OECD, 2020[23]), and on improving governance through policy evaluation (OECD, 2020[21]) (See more in Box 1.2).

Box 1.2. An overview of related OECD work on Evidence-Informed Policy-Making

From a governance perspective, the OECD has also focused on building capacity for evidence informed policy making, through appropriate use of evidence, and has also analysed the institutionalisation, quality and use of policy evaluation systems:

1. **Building capacity for evidence-informed policy-making**, this report analyses the skills and capacities governments require to reinforce evidence-informed policy-making (EIPM) and the capacity to engage with stakeholders. At the individual level, the report presents a core skillset, including: the capacity for understanding; obtaining; assessing; using; engaging with stakeholders; and applying evidence. At the organisational level, the report discusses diagnostic tools to evaluate and build governmental capacities for EIPM as well as broader approaches to strengthen an evidence driven culture, such as the implementation of champions for an evidence driven approach, such as chief evaluators, chief economists or statisticians in Ministries and agencies. (OECD, 2020[23]).
2. **Improving governance through policy evaluation**, this report presents the results of the first significant cross-country survey of policy evaluation practices covering 42 countries. This report adopts a systemic approach around institutionalisation, quality and use of evaluation across countries. It discusses the main actors, the key institutional frameworks and includes a wealth of concrete experiences and good practices to make evaluation happen. The analysis of how to ensure the quality of evaluations and how to institutionalise them in the policy process are also related to the discussion of standards and principles for the governance of evidence in the current report (OECD, 2020[21]).

From a well-being perspective, the OECD has also conducted work on the policy use of well-being metrics, analysing specifically the institutional frameworks through which these metrics are being integrated in policy making (Exton and Shinwell, 2018[42]). In the environmental area, the report on Cost Benefit Analysis and the environment, also provides an overview of methods and conditions for use, and is also referenced in the standards section of this report (OECD, 2018[43]). Other contributions flagged in the text above refer to the evaluation of regional development policy or of industrial policy. As an evidence driven organisation, the OECD focuses on evidence driven processes in many sectoral

areas. The contribution of the approaches highlighted here is to focus on the institutional or methodological conditions that best allow for evidence to be incorporated into policy processes.

Source: Adapted from OECD (2020[23]), Building Capacity for Evidence-Informed Policy-Making: Lessons from Country Experiences https://doi.org/10.1787/86331250-en. OECD (2020[21]), *Improving Governance with Policy Evaluation: Lessons From Country Experiences*, https://doi.org/10.1787/89b1577d-en.

Methodology

Conceptual framework

The elaboration of the conceptual framework was prepared following an initial literature review on evidence-informed policy-making, accompanied by an early engagement from a set of experts from OECD countries, and complemented by an exploratory expert meeting in October 2018. The list of the experts is available in the acknowledgements and their help is gratefully acknowledge.

- **Principles for the good governance of evidence:** This outlook was informed by the research of Justin Parkhurst's (2017[37]) in "The politics of evidence: from evidence-based policy to the good governance of evidence". This led to the extraction of key principles (listed in below), which were reviewed, expanded and refined in cooperation with experts as well as with the OECD public governance community at large. This provided a framework to extract information from a range of other sources relevant to the good governance of evidence.
- **Standards of evidence**: The starting point was "*Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research*" Gottfredson (2015[44]), which was complemented with expert inputs and other OECD sources[3]. This provided a framework to synthesise and analyse other approaches to the standards of evidence, along with existing mapping efforts that had been done originally in the UK and which have now been expanded to an international level through the current report (Puttick, 2018[33]) (Breckon et al., 2019[45]) (Gough and White, 2018[35]). This led to the extraction of standards of evidence (listed in below) which the report then used as a framework in order to extract information from a range of other relevant sources.

Desk research

Comprehensive and iterative desk research was used to examine and synthesize existing approaches to both the principles and standards of evidence. This included the analysis of the clearinghouses and organizations' websites and their available online resources, restricted to the information published in English that was available to the Secretariat by the end of 2019. Therefore, to be included in the report the approaches had to be a) publicly available with accessible resources that could be accessed and assessed and b) available in English.

Engaging with experts

This report was also supported by thorough engagement with experts. Experts came from across OECD countries and from a range of organisations, including Ministries of Finance, Cabinet Offices, What Works Centres, Science Advisors, and other knowledge brokers with an interest in the principles and standards of evidence.

The expert group met twice over the course of the project and the Secretariat maintained an iterative engagement with the expert community throughout the project. The first meeting took place in October 2018 and aimed to integrate different approaches and country's experiences towards the good governance

of evidence, to shape the framework for the study. The second meeting took place in June 2019 and this meeting focused on the findings from the mapping exercise and offered significant opportunity for expert feedback. The expert group's feedback has been incorporated into this version of the report while experts were mobilised again twice in 2020.

Country coverage

Principles for the Good Governance of Evidence

Overall, the report mapped 29 approaches that covered one or more principles for ensuring the appropriate use of evidence. The sources come from a range of OECD countries: New Zealand, Japan, United Kingdom, and United States, as well as the European Union, the OECD, and from the International Science Council- INGSA. Thirteen approaches are from United Kingdom (UK), five from the European Union, four from United States, two from New Zealand and the OECD, and three from Malawi, Japan and INGSA (See Overview of the Principles for the use of Evidence across Jurisdictions). Whilst not part of the formal mapping of principles presented in Annex, the report also draws on a broader range of examples from a range of OECD jurisdictions including from Canada, Finland, France Norway, OECD and Portugal among others.

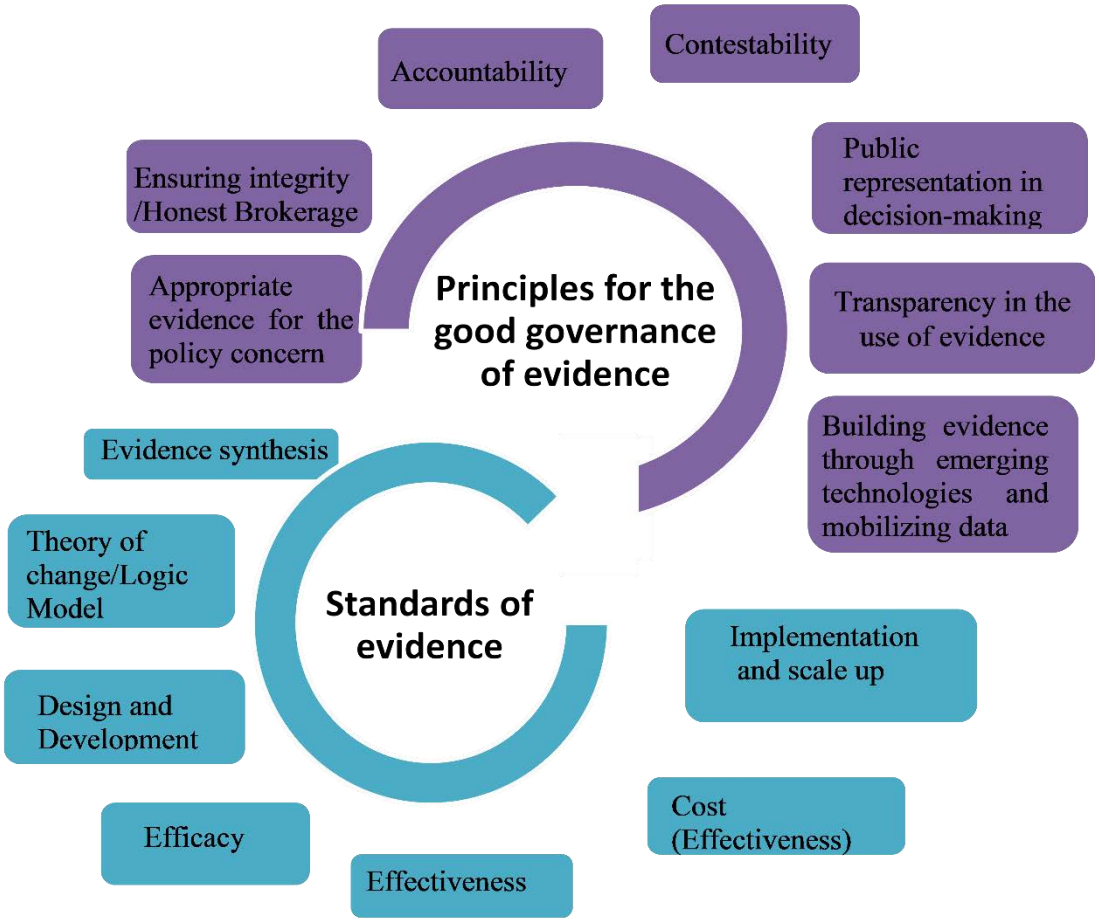
Standards of evidence

This report has mapped 50 standards of evidence, which come mainly from seven OECD countries: Australia, Canada, Germany, New Zealand, Spain, United Kingdom, and United States, as well as approaches from the European Union, and one with an international focus. Twenty-three approaches are from United States (USA), fourteen from United Kingdom (UK), and eight from Canada, Australia, Germany, New Zealand, and Spain. The European Union has four approaches drawn from existing EU agencies and the Joint Research Center of the European Commission (See Mapping of Existing Standards of Evidence across a range of Jurisdictions). Whilst not part of the formal mapping of principles presented in annex, the report also draws on a broader range of examples from a range of OECD jurisdictions including from Canada, Finland, France Norway, OECD and Portugal among others.

Consolidated overview of the principles and standards

This section presents a consolidated overview of the *Principles for the Good Governance of Evidence* as well as the related *Standards of evidence*. Figure 1.1 below presents the analytical framework while the following sections outline each of the principles and standards.

Figure 1.1. Mapping of the principles and standards for the good governance of evidence



Source: Research by the authors, expending and drawing on Parkhurst (2017[37]).

Principles for the good governance of evidence

Appropriate evidence for the policy concern

Evidence should be selected to address multiple political considerations; useful to achieve policy goals; be considerate of possible alternatives; and consider the local context. It should be useful for learning how to improve policy design, identifying both the positive and negative effects.

Ensuring integrity (honest brokerage)

Individuals and organisations providing evidence for policy-making need processes to ensure the integrity for such advice, managing conflicts of interest and ethical conduct to avoid policy capture and maintain trust in evidence and evidence-informed processes for policy-making. This notion of integrity includes both the processes used to select and analyse the evidence, but also the processes through which the advice is then provided to policy-making.

Accountability

Accountability in decision-making means that the agent setting the rules and shape of official evidence advisory systems used to inform policymaking should have a formal public mandate, and the final decision

authority for policies informed by evidence should lie with democratically representative and publicly accountable officials.

Contestability

Evidence must be open to critical questioning and appeal, which can include enabling challenges over decisions about which evidence to use. The challenge function can be organised within government, and through organisations such as chief economists, central review boards, or government science advisors.

Public representation in decision-making

Public engagement enables stakeholders and members of the public to bring their multiple competing values and concerns to be considered in the evidence utilisation process, even if not all concerns can be selected in the final policy decision.

Transparency in the use of evidence

Evidence should be clearly visible, understandable and open to public scrutiny. The public should be able to see how the evidence used to inform a decision was collected, analysed, and used in the decision-making process and for what purpose.

Building evidence through emerging technologies and mobilising data

The use of data and emerging technologies for the purpose of evidence-informed policy making needs to pay due attention to the increase in data governance instruments that address ethics, protection, privacy and consent, transparency and security; and to the development of relevant data standards that build evidence. It is important that the data and AI processes that build evidence are designed, generated, applied and used in ethical ways that respect the rights of individuals so that resulting evidence contributes to the fostering of public trust.

Standards of evidence

Standards concerning evidence synthesis.

Evidence syntheses provide a vital tool for policy makers and practitioners to find what works, how it works, and what might do harm based on thorough literature reviews that help to ensure good knowledge management of the existing and previous research. As with primary studies readers can and should appraise the quality of evidence synthesis.

Given the breadth of the published literature, including impact evaluations, RCTs, being published each year, knowledge management is essential as it becomes more difficult for policy makers and practitioners to keep abreast of the literature. Furthermore, policies should ideally be based on assessing the full body of evidence, not single studies, which may not provide a full picture of the effectiveness of a policy or programme.

Theory of change and logic underpinning an intervention: Should the intervention work?

A theory of change can be defined as a set of interrelated assumptions explaining how and why an intervention is likely to produce outcomes in the target population. A logic model sets out the conceptual connections between concepts in the theory of change to show what intervention, at what intensity, delivered to whom and at what intervals would likely produce specified short term, intermediate and long term outcomes. In some cases, a single theory of change might be difficult to identify due to multiple and

complex interactions, it might be difficult to identify a unique course of action and the underlying policy goals could be multiple and conflicting; this should not impede the activation of evidence processes.

Design and development of policies and programmes: Can it work?

Standards concerning the design and development of policies and programmes focus on evidence that tests the feasibility of delivering a policy in practice. At the design and development stage, analysts are often perform important work in testing theories of change and logic models, carrying out process evaluations and pre/post studies.

Efficacy of an intervention: Does it work?

Once an intervention has been identified as 'promising' in preliminary research, many standards of evidence emphasise the need for rigorous efficacy testing. Efficacy studies typically privilege internal validity, which pertains to inferences about whether the observed correlation between the intervention and outcomes reflect and underlie causal relationship. In order to maintain high internal validity, efficacy trials often test an intervention under 'ideal' circumstances. This can include a high degree of support from the intervention developer and strict eligibility criteria thus limiting the study to a single population of interest.

Effectiveness of interventions: Does it work in the real world?

Efficacy trials often tell us little about the impact of an intervention in 'real world' conditions, because the evaluation is often overseen by the developer of the policy or programme, with a carefully selected sample. What are the benefits or damages, independently from the policy goals? Therefore, standards of evidence often stipulate that a policy or programme demonstrates effectiveness, in studies where no more support is given than would be typical in 'real world' situations. This requires flexibility in evaluation design to address cultural, ethical and practice challenges. During policy implementation, evidence is useful to understand for whom its work and for whom it does not work. Therefore it is important to learn how to maximize benefits and minimize damages, also within a no policy change scenario.

Cost effectiveness of interventions: Is it worth it?

Positive impacts at a very high price may not be in the interests of governments and citizens. Using economic evidence is important to demonstrate value for money for public programmes in a context of continued fiscal constraints. Increased understanding of interventions that achieve impact at too high a price would enable decision-makers to make more efficient decisions.

Implementation and scale up of interventions

Knowledge of 'what works' – of which policies and programmes are effective, is necessary but not sufficient for obtaining outcomes for citizen. Increasingly, there is recognition that 'implementation matters'- that the quality and level of implementation of an intervention of policy is associated with outcomes for citizens. Thus, it is important to understand the features of policies and programmes, of the organisation or entity implementing them, along with the other factors that are related to adoption, implementation and sustainability of a policy or programme. It can also be important to ensure a systematic monitoring framework to contribute to the process of implementation and scaling up. This enables practical guidance to enable successful implementation.

References

Association of Children's Welfare Agencies (2018), *Using Evidence in Practice: The Role of Intermediaries*, http://www.acwa.asn.au/using-evidence-in-practice-the-role-of-intermediaries (accessed on 25 March 2020). [32]

Breckon, J. et al. (2019), *Evidence vs Democracy*, Alliance for Useful Evidence, http://www.alliance4usefulevidence.org/join (accessed on 22 March 2019). [45]

Burkhardt, J. et al. (2015), "An overview of evidence-based program registers (EBPRs) for behavioral health", *Evaluation and Program Planning*, Vol. 48, pp. 92-99, http://dx.doi.org/10.1016/J.EVALPROGPLAN.2014.09.006. [28]

Castro, F., M. Barrera and L. Holleran Steiker (2010), "Issues and Challenges in the Design of Culturally Adapted Evidence-Based Interventions", *Annual Review of Clinical Psychology*, Vol. 6/1, pp. 213-239, http://dx.doi.org/10.1146/annurev-clinpsy-033109-132032. [19]

Crato, N. and P. Paruolo (2019), *Data-driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design, Springer. ISBN 978-3-319-78461-8, https://doi.org/10.1007/978-3-319-78461-8*, Springer Nature, http://dx.doi.org/10.1007/978-3-319-78461-8 (accessed on July 19, 2019). [14]

D'Ancona, M. (2017), "Post truth", in *Post truth: the new war on truth and how to fight back*, Ebury Press, London. [6]

Duflo (2017), *The economist as plumber*, National Bureau of Economic Research, https://www.nber.org/papers/w23213. [20]

Ellen, M. et al. (2013), "What supports do health system organizations have in place to facilitate evidence-informed decision-making? a qualitative study", *Implementation Science*, Vol. 8/1, p. 84, http://dx.doi.org/10.1186/1748-5908-8-84. [30]

Exton, C. and M. Shinwell (2018), *Policy use of well-being metrics: Describing countries' experiences*, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/DOC(2018)7&docLanguage=En. [42]

Gaarder, M. and B. Briceño (2010), "Institutionalisation of government evaluation: balancing trade-offs", *Journal of Development Effectiveness*, Vol. 2/3, pp. 289-309, http://dx.doi.org/10.1080/19439342.2010.505027. [22]

Gough, D., S. Oliver and J. Thomas (2013), *Learning from Research: Systematic Reviews for Informing Policy Decisions: A Quick Guide.*. [16]

Gough, D. and H. White (2018), *Evidence standards and evidence claims in web based research portals*, Centre for Homelessness Impact, https://uploads-ssl.webflow.com/59f07e67422cdf0001904c14/5bfffe39daf9c956d0815519_CFHI_EVIDENCE_STANDARDS_REPORT_V14_WEB.pdf (accessed on 8 March 2019). [35]

Head, B. and J. Alford (2015), "Wicked Problems: Implications for Public Policy and Management", *Administration & Society*, Vol. 47/6, pp. 711-739, http://dx.doi.org/10.1177/0095399713481601. [1]

Langer, L., J. Tripney and D. Gough (2016), *The Science of Using Science Researching the Use of Research Evidence in Decision-Making e PPI EPPI-Centre*, http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3504 (accessed on 31 October 2019). [11]

Leadbeater, B. et al. (2018), "Ethical Challenges in Promoting the Implementation of Preventive Interventions: Report of the SPR Task Force", *Prevention Science*, pp. 1-13, http://dx.doi.org/10.1007/s11121-018-0912-7. [39]

Means, S. et al. (2015), "Comparing rating paradigms for evidence-based program registers in behavioral health: Evidentiary criteria and implications for assessing programs", *Evaluation and Program Planning*, Vol. 48, pp. 100-116, http://dx.doi.org/10.1016/J.EVALPROGPLAN.2014.09.007. [34]

Meyer, M. (2010), "The Rise of the Knowledge Broker", *Commentary Science Communication*, Vol. 32/1, pp. 118-127, http://dx.doi.org/10.1177/1075547009359797. [27]

Moore, J., B. Bumbarger and B. Cooper (2013), "Examining Adaptations of Evidence-Based Programs in Natural Contexts", *The Journal of Primary Prevention*, Vol. 34/3, pp. 147-161, http://dx.doi.org/10.1007/s10935-013-0303-6. [18]

OECD (2020), *Building Capacity for Evidence Informed Policy Making*, https://doi.org/10.1787/86331250-en.'. [23]

OECD (2020), "Digital Government in Chile – Improving Pubilc Service Design and Delivery", *OECD Digital Government Studies*, https://doi.org/10.1787/b94582e8-en. [15]

OECD (2020), "Ensuring data privacy as we battle COVID-19", *OECD Policy Responses to Coronavirus (COVID-19)*, http://www.oecd.org/coronavirus/policy-responses/ensuring-data-privacy-as-we-battle-covid-19-36c2f31e/. [10]

OECD (2020), *Improving Governance through Policy Evaluation: : Lessons From Country Experiences*, OECD Publishing, https://doi.org/10.1787/89b1577d-en. [21]

OECD (2019), *Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies-*, https://doi.org/10.1787/276aaca8-en. [8]

OECD (2019), *Government at a Glance*. [3]

OECD (2019), *Recommendation of the Council on Health Data Governance*. [9]

OECD (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/059814a7-en. [25]

OECD (2018), *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use*, OECD Publishing, https://doi.org/10.1787/9789264085169-en. [43]

OECD (2017), "Making policy evaluation work: The case of regional development policy", *OECD Science, Technology and Industry Policy Papers*, https://dx.doi.org/10.1787/c9bb055f-en. [40]

OECD (2017), "Trust and Public Policy", in *Trust and Public Policy: How Better Governance Can Help Rebuild Public Trust*, OECD Publishing, Paris. [2]

OECD (2016), *Open Government: The Global Context and the Way Forward*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264268104-en. [12]

OECD (2015), *Estonia and Finland: Fostering Strategic Capacity across Governments and Digital Services across Borders*, OECD Publishing, https://doi.org/10.1787/9789264229334-en. [24]

OECD (2014), "Evaluation of Industrial Policy: Methodological Issues and Policy Lessons", *OECD Science, Technology and Industry Policy Papers*, Vol. 16, https://doi.org/10.1787/5jz181jh0j5k-en. [41]

OECD-DAC (2009), "Guidelines for Project and Programme Evaluations", https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Projektabwicklung/Englisch/Guidelines_for_Project_and_Progamme_Evaluations.PDF (accessed on 20 September 2019). [13]

Olejniczak, K., E. Raimondo and T. Kupiec (2016), "Evaluation units as knowledge brokers: Testing and calibrating an innovative framework", *Evaluation*, Vol. 22/2, pp. 168-189, http://dx.doi.org/10.1177/1356389016638752. [26]

Oliver, K. et al. (2014), "A systematic review of barriers to and facilitators of the use of evidence by policymakers", *BMC Health Services Research*, Vol. 14/1, http://dx.doi.org/10.1186/1472-6963-14-2. [29]

Oliver, K. and W. Pearce (2017), "Three lessons from evidence-based medicine and policy: increase transparency, balance inputs and understand power", *Palgrave Communications*, Vol. 3/1, p. 43, http://dx.doi.org/10.1057/s41599-017-0045-9. [38]

Paris (ed.) (2018), *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use*, OECD Publishing, https://doi.org/10.1787/9789264085169-en. [46]

Parkhurst, J. (2017), *The politics of evidence : from evidence-based policy to the good governance of evidence*, Routledge, London, http://researchonline.lshtm.ac.uk/3298900/ (accessed on 23 November 2018). [37]

Parkhurst, J. and S. Abeysinghe (2016), "What Constitutes "Good" Evidence for Public Health and Social Policy-making? From Hierarchies to Appropriateness", *Social Epistemology*, Vol. 30/5-6, pp. 665-679, http://dx.doi.org/10.1080/02691728.2016.1172365. [36]

Prange-Gstöhl, H. (2016), "Eroding societal trust: a game-changer for EU policies and institutions?", *Innovation: The European Journal of Social Science Research*, Vol. 29/4, pp. 375-392, http://dx.doi.org/10.1080/13511610.2016.1166038. [4]

Proctor, E. et al. (2019), "Intermediary/purveyor organizations for evidence-based interventions in the US child mental health: Characteristics and implementation strategies", *Implementation Science*, Vol. 14/1, p. 3, http://dx.doi.org/10.1186/s13012-018-0845-3. [31]

Puttick, R. (2018), *Mapping the Standards of Evidence used in UK social policy Alliance for Useful Evidence About Nesta About The Alliance for Useful Evidence*, http://www.alliance4usefulevidence.org/join (accessed on 17 September 2018). [33]

Society for Prevention Research Standards of Evidence (2015), "Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation", *Society for Prevention Research Standards of Evidence*, Vol. 16/7, pp. 893-926. [44]

The Lancet (2020), "COVID-19: a stress test for trust in science", Vol. 396/10254, https://doi.org/10.1016/S0140-6736(20)31954-1. [5]

Torgerson, C. and D. Torgerson (2003), "The Design and Conduct of Randomised Controlled Trials in Education: Lessons from health care", *Oxford Review of Education*, Vol. 29/1, pp. 67-80, http://dx.doi.org/10.1080/03054980307434. [17]

Yarborough, M. (2014), *Taking steps to increase the trustworthiness of scientific research*, FASEB, http://dx.doi.org/10.1096/fj.13-246603. [7]

Notes

[1] See www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf

[2] As reflected in the work of the so called COCHRANE organisations, https://www.cochrane.org/evidence.

[3] See report on cost benefit analysis by the OECD Environment directorate.

2 Principles for the good governance of evidence

Deliberation of technical features of evidence are necessary but not sufficient for an evidence-informed approach to policy-making to deliver results. Policy and practice decisions actors should consider broader considerations, such as ethics, values, and political considerations to think critically about what evidence is needed in a particular context. This chapter presents existing principles for the good governance of evidence, including: appropriate evidence for the policy concern; ensuring integrity (honest brokerage); accountability; contestability; public representation in decision-making; transparency in the use of evidence; and building evidence through emerging technologies and mobilising data. Each subsection offers details of why these principles are important and a summary of the existing approaches.

Appropriate evidence for the policy concern

Evidence should be: selected to address multiple political considerations; useful to achieve policy goals; considerate of possible alternatives; understanding of the local context. It should be useful for learning how to improve policy design, identifying both the positive and negative effects.

Why is it important?

The notion of appropriateness is intended to supplement considerations of evidentiary quality (discussed in the standards chapter). The quality of the evidence remains critical, but the appropriate way to judge the quality of the evidence should be chosen only after consideration of which evidence is most useful to the policy concern (Parkhurst, 2017[1]). Ensuring that evidence is appropriate to the policy question at hand can also create a virtuous circle, creating increased demand and use of policy-relevant evidence.

Summary of the mapping of existing approaches

Approaches to the good governance of evidence consider several components to ensuring the availability of appropriate evidence, notably that evidence addresses multiple policy concerns, that evidence is constructed in ways that best service policy goals and that evidence is applicable to the local context. These can be considered as three interrelated concepts that ensure evidence is appropriate to the policy question and context (Figure 2.1).

Figure 2.1. Appropriate evidence for the policy context



Source: Adapted from Parkhurst (2017[1]).

Evidence that addresses the multiple policy concerns at stake

Some approaches emphasise the importance of ensuring that evidence addresses the multiple relevant social and policy concerns at stake (Parkhurst and Abeysinghe, 2016[2]) (Parkhurst, 2017[1]). Governments engage in a range of activities to improve the appropriateness of evidence generated in policy. In an analysis of initiatives in the UK and the USA, Shaxson (2019[3]) identifies the importance of strategies, policies and plans at different level of government which set out the priorities for evidence collection. Shaxson recommends that these:

- Cover all types of evidence.

- Demonstrate relevance to current and future policy directions and risks.
- Are open and transparent to encourage public engagement around evidence.
- Are updated, via broad-based engagement, on a regular basis.

OECD's 'New Approaches to Economic Challenges' has carried out work on complexity and policy-making (Love and Stockdale-Otárola, 2017[4]). This work recognises that policy makers must pay attention to the interlinkages between policy areas and related objectives, whilst improving evidence on the simultaneous movement of various targets and policy levers. Simultaneously, scientists and researchers must provide policy makers with evidence that makes policy makers aware of the complexity of the systems they are dealing with, whilst avoiding the sentiment that the systems are so complex that no one can possibly understand or influence them.

Evidence that best informs policy goals

Many approaches also emphasise the importance of ensuring that evidence should be sourced, created and analysed in ways that help decision-makers reach policy goals. Parkhurst (2017[1]) argues that what constitutes 'good' evidence for policy-making needs to be reframed as a question of policy appropriateness. This avoids over-simplistic applications of evidence hierarchies for questions they are not designed to address, and to help reconsider which evidence is most important to inform policy decisions.

The US Office for Management and Budget (2017[5]) highlights the importance of 'relevance' meaning that evidence must actually be used to inform decision- making. This means that findings should be presented in a format that is easy to interpret and apply. Ensuring relevant evidence can be facilitated by strong partnerships and collaborations among evidence producers, policy makers and service providers. The importance of early engagement between evidence experts and partner organisations in framing appropriate and relevant research questions is also highlighted by the UK Government Office for Science (2010[6]).

A range of strategies for ensuring that the information is easy for policy makers to interpret and apply comes from the fields of knowledge translation and knowledge brokerage. This includes a range of strategies such as decision making toolkits, guidelines and recommendations designed to facilitate solutions to particular policy problems (Langeveld, Stronks and Harting, 2016[7]). For example, many of the organisations reviewed in chapter 3 make use of standards of evidence to present evidence in an attractive and understandable way. For example, the Health Evidence (Canada) (2018[8]) offers the *Quality assessment tool* to help public health service to identify the best evidence available. The tool contains a questionnaire/checklist to review the studies, a dictionary for definition of terms and instructions for completing each section, and an overall assessment of the methodological quality of the review.

Evidence that is applicable to the local policy context

The applicability or transferability of evidence to a local context receives attention from a large variety of sources. Most stress the fact that 'context matters' so that even the most carefully studied and well-evidenced policies may not be successful in every context (Wandersman et al., 2000[9]; Leviton and Trujillo, 2017[10]). The case of Nurse Family Partnership illustrates the importance of context, with a heterogeneous pattern of results across different implementation contexts.

Box 2.1. Context Matters: Transporting Nurse Family Partnership to different contexts

Nurse Family Partnership is a licensed intensive home-visiting intervention developed in the United States and introduced into practice in England as *Family Nurse Partnership.*

The intervention has been shown to be effective in numerous RCTs in the USA (Eckenrode et al., 2010[11]; Olds et al., 2003[12]) and has also shown positive results in trial in the Netherlands (Mejdoubi and Midwifery, 2014[13]). However, when transferred to England it showed disappointing results with no additional short-term benefits to primary outcomes over and above those associated with usually provided health and social care (Robling et al., 2016[14]).

Why might this be the case? In the Netherlands, the intervention was modified so that it was focused on high-risk families. In the United Kingdom context, usual care already includes universal provision in the form of the *Healthy Child Programme*. It has been recommended that the Family Nurse Partnership be refocused in England to focus on higher risk families and by incorporating other activities to improve the effectiveness of the intervention (Barlow et al., 2016[15])

Source: adapted from (OECD, 2020[16]), Delivering evidence based services for all vulnerable families: A review of main policy issues.

These issues of applicability or transferability of evidence are closely related to themes discussed in section on implementation and scale up of interventions. Several clearinghouses discussed in that section provide resources that enable an assessment of evidence produced in one context and its applicability to a different context. For example the California Evidence-Based Clearinghouse for Child Welfare has provided a practical guide that enables organisations to evaluate their needs, make decisions about which new interventions might be suitable and plan for implementation and sustainability (2018[17]). The Norwegian Institute of Public Health has recently developed the TRANSFER approach to support the assessment of systematic review findings to the context of interest in the review, based on comprehensive list of factors that influence transferability of evidence (Munthe-Kaas, Nøkleby and Nguyen, 2019[18]).

Other important tools also exist to help decision makers apply evidence in context. (The GRADE Working Group, 2015[19]; Lavis et al., 2010[20]; Moberg, Alonso-Coello and Oxman, 2015[21]). One specific approach is the *Conclusions on Minimum Quality Standards in Drugs Demand Reduction form the Council of the European Union* (2015[22]), where they suggest appropriate evidence-based treatment should be tailored to the characteristics and needs of service users and is respectful of the individual's dignity, responsibility and preparedness to change.

Key questions

- Does the evidence selected address the multiple policy concerns at stake?
- Are there plans, which set out government priorities for data collection, to generate appropriate evidence, either through administrative or survey data?
- Does the evidence selected best serve the policy goals?
- Does the evidence collected help to determine a range of policy options including alternatives and the most appropriate option(s)?
- Is the evidence collected applicable in the local policy context?

Ensuring integrity (honest brokerage)

Individuals and organisations providing evidence for policy-making need processes to ensure the integrity for such advice, managing conflicts of interest and ethical conduct to avoid policy capture and maintain trust in evidence-informed processes for policy-making. This notion of integrity includes both the processes used to select and analyse the evidence, but also the processes through which the advice is then provided to policy-making.

Why is it important?

Ensuring the integrity of the interface between evidence production and decision making is critical to avoid the potential manipulation of evidence. Therefore, knowledge brokers need to develop a credible, legitimate and respected position in the eyes of both decision makers and evidence producers (Rantala et al., 2017[23]).

The term “Honest broker” refers to producers of evidence facilitating complex political decisions from a neutral position, aiming to understand and expand the scope of choice available instead of advocating a certain policy (Wilsdon, Saner and Gluckman, 2018[24]; Gluckman, 2014[25]) (Rantala et al., 2017[23]).

Ensuring integrity is particularly important in the 21st century as an increasing number of organisations start using emerging technologies to help them make better informed decisions. In the case of Artificial Intelligence (AI) systems, there is a heavy reliance on training data to develop the solutions, which may reduce the space for human decision making. Handling the preparation of data with integrity is critical to achieving results that are more accurate and to create trust in the application of AI. The OECD Principles on Artificial Intelligence promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values. These principles complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct (OECD, 2019[26]).

Summary of the mapping of existing approaches

Approaches to the good governance of evidence consider several facets of integrity, notably honest brokerage, the importance of humility of advice, ethical conduct and managing conflicts of interest.

Honest brokerage

Many jurisdictions have guidance concerning appropriate conduct for public officials, which include material relevant to integrity and honest brokerage. For example, the UK’s Committee on Standards in

Public Life (2019[27]) stipulates that public officials must avoid placing themselves under any obligation to people or organisations that might try inappropriately to influence them in their work. It is also important that those providing evidence for policy-making do so in a way that protects their independence from political interference (Gluckman, 2014[25]).

Other approaches consider the integrity specifically in relation to those providing evidence to decision makers. A case study in Finland identified four key roles and thirteen critical attributes in managing the challenges of honest brokerage in the political water supply decision-making process in Finland (Table 2.1).

Table 2.1. Honest brokerage roles and their attributes in Finland

Honest Brokerage	Rationale	Specific attributes of the brokerage roles
Knowledge provisioning	Need to provide clear, objective, neutral, and timely information	Information provider: to provide timely information and decision support to stakeholders during the process Interdisciplinary: to provide holistic picture about complex issues from various disciplines Clear communicator: to use clear language, illustrative methods, and concrete policy paths in making uncertainties and risks tangible Neutral: to be humble and objective by avoiding conveying researcher's own values and unintentionally promoting an agenda preferred by the broker
Trust and relationships	Need to build trust and relationships between brokers and stakeholders	Empathetic: to listen to all parties patiently and show understanding for their concerns and values Trustworthy: to treat confidential information ethically and to encourage transparency Reciprocal: to attract busy politicians to discuss and exchange ideas with the broker by providing useful information in return
Leadership	Need to keep the process on track and drive issues forward	Facilitative: to make sure the policy discussion evolves by providing clear answers for all open questions and queries Courageous: to expose broker's own credibility to open criticism by leading discussions and correcting misunderstandings
Intermediary	Need to balance and mediate between various positions	Mediative: to sketch alternative solutions and search for mutual gains Deliberative: to provide opportunities for stakeholders and the public to engage in dialogue and joint knowledge construction in order to balance the prevailing interest-based positions Detective: to be cognizant about hidden motives, political games, and interpersonal ties behind advocacy strategies Adaptive: to be sensitive and ready to

Source: Adapted from (Rantala et al., 2017[23]).

Humility

The need for humility in the provision of advice is recognised as a critical component of the role of honest brokers. Humility is defined as the striving for an accurate view of oneself and one's role, combined with an appropriate respect for both the power and limitations of science and technology. The International Network for Government Science Advice (INGSA) provides guidelines on three aspects of humility (Wilsdon, Saner and Gluckman, 2018[24]):

- Humility of competence respects the limits of knowledge and understanding. It requires those engaged in scientific advice to be open to new ways of knowing and the need to learn from others, with different disciplines, traditions, ages, genders, or geopolitical perspective.
- Humility of motive respects that advisers have their own history, culture, and political motives. Bias and conflict of interest are inescapable.

- Humility of role respects that a division of labour in decision-making is legitimate and that the process is non-linear. Scientific and technical input is often essential but rarely sufficient for complex, high-stakes decisions. Policy-making always involves considerations beyond the evidence.

Ethical conduct

Ensuring that there are robust procedures for maintaining the integrity of advice provided to decision makers, including ethical conduct and managing conflicts of interest is a further critical aspect of 'honest brokerage'.

Many of the approaches consider how ethical values and safeguarding can be assured. The criteria to select best practices adopted by the European Commission's Directorate General for Health and Food Safety covers such ethical aspects. These include ensuring that conflicts of interests are clearly stated along with measures taken to address them. In the US, the Society for Prevention Research has recently extended its work to consider ethical challenges inherent in the promotion of evidence-based interventions (see Box 2.2). The OECD itself has general recommendations on 'Managing Conflict of Interest in the Public Service' (OECD, 2003[28]) and 'Principles for Managing Ethics in the Public Service' (OECD, 2000[29]). Ethical conduct principles also extend to relationships between policymakers and providers of evidence, as well as effective communications to the public (Mair et al., 2019[30]). This can help limit undue pressure of policymakers on research providers that might compromise the quality of the evidence and distort trust in evidence-informed policy-making processes.

Box 2.2. Ensuring the ethical actions of researchers providing advice on policy implementation

The ethical challenges involved in the implementation of evidence-based interventions has been explored by the Society for Prevention Research in the US. Seven value statements were developed to promote the ethical conduct of implementation scientists, who should:

- Be guided fundamentally by intent to maximize benefits and prevent harm to both individuals and public wellbeing and welfare.
- Respect the rights of those whose lives they hope to improve and empower them to make decisions concerning issues that affect them.
- Maintain high standards of transparency in representing themselves to stakeholders and in disseminating scientific findings related to evidence-based practices.
- Provide accurate and complete information about the generalizability of available evidence, available choices, and costs of evidence-based policies to enhance the capacity of communities to make informed decisions regarding their adoption or scale-up of preventive intervention.
- Disclose financial and professional conflict of interests or limitations of expertise when presenting the scientific findings to stakeholders that affect programme adoption, dissemination, and implementation strategies.
- Promote ongoing communication, transparency, accountability, reliability, and reciprocity in relationships with all partners across the phases of implementation of preventive interventions.
- Anticipate and respect diverse values, beliefs, and cultures of the community or population engaged in implementing an intervention.

Source: Adapted from Leadbeater (2018[31]).

Many approaches focus on ensuring that policy evaluations and other policy related research are conducted to safeguard the dignity, rights, safety and privacy of participants (e.g. OMB guidance). Given the increase in use of administrative data for evaluation purposes, issues of informed consent are raised when information provided by citizens is not being used for the original purposes to which the citizens consented. For example, some secondary analyses use individuals' identifiers to link public data to existing longitudinal datasets. When consent was given for the original purposes for collecting the data, but not for the use of the linked data, considerations of the need to balance potential benefits of the research with respect for individuals' autonomy and self-determination are of concern (Leadbeater et al., 2018[31]).

Conflicts of interest

Conflicts of interest is addressed in existing principles both as a general feature of those involved in public service and as more specifically relating to those involved in providing evidence for decision making. A key part of standards of public life is that officials do not act or take decisions in such a way as to gain financial or other material benefits. The UK Committee standards of public life stipulate that officials must declare and resolve any interests and relationships that could give rise to conflicts of interest. In France, a High Authority for the Transparency of Public Life, ensures that officials carrying high level political responsibilities cannot be placed in a situation of conflicts of interest with a wide range of application.[1] There is also a literature on managing conflicts of interest in the evidence synthesis process (Healy, 2019[32]; Sturmberg, 2019[33]) and many clearinghouses and evidence repositories who conduct evidence synthesis have policies and procedures for managing conflicts of interest. The OECD's approach is described in Box 2.3.

Box 2.3. OECD Guidelines for managing conflict of interest in the public service

1. Identify relevant conflict of interest situations
2. Demonstrate leadership commitment
   a. Organisations should take responsibility for the effective application of their Conflict of Interest policy.
3. Create a partnership with employees: awareness, anticipation and prevention
   a. Ensure wide publication and understanding of the Conflict of Interest policy.
   b. Review 'at-risk' areas for potential conflict of interest situations.
   c. Identify preventive measures that deal with emergent conflict situations.
   d. Develop an open organisational culture where dealing with conflict of interest matters can be freely raised and discussed.
4. Enforce the Conflict of Interest policy
   a. Provide procedures for establishing a conflict of interest offence, and proportional consequences for non-compliance with Conflict of Interest policy including disciplinary sanctions.
   b. Develop monitoring mechanisms to detect breaches of policy and take into account any gain or benefit that resulted from the conflict.
   c. Co-ordinate prevention and enforcement measures and integrate them into a coherent institutional framework.
5. Initiate a new partnership with the business and non-profit sectors

a. Create partnerships for integrity with the business and non-profit sectors by involving them in the elaboration and implementation of the Conflict of Interest policy for public officials
b. Anticipate potential conflict of interest situations when public organisations invite the involvement of persons representing businesses and the non-profit sector.
c. Raise awareness of the Conflict of Interest policy when dealing with other sectors, and include safeguards against potential conflict of interest situations when co-operating with the business and non-profit sectors.

Source: Adapted from OECD (2003[28]), Recommendation of the council on guidelines for managing conflict of interest in the public service.

Addressing the risk of capture and bias in the policy process

There is also a need to address the potential for bias as external voices often try to intervene in the policy-making process to preserve or promote specific interests. OECD's 'Guidelines for Managing Conflict of Interest in the Public Service' respond to a growing demand to ensure integrity and transparency in the public sector (OECD, 2003[28]). The primary aim of the Guidelines is to help countries, at central government level, consider Conflict of Interest policy and practice relating to public officials. Demands for transparency in public decision-making have also led to concerns over lobbying practices. In 2009, the OECD reviewed the data and experiences of government regulation, legislation and self-regulation, leading to '10 Principles for Transparency and Integrity in Lobbying'. These issues are also pertinent to evidence-informed policy-making and capacity building. Commercialisation of capacity building activities can also create pressure to overstate the benefits, leading to erosion in confidence if expectations are not met (Leadbeater et al., 2018[31]).

Moreover, the rapid spread across governments of emerging technologies such as AI systems in decision-making processes raises pressing issues in acknowledging and mitigating bias in algorithms to ensure integrity and maintain trust. To illustrate this point, the federal government of Canada explored the responsible use of AI in government, establishing an Algorithmic Impact Assessment (AIA) tool in order to assist designers evaluate the suitability of their AI solutions and creating a set of guidelines to complement it (Government of Canada, 2019[34]). This will be further discussed in Building evidence through emerging technologies and mobilising data.

Given that misperceptions affect the success of policy interventions, and eventual trust and buy-in of stakeholders and the public (Mair et al., 2019[30]), it is vital that perceived and alleged conflicts of interest are also proactively tackled and clarified for the public.

Key questions

- Are there provisions to ensure appropriate conduct of public officials involved in providing evidence for decision-making?
- Are procedures in place to ensure the impartiality of those providing evidence to decision makers?
- Are robust procedures in place for disclosing and managing potential perceived and real conflicts of interests of those providing evidence to decision makers?
- Are experts providing evidence able to provide objective, clear, neutral and timely information?
- Are procedures in place to ensure that evidence experts respect the limits of their own knowledge and understanding and communicate the limitations of evidence?

Accountability

Accountability in decision-making means that the agent setting the rules and shape of official evidence advisory systems used to inform policymaking should have a formal public mandate, and the final decision authority for policies informed by evidence should lie with democratically representative and publicly accountable officials.

Why is it important?

Accountability reflects the extent to which advisers and public officials are answerable for their actions. This also matters from the perspective of open and inclusive government, as reflected in the OECD Recommendation of the Council on Open Government (OECD, 2017[35]).

Summary of Mapping Approaches

Approaches to accountability in decision-making have focused on creating clear distinctions between the roles of advisors and public officials, and ensuring elected officials have leadership over the evidence advisory process.

Clear Roles of Policy Makers and Advisors in the Evidentiary Advisory Process

Creating clear distinctions between the roles of policy makers and advisors within the policy-making process ensures that the final decision rests with those who are democratically elected and whose actions are accountable to the public. As governments have a democratic mandate, public officials must take into account “a wide range of factors and recognise that science is only part of the evidence government must consider in developing policy” (Government Office for Science, 2010[36]).

The role of advisors is to present evidence to policy makers whereas the role of policy makers and elected officials is to define policy and choose between options with different trade-offs (Gluckman, 2014[25]). To ensure accountability, The Nuffield Council for Bioethics states that “there should be explicit acceptance and acknowledgement of where responsibility for governance lies and how it might legitimately and democratically be influenced” (Nuffield Council on Bioethics, 2012[37]).

In many countries governments will appoint ministerial advisors to “increase the responsiveness of government and help address strategic challenges faced by government leaders” (OECD, 2011[38]). Ministerial advisors will gather evidence and co-ordinate with stakeholders within and outside government,

and will advise the elected officials based on the evidence. For more information on the ministerial advising position and the associated challenges, see Box 2.4.

Box 2.4. Creating a Governance Framework for Ministerial Advisors

The OECD conducted a survey of 27 countries on their ministerial advising systems and found that there remains considerable room for improvement in developing a clear governance framework for ministerial advisors. The survey findings point to the following as potential avenues of reform:

- Clearly stating advisors' job descriptions, delineating their power and functions as distinct from those of senior public servants and setting the boundaries they may not overstep;
- Setting clear standards of integrity specifically for ministerial advisors and ensuring that they disclose their private interests pro-actively to identify and prevent conflict of interest;
- Increasing transparency not only in respect to the number of advisors but of their overall cost, profiles and competencies;
- Clarifying the accountability structure governing ministerial advisors.

Source: Adapted from OECD (2011[38]), Ministerial Advisors: Role, Influence and Management

Leadership in the Evidentiary Advisory Process

In the evidentiary advisory process, the democratically elected officials should have leadership over the structure and process. The National Institute for Health and Care Excellence highlights the importance of effective leadership for accountability in their guidelines. This means, having 'visible, proactive and inspiring leadership' (National Institute for Health and Care Excellence, 2013[39]). The policy makers and elected officials should build a sense of trust while maintaining clear boundaries between those who supply evidence and those who make decisions based on the evidence supplied. It should be clear to the public and to all of those involved in the policy-making process that policy makers and elected officials have authority over the policy-making process and the final policy decisions.

The process through which evidence is provided to policy makers should be created by those who are democratically elected (Shaxson, 2019[3]). The process should be clearly defined with a clear starting and ending point for evidence provision and policy advising. After which, policy makers deliberate and make policy decisions, for which they are accountable to the public.

The ministerial advisory system should be managed by government officials. Government officials can do so through playing a role in appointing advisors, and providing guidance on the ministerial advisors' functions and terms of employment. Government officials can also set clear standards of integrity for ministerial advisors (OECD, 2011[38]). In enacting these measures, government officials can ensure that the advisory system that provides evidence is structured and managed by democratically elected officials.

Aside from individual ministerial advisors, many governments are also given advice from advisory bodies and have sought to create clear roles and a separation between government and advisory bodies to ensure the independence of policy advice. In a survey of 15 countries conducted by the OECD, it was found that many countries have put measures in place to regulate the division of roles between the advisory bodies and elected officials, including regulations on conflict of interest, ethics and corruption (OECD, 2017[40]).

While the separation of responsibilities and independence of advisory bodies are crucial, the right balance must be reached between the extremes of isolation and dependence. Co-creation, iteration and building effective operational relationships are important for understanding the needs of policymaking by evidence providers, and helps better respond to complexity of the tasks (Mair et al., 2019[30]) (OECD, 2017[40]).

Key questions

- Are there clear policies and procedures in place clarifying the responsibilities of those providing evidence to the policy-making process?
- Are the roles of advisors and policy makers clearly defined and understood?
- Are the structure and processes involved in the evidence advisory processes obliged to give an account and report on their activity?

Contestability

Evidence must be open to critical questioning and appeal, which can include enabling challenges over decisions about which evidence to use.

Why is it important?

The need for contestability – in the form of appeals processes and opportunity for public debate is premised on the principle that having data, evidence and arguments open to questioning and challenge is a key element of the scientific process (Hawkins and Parkhurst, 2016[41]) and a critical part of good governance to ensure trustworthy participatory processes. Furthermore, evidence and its policy implications are typically uncertain and open to debate and disagreement, suggesting that a range of voices and views need to be heard in the process of moving from the evidence base to tangible policy solutions.

Summary of the mapping of existing approaches

Processes for when evidence is uncertain and contested.

Most of the approaches dealing with contestability, highlight the importance of ensuring an openness to critical questioning, including both the evidence that has been used and also the processes through which it has informed decision-making (Shaxson, 2019[3]).

Many of the evidence based on clearinghouses and What Works centres, would have some form of stakeholder input to provide a critical contribution and scrutiny about the evidence that underpins an assessment. This can be formalised in terms of entities such as the 'Stakeholder Advisory Review Groups' that exist for the reviews carried out by the Campbell Collaboration (2019[42]). The identification and involvement of a *Stakeholder Advisory Review Group* can have a number of positive benefits, which include assisting the review team by providing critical challenges about the decisions in the systematic review process, including the inclusion and exclusion criteria in order to maintain relevance to the stakeholder audience.

The European Union has a number of approaches to using stakeholder engagement to tackle the contestability and uncertainty of evidence used in decision-making[2]. The Commission's Principles and guidelines on the collection and use of expertise (European Commission, 2002[43]) note that many policy

decisions must be made on contentious issues in the face of uncertainty. This can lead to decision makers being confronted with conflicting expert opinions, coming variously from within the academic world, emanating from different starting assumptions, and different objectives. The Commission provides several practical questions to help Commission departments' design arrangements appropriate to circumstances of specific cases. These practical questions include the following:

- Is the advice properly substantiated and documented?
- Should the advice be submitted to other persons for comments or validation?
- Will this be a scientific peer review?
- Is it appropriate to submit the advice to scrutiny and comments from a wider circle of experts and interested parties?
- Have arrangements been put in place to record and assess unsolicited comments once advice has been published?
- Does the issue require interaction between the experts, interested parties and policy-makers?

The European Food Safety Authority (European Food Safety Authority, 2014[44]) has produced its own detailed guidance on 'expert knowledge elicitation (see Box 2.5).

Box 2.5. The European Food Safety Authority's Guidance on Expert Knowledge Elicitation in food and Feed Safety Risk Assessment

In the EU, risk assessment in food and feed safety is the responsibility of the European Food Safety Authority (EFSA). In 2012, a working group was established to develop guidance on expert knowledge elicitation appropriate to EFSA's remit.

In an ideal world, quantitative risk models should be informed by systematically reviewed scientific evidence. In practice, empirical evidence is often limited and in such cases, it is necessary to turn to expert judgement. However, psychological research has shown that unaided expert judgement of the quantities required for risk modelling - and particularly the uncertainty associated with such judgements - is often biased, thus limiting its value.

To address this issue, methods have been developed for eliciting knowledge from experts in as unbiased a manner as possible. EFSA's guidance first presents expert knowledge elicitation as a process beginning with defining the risk assessment problem, moving through preparation for elicitation (e.g. selecting the experts and the method to be used) and the elicitation itself, culminating in documentation.

Three detailed protocols for expert knowledge elicitation are provided - that can be applied to real-life questions in food and feed safety - and the pros and cons of each of these protocols are examined.

- The Sheffield Protocol with group interaction of experts (behavioural aggregation)
- The Cooke Protocol with use of seed questions for the calibration of experts (mathematical aggregation)
- The Delphi protocol on written individual expert elicitation (i.e. remote) with feedback loops (mixed behavioural and mathematical aggregation).

The guidance also contains principles for overcoming the major challenges to expert knowledge elicitation: framing the question; selecting the experts; eliciting uncertainty; aggregating the results of multiple experts; and documenting the process.

Source: Adapted from European Food Safety Authority (European Food Safety Authority, 2014[44]), Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment.

A further set of issues about the contestability of evidence concerns the erosion of the authority of science in the context of social media use, the increased diffusion of fake news, and an increasingly polarised values based political debate (Allcott and Gentzkow, 2017[45]). Social media outlets such as Facebook and Twitter have a fundamentally different structure to earlier media technologies. It is now possible for content to be relayed among users without significant third party verification, fact checking or relevant editorial process (Allcott and Gentzkow, 2017[45]). This unverified material makes it both easier to diffuse 'fake news' as well as cast doubts on legitimate news. This news imposes both private and societal costs by making it more difficult for citizens to infer the true state of the world. This is compounded by the so called 'echo chamber effect', whereby online communication leads to selective exposure and ideological segregation (Barberá et al., 2015[46]). Initiatives such as the International Fact Checking Network are using algorithms to identify fake content and validating information sources - two approaches that have been developed to address these issues (Figueira and Oliveira, 2017[47]).

Other initiatives work on bridging the gap between the cultures of the various stakeholders and providing insight into the use of evidence in political behaviour. For example, France Stratégie, who questions the difficulties involved in the production, distribution, and circulation of 'correct' information from the point of view of elected representatives, researchers, administrators, community leaders, journalists, and think tanks (France Stratégie, 2019[48]), or the French Council of State which recommends to use evaluations as part of democratic deliberations (Conseil d'Etat, 2020[49]). Similarly, the European Commission also examines how evidence-informed policymaking processes influence political behaviour (Mair et al., 2019[30]).

Processes for stakeholder engagement in participatory evidence assessments and recommendations

Many of the evidence-based clearinghouses conducting health technology assessments and assessments of social services have procedures for relevant stakeholders to request a reassessment of the judgment reached about an intervention. For example, the US Clearing House HomVEE has a detailed process that enables stakeholders to request a reconsideration of 'evidence determinations' (see Box 2.6). The UK Early Intervention Foundation has a similar process, allowing stakeholders to request a reassessment of an evidence rating if they feel that the standards of evidence have not been properly applied (Early Intervention Foundation, 2017[50]).

The notion of contestability is also relevant to evaluations carried out by other elements of the government ecosystem. The International Standards of Supreme Audit Institutions (INTOSAI, 2016[51]) aim to help Supreme Audit Institutions and other entities in charge of evaluation to analyse neutrally and independently the utility of a public policy. This guidance recommends a 'Clearing Stage' in which a draft report containing the results and the analyses are shared and discussed with stakeholders of the policy evaluated. The purpose of this exercise is to ensure that the provisional analyses and conclusions are accurate and that the opinions of stakeholders can be gathered and integrated where appropriate, thus ensuring accountability (INTOSAI, 2016[51]). This form of stakeholder engagement is also a critical part of Regulatory Impact Assessment, in order to elicit feedback from citizens and other affected parties so that regulatory proposals can be improved and broadly accepted by society (OECD, 2019[52]; OECD, 2018[53]).

Box 2.6. HomVEE's 'Requests for Reconsideration of Evidence Determinations'

In the event that a US State, a researcher, an intervention developer or other stakeholder believes that, the Department for Health and Human Services (HHS) criteria for evidence of effectiveness for a particular intervention, contains an error these can be raised with the review team.

A 'request for reconsideration of the evidence based determination' can be requested based on misapplication of the HHS criteria, or missing information, or errors on the HomVEE website.

The HHS review team considers the request and, if approval is granted a re-review team composed of members external to the original team conducts a the new independent review. The re-review team will provide assurance that they do not have any actual or perceived conflicts of interest. The re-review team is certified and trained in the HomVEE standards.

The re-review team utilizes the original empirical articles (see the model reports), any information submitted by the individual raising the concern, the original review team's reports, and make any needed queries to the original team. The goal will be to issue a final decision as to whether the standards were accurately applied or not within 60 days of the submission of the request for review. Following the decision, the requester will be notified of the decision and, if necessary, any adjustments to the model reports or HomVEE website will be made.

Source: Adapted from HomVEE (2018[54]), Assessing Evidence of Effectiveness.

In the health area in the UK, the National Institute for Clinical Excellence (NICE,) has developed a few principles that guide the development of NICE guidelines and standards. One of NICE's principles stipulates that people interested in a topic area should be given the opportunity to comment on and influence NICE's recommendations (See more in Box 2.7). In the area of Sustainable Development, the French National Institute for Industrial Environment and Risks (INERIS/CDDEP) has also developed guidelines for engaging with stakeholders[3].

Box 2.7. Stakeholder engagement in NICE recommendations

NICE recommendations are based on complex considerations of the evidence by NICE committees, and it is important therefore that a wider group of stakeholders also have the opportunity to comment.

This wider consultation helps ensure the validity of the final recommendations. The principles of the NHS Constitution also require NICE to be accountable to the public, communities and patients that NICE serves, and to make decisions in a clear and transparent way.

NICE's guidance and standards are therefore developed using a process that takes into account the opinions and views of the people who will be affected by it. NICE consults openly with organisations that represent people using services, carers and the wider public as well as health and social care professionals, NHS organisations, industry, social care businesses and local government.

NICE's advisory committees consider and respond objectively to comments and, where appropriate, amend the recommendations.

Source: adapted from NICE (National Institute for Health and Care Excellence, 2013[39]).

Key questions

- Is the evidence used in decision making open to critical questioning, as well as the underlying process through which it was used?
- Are there opportunities for opening up processes through stakeholder engagement, to ensure participatory processes, including with experts, in policy areas where the evidence is contentious and uncertain?
- Are there procedures to ensure that knowledge coming from experts on contested issues is as unbiased as possible?
- Are there opportunities for stakeholders to comment on and influence in the recommendations produced based on an evidence?

Public representation in decision making

There should be public engagement that enables stakeholders and members of the public to bring their multiple competing values and concerns to be considered in the evidence utilisation process, even if not all concerns can be selected in the final policy decision.

Why is it important?

Government has long used engagement and participation to earn trust and overcome complexity. Traditionally engagement has focused on getting buy-in from communities for a policy. But engagements should not focus solely on buy-in, and 'managing' citizens and stakeholders and their expectations, looking to minimise opposition. Rather, public servants should see the public as a source of expertise, and that engaging with them (particularly through innovative engagement methodologies like public deliberation) can forge a partnership to overcome complexity. Finally, engagement helps in supporting countries strategies and initiatives towards Open Government (i.e. transparency, integrity, accountability and stakeholders' participation), in line with the OECD Recommendation of the Council on Open Government (OECD, 2017[35]), as a compound for good governance, democracy that promotes inclusiveness (OECD, 2016[55]).

Summary of the mapping of existing approaches

Guidelines and standards that discuss public deliberation focus on stakeholder and public consultations, and public inclusion in evidence synthesis.

Stakeholder and public consultations

Many of the approaches focus on the importance of including a variety of stakeholders in the policy-making process. The OECD encourages public deliberation that includes a wide range of stakeholders in public consultations to ensure that the policy "serves the public interest and is consistent with societal values" (OECD, 2017[56]). Guidelines published by Oxford University also emphasise including a larger range of stakeholders to increase public participation in the policy process as well as in the creation and use of evidence (Oliver and Pearce, 2017[57]). Such as these initiatives, other sources specify that consultations enable stakeholders to reach consensus on research questions and shape research agenda decisions; and it can be a factor for successful applications (Ferri et al., 2015[58]).

Stakeholder engagement can include a variety of groups in the public, with guidelines each emphasising the importance of different groups. The European Science organisation EuroScientist, highlights the need to include stakeholders from the private sector, civil society organisations and NGOs in public deliberations (EuroScientist, 2017[59]). (What Works Clearinghouse, 2020[60]) (What Works Clearinghouse, 2020[60]) Ferri et al (2015[58]) explore interventions' effectiveness together with a stakeholder's consultation to ensure a broader participation in the definition of research needs. In this in-depth analysis, the authors report international initiatives such as James Lind alliance (2019[61]), a non-profit initiative dedicated to stakeholders' involvement in research priority setting, which enables patients, carers and clinicians to reach consensus on research questions, and it explicitly excludes the pharmaceutical industry and non-clinicians researchers.

It is also important to ensure that the perspectives of the people receiving services are consulted, including marginalised groups, to understand how they feel about the policy and ways in which they feel policy will affect them (Bond, 2018[62]). The Society for Prevention Research also includes the need for consultations with communities, institutions and public agencies (Leadbeater et al., 2018[31]).

Public deliberation can allow politicians to understand what the public thinks on different issues and allows for the public to more thoroughly discuss their views and what they want compared to public feedback through polls or referendums (Chwalisz, 2017[63]). One example of a forum for public deliberation is mini-publics where a group of randomly selected individuals (usually 24 to 48 people) will come together and deliberate on an issue for a period of time (usually 2 to 4 months) and will try to come up with a solution (Chwalisz, 2017[63]). Another interesting example are the *Citizens' Assemblies* that have been put in place in France to identify ways to address climate change[4].

For more information on mini publics, see Box 2.8.

Box 2.8. Using Mini-Publics for Deliberation

In the public deliberation process, some governments will form mini-publics, where a group of citizens will come together and deliberate on a certain policy or issue. The organisation New Democracy has collected data on mini-publics in Australia through participant surveys and has received very positive feedback.

- In the deliberation process, the mini-public members establish their own agreed behavioural guidelines, set the criteria for evaluation, gather and test information, brainstorm solutions, prioritise the possibilities, agree on recommendations and account for minority opinions when consensus is not found, and collectively writing a report.
- Post-deliberation, many departing citizens speak of challenging but surprisingly respectful conversations, despite individual differences; deep exploration of issues, with a shared motivation to solve a problem; and an enhanced ability to think critically.
- In Anonymous feedback post-deliberation, participants usually say they would do it again and they want decision makers to make many more, similar opportunities available to their fellow citizens.

Source: Adapted from Carson (2017[64]), New Democracy Research and Development Note: Deliberation

There are many benefits to public deliberation; however, policy makers should be aware of and acknowledge the "potential for bias and vested interests" (EuroScientist, 2017[59]) . Public deliberation

entails listening and including the views of the public; policy makers must take into account the many diverse views but policy makers do not necessarily need to agree with or integrate the opinions of the public groups into the final policy. The appropriate choice of stakeholder engagement methods will depend, therefore, on the rigour and relevance of the knowledge already available about a specific issue and its context in the implementation stage (Oliver et al., 2018[65]).

Stakeholder Engagement in Evidence Synthesis

Evidence syntheses, such as systematic reviews, attempt to summarise the best available evidence on a specific issue, using transparent procedures to locate, evaluate and integrate the research findings. The technical aspects of carrying out evidence synthesis are described in chapter 3, but a further issue that has received attention is how to ensure appropriate stakeholder engagement in the process of creating evidence synthesis. Stakeholder engagement in the evidence synthesis process can ensure that research findings are more effectively put into practice (Pollock et al., 2018[66]). Stakeholder engagement can also ensure that the research has 'real-world' relevance and applicability. Nonetheless, other authors suggest that the appropriate methods for stakeholder engagement with knowledge production will depend on the clarity and consensus about core concepts shaping a research, and whether or not the purpose is to generate research findings that are generalizable beyond the context of the research setting (Oliver et al., 2018[65]).

Stakeholder engagement can be done through several different means. In a review that analysed 291 systematic reviews, the most common ways that stakeholders were involved in the systematic review process was through one-off face to face meetings with targeted participants, general meetings open to the public and online surveys (Pollock et al., 2018[66]). The systematic reviews varied in how they engaged stakeholders in the systematic review process. Some reviews included stakeholders throughout the review process or during different stages of the process, while other reviews only included stakeholders in interpreting the results after the evidence was synthesised (Pollock et al., 2018[66]). Overall, the evidence is still unclear as to what the best ways are for including stakeholder participation in evidence synthesis.

Key questions

- Are a variety of stakeholders given the opportunity to give input or feedback during the policy-making process?
- Are there mechanisms to ensure meaningful engagement, so that stakeholders can discuss their views, including deliberative techniques such as mini publics?
- When engaging with stakeholders, are policy makers aware of the potential biases and vested interests of the different stakeholders?
- Are there mechanisms in place to ensure that the evidence collected is presented in a format that is easy to interpret and apply?

Transparency in the use of evidence

Evidence should be clearly visible and open to public scrutiny. The public should be able to see how the evidence base informing a decision is identified and utilised and for what purpose.

Why is it important?

Transparency in the gathering and use of evidence is an essential part of the democratic process. Ensuring transparency can help to build trust with the public and experts in the policy-making process (McDonald, 2010[67]). Transparency also enables accountability and facilitates stakeholder participation (OECD, 2019[68]).

Summary of the mapping of existing approaches

Transparency is well covered in existing approaches to the use of evidence, which includes the transparency of decision making and the transparency in disclosing information.

Transparency of decision making

Many guidelines and principles align with Parkhurst's definition, emphasising that policy-makers should make information available to the public, in a form that is meaningful and understandable to the citizens and stakeholders concerned. In order to be transparent, governments need to disclose and make accessible to the public relevant government information and data through explicit open government data strategies and through securing access to information (OECD, 2016[55]).

Transparency in decision-making requires policy makers to explain and justify the process and purpose of evidence gathering. This includes disclosing how policy makers decide what evidence to use or not use in their policy creation (Ministry of Health). The Commission of the European Communities, in their report on principles and guidelines for evidence use in policy-making, stress that policy makers should be able to explain and justify their decisions on "the way issues are framed, experts are selected, and results handled" (Commission of the European Communities). This includes disclosing the framing and methodology of the evidence, assumptions and limitations, and any complexities or contentions within the findings of the evidence used in the decision-making process. As well, policy makers should disclose how they weigh the evidence and determine what evidence is useful based on the transparent presentation of the evidence. For more information on transparency in presenting evidence, see Box 2.9.

Box 2.9. Transparency in the Presentation of Evidence

The Royal Society of the UK, in their principles on the use of evidence in policy-making, outline key ways to ensure that those who are advising policy makers on research are transparent in their portrayal of evidence. This is important because evidence that is transparent in its sources and creation "is likely to be more credible, replicable and useful."

In presenting evidence, it should:

- Clearly describe the research question, methods, sources of evidence and quality assurance process,
- Communicate complexities and areas of contention,
- Acknowledge assumptions, limitations and uncertainties, including any evidence gaps,
- Declare personal, political and organisational interests and manage any conflicts.

Source: Adapted from The Royal Society (2018[69]), Evidence synthesis for policy a statement of principles

When policy makers are not the ones who selected evidence to consider, but instead are presented with synthesised evidence, policy makers should be transparent about the evidence synthesis process. Transparency in evidence synthesis ensures that it is made clear how the research question was formulated and how evidence was selected to be included in the synthesis. For more information on transparency in evidence synthesis, see Box 2.10.

Box 2.10. Transparency in Evidence Synthesis

The European Food Safety Authority created a list of principles for the synthesis of evidence and outlined how to ensure transparency in the synthesis process. Those who synthesize information should have a protocol that explains and defines the following:

- The review question and objective
- The criteria for study inclusion or exclusion
- The methods for searching research studies, methodological quality of the included studies, and synthesising the data from the included studies.

Source: Adapted from European Food Safety Authority (2010[70]), Application of systematic review methodology to food and feed safety assessments to support decision making.

Many of the guidelines emphasise the importance of policy makers being transparent in their reasoning for the policy decisions they make. The EU, in their Declaration on Ethics and Principles for Science and Society Policy-Making, state that transparency within the decision-making process includes policy makers disclosing all sources of input, including non-scientific, that were considered when making policy decisions (EuroScientist, 2017[59]). The policy-making process is complex and policy makers must consider many inputs including community desires, public opinion, budget, timeline, etc. The different inputs that were factors in the decision-making process and the way in which these factors were weighed should be disclosed to the public so that they are able to understand how a final decision was made.

Transparency in disclosing information

To be transparent governments need to take steps to ensure that information is available and accessible to the public. To ensure that the public can access the information, policy makers should use a variety of channels to disseminate the information. A number of guidelines propose ways for policy makers to disseminate information and OECD has been generally proactive in this regard. The OECD work on *Enabling the strategic use of data for productive, inclusive and trustworthy governance* explores how to make sure data and digital technologies (existing and emerging) are used to enact openness, such as the use of trends and patterns in order to mitigate emerging risks and respond to developing crises; and the use of data to understand problems, engage the public and provide access to insights for improving public services that meet user needs, while creating the conditions for robust, evidence-based policy making (van Ooijen, Ubaldi and Welby, 2019[71]).

The Commission of the European Communities recommends that policy makers are proactive in their communication and "constantly seek ways to better publicise and explain its use of expertise to interested parties and the public at large" (European Commission, 2002[43]). The US' Office for Management and Budget advise that information should be made easily accessible through posting information online (Office of Management and Budget, 2018[72]). The OECD's report on Scientific Advising found that many advisory boards tend to use the more traditional forms of communication such as printed reports or online editions.

However, to reach a broader audience, policy makers and advisory boards should also use sources that are more frequently read by the public including traditional media (newspaper and television), as well as social media (OECD, 2015[73]). Using these sources can facilitate a larger discussion around the policy and can include a broader section of society.

In creating transparency, policy makers need to ensure that the information they make available is understandable to the public. To make information more understandable, The Royal Society recommends that policy makers use plain language, avoid jargon, and present the information in a clear, concise and objective manner (The Royal Society, 2018[69]). This will ensure that the information is more accessible to the public, including individuals who are not scientific experts. However, policy makers still need to also make available the more technical and detailed information about the evidence so that those in the public who do have a greater understanding of the topic can engage with the technical aspects of the evidence

Key questions

- Is the evidence underpinning policy advice clearly accessible?
- Is the underlying data made publicly available through explicit open government data commitments?
- Is it possible to justify decisions with reference to how issues were framed, how experts were selected and how evidence was interpreted?
- Is information presented to the public in a clear and understandable fashion?

Building evidence through emerging technologies and mobilising data

Why is it important?

Due to the fast advancement of digital technology, a wealth of standards currently being developed aim at encouraging and guiding governments' use of emerging technologies and data to develop, govern, and improve data and hence evidence generation in the public sector, relying on the new powers of data (e.g. Big Data, Open Data) and artificial intelligence. This chapter offers a preliminary overview of these issues, while acknowledging that this remains a fast moving field. The goal is mainly to ensure that the reader can be broadly aware of both the opportunities and challenges offered by these technologies and by the new data environment, while ensuring that these can be appropriately mobilised to feed into policy relevant evidence informed decision making processes.

Although these set of standards can be seen as a "softer" option to regulations (Wagner, 2018[74]), setting and following them are particularly important because this ensures both efficiency, consistency and ethical behaviours of public servants when they are handling government held data, as well as of those in charge of processes where these data are fed into algorithms and emerging technology systems supporting policy processes to inform policy decision making. This is important to maintain trustworthy processes in an area of rapid technological development.

Summary of the mapping of existing approaches

Achieving the promise of emerging technologies through appropriate institutional set ups

With emerging technologies playing an important role in decision making processes and in the generation, access, sharing and use of data and evidence, organisations need to gain trust from citizens in their responsible behaviours and practices. For this, governments have often established data governance arrangements, inclusive of independent bodies, and specific legislations and frameworks with the following four dimensions: ethics, privacy and consent, transparency and security. The goal is that the outcomes of the emerging technologies can represent valuable contributions to evidence-informed policy-making while adhering to values and respecting the rights of individuals.

Since something lawful does not necessary mean ethical, many governments rely on independent bodies and frameworks to ensure data and information are generated, accessed, shared and used in a responsible manner. Bodies operating at arms' length from government support good data governance practices across public sector entities to further build their capability to use data and manage it as a valuable strategic asset for instance by to reducing access barriers to data, implementing data standards and experimenting with new methodologies and data itself in a safe environment whilst using the data ethically and responsibly. These bodies have been set up in many OECD countries. For example, in Portugal, the National Commission for Data Protection (CNPD) is an independent entity with powers of authority extending throughout the country. It supervises and monitors compliance with the laws and regulations in the area of personal data protection, with strict respect for the human rights and the fundamental freedoms and guarantees enshrined in the Constitution and the law. Public and private entities have to notify the CNPD regarding any personal data treatment made by them. Another example is the CNIL in France (CNIL, National Commission on Information and Liberty). The latest OECD Digital Economy Outlook offers an overview of recent developments (OECD, 2020[75]), where the majority of countries reported to have in place a type of legislation for privacy and personal data protection. Even so, the report highlights the effort of many countries to address the collection, processing and sharing of personal data to support COVID-19 by endorsing privacy-enhancing solutions such as systems to manage encrypted data to prevent the access to personal data.

Countries face a number of key challenges to achieve a Data Driven Public Sector (2019[76]). One of them is to guarantee the interoperability of data. Interoperability is a critical issue at both the national and regional level to maximise the potential of data. The European Interoperability Framework looks to establish and adopt universal data standards to member states of the European Union. In Korea this is done through the Enterprise Architecture for managing data resources and in Italy, the National Digital Data Platform (PDND) is aimed at improving interoperability of national public information assets.

Developing frameworks and principles for the data governance

Data privacy, consent, and agency are also fundamental for public trust when data is being handled to inform evidence. Due to the increasing amount of data, and particularly sensitive data, being used and held by governments to create evidence for better policy-making, the careful, safe and responsible treatment of that data is essential (OECD, 2019[76]). In 2008, the OECD Recommendation for Enhanced Access and More Effective Use of Public Sector Information had addressed issues of openness, access and transparent conditions for use, as well as quality and integrity among others[5]. This recommendation provided a framework for the broader and more effective use of public sector information, such as: the presumption of openness as the default rule to facilitate access and re-use; integrity; access and transparent conditions for re-use; and sharing best practices to educate users and re-users, within other principles. In 2014, the OECD Recommendation on Digital Government Strategies (2014[77]) made explicit reference to the need for developing frameworks to enable, guide, and foster access to, use and re-use of

the increasing amount of evidence, statistics and data concerning operations, processes and results in the public sector.

Given the growing use of emerging technologies in the public sector for information generation, the transparency of data use, the purpose, the algorithmic systems it informs, and the decisions made once applied are essential for accountability purposes. Indeed, citizens may not be informed about the data being used, how and by whom (Saidot, 2019[78]). This is why transparency in the use, access and sharing of data matters to ensure quality and reliability, contributing to trust in the evidence generated using such data (Ubaldi et al., 2019[79]). Failing to make data or algorithms transparent can also result in difficulties in standardizing evidence, with possible biased results.

Data governance instruments such as frameworks, guidelines and principles provide users with information, resources and approaches to assist them achieve ethical practices and decisions making in handling data, particularly at the practitioner level for they promote self-regulation and control in practice. They are not intended to be prescriptive but aim at widening common understanding and work through ethical concerns.

Within the European Union, the General Data Protection Regulation (GDPR) came into force in May 2018. This regulation guarantees the rights of the data subject to be informed about the use of their data, to access, to edit, to remove, and to restrict the use of their data, and the portability of their data. The regulation intends to protect citizens / data subjects with regard to the processing of their data.

The United Kingdom developed the Data Ethics Framework to guide policymakers and data analysts in the ethical implications of the work they are undertaking. This framework provides a foundation for the work being done in the field of data science, requiring that all activity be as open and accountable as possible (GOV.UK, 2019[80]).The US Federal Data Strategy centres builds on the Evidence-Based Policy making act from 2018 (OECD, 2019[76]). The mission of the Federal Data Strategy is to enhance the value of federal data for the mission, service, and the public good by guiding the Federal agencies in the areas of: Ethical Governance, Conscious Design, and Learning Culture. The strategy comprises principles, practices, and an annual action plan to guide federal data management and use. The practices include guidance on issues such as how to build a culture that values data and on how to promote Efficient and Appropriate Data Use. Similarly, in Canada, the Canadian Data Strategy intends to support federal public services foster a more strategic use of data while protecting citizens' privacy and building on current federal data initiatives to ensure complementarity, coherence and transparency, so that emerging opportunities are understood and quickly acted upon. New Zealand has established Principles for the safe and effective use of data and analytics supported by the Government Chief Data Steward (See Box 2.11). Chief Other countries have established specific posts with the responsibility of helping government to better use data as a resource across government, including:

- A nominated Data Protection Officer is part of the regulatory provisions of the UK Government Data Protection Regulation. There are sanctions for organisations if the Data Protection Officer is not properly resourced or supported.
- In France, the Lemaire Act, serves the purpose of promoting greater transparency: it aims to ensure a trustworthy public service of data by encouraging innovation and building a framework of trust that guarantees the rights of users while protecting their personal data (Dreyfus, 2019[81]).
- In Ireland, the National Research Ethics Committees Bill (2019[82]), along with parallel secondary legislation on clinical trials, promotes a streamlined, regulated, and fit-for-purpose model for the ethical review of health research projects, such as "*whether there are adequate safeguards in place to protect the privacy of individuals participating in the health research and the confidentiality of their personal data* (Head 23)" where it is not a requirement to provide personal data (Head 33). Equivalent laws exist in many EU and OECD countries.

Although making the processing of data access, sharing and use accessible and transparent is important in maintaining public trust, it is equally crucial to ensure data security. Efforts put in place to secure the processing and protection of data are essential. The increasing number of sophisticated hackers is a challenge that needs to be addressed starting by strengthening digital security in a comprehensive manner, both in the public sector and engaging with citizens. Government needs to not only protect itself but also to help citizens to understand how to keep themselves safe through their online interactions.

Box 2.11. New Zealand Principles for the safe and effective use of data and analytics

The Government Chief Data Steward (GCDS) role supports the use of data as a resource across government to help deliver better services. The GCDS aims to facilitate and enable a joined-up approach across government. As well as developing policy and infrastructure, the GCDS provides support and guidance so agencies can use data effectively, while maintaining the trust and confidence of citizens.

Within its lead roles on data, the Chief Government Data Steward and the Privacy Commissioner have developed the *Principles for the safe and effective use of data and analytics*. These principles support the development of government agencies' guidance on the use of data and analytics for decision-making, such as:

- **Deliver clear public benefit:** The use of data and analytics must have clear benefits for the population**;**
- **Focus on people:** Keep in mind the people behind the data and how to protect them against misuse of information**;**
- **Ensure data is fit for purpose:** Using the right data in the right context can substantially improve decision-making and analytical models, and will avoid generating potentially harmful outcomes**;**
- **Maintain transparency:** Transparency supports collaboration, partnership, and shared responsibility, and is essential for accountability**;**
- **Retain human oversight:** Ensure significant decisions based on data involve human judgement and evaluation and regularly reviewed decision-making processes to make sure they're still fit for purpose**; and**
- **Understand the limitations:** Developing data capability helps to create depth of understanding and to implement the most useful data tools while keeping any limitations in mind.

Source: New Zealand Government (2020[83]), Government Chief Data Steward (GCDS) https://www.data.govt.nz/about/government-chief-data-steward-gcds/. Principles for safe and effective use of data and analytics, https://www.stats.govt.nz/about-us/data-leadership/ (Data.gov.nz, 2019[84]).

Balancing the dynamics of innovation and new technologies through good governance

Recent OECD's work on the regulation and governance of innovation[6] recognises that quickly evolving technologies are shifting control away from governments. The challenge is to maintain a balance between innovation and other values such as privacy, transparency and accountability in a complex and rapidly evolving context. The shift in power from the government to a broad range of non-governmental actors requires multi-stakeholder involvement as these non-government actors have had increasingly important roles in representing a variety of interest in the digital society. The inherent challenge is that social and economic actors are increasingly influenced by processes that are invisible to governments. For example, the proliferation of algorithms, with their 'black box' nature, creates information asymmetries between the

public and private sector, which raise complex accountability, transparency and regulatory concerns. This is all the more the case as such algorithms have an impact on the data and the evidence that are produced as a result. In cases where companies are using Machine Learning algorithms, it can be extremely difficult for regulators to determine what an algorithm is doing and why it is doing it. One possible solution might be to archive versions of the algorithm and to make them available to regulators under a non-disclosure agreement. Non-government actors have an increasingly important role in representing different interests for setting norms in the digital society that need to be reconciled. Consequently, multi-stakeholder approaches are important in setting societal and regulatory goals.

The use of advanced data analytics, with Machine Learning and Artificial Intelligence will require to make objectives explicit, exposing policy trade-offs that had previously been implicit and obscured (Coyle D., 2020[85]). Many of the Machine Learning systems may work as black boxes, with an implicit bias that results from the way that they have been specified in terms of users' experience. The question is the extent to which weighting decisions and trade-offs can be left with such systems that can only maximise broad utilitarian goals. This will hamper the demands for explaining, justifying and maintaining accountability of the evidence that can be produced through such approaches. These are of course frontier issues in terms of developing policies fit for the time, but which will have to be resolved if these technologies are to play a substantive role in the future beyond mere enforcement decisions such as in the health, police or justice areas.

The development of principles for the use of emerging technologies and new data and their impact on evidence-informed policymaking

The demand for principles has increased massively because of the growing use of emerging technologies, which also directly impacts the generation and use of evidence for policymaking. It is important to note the increasing focus on establishing ethical principles as a supporting framework for developing public policy in a way that avoids setting regulations that might in return have a negative impact on data access and sharing. Many private organisations use options for self-regulation to implement these ethical concerns.

Work is also undertaken at sectoral level, particularly in the health and social services area, where data is critical for evidence informed decision making and yet where many ethical issues arise. The Recommendation of the OECD Council on Health Data Governance (OECD, 2017[56]) calls upon countries to develop and implement frameworks that secure privacy while enabling health data uses that are in the public interest. It recommends that governments establish and implement a national health data governance framework that includes consideration of issues of informed consent and appropriate alternatives. The purpose is to provide clarity on whether individual consent to the processing of their personal health data is required and, if so, the criteria used to make this determination; what constitutes valid consent and how consent can be withdrawn. The New Zealand Social Investment Agency and Statistics New Zealand are developing a shared set of rules for the safe, ethical, and transparent use of social sector data (Cabinet Social Policy Committee, 2017[86]). In order to generate useful data for evidence informed policy making, analysts and researchers often need access personal data across Integrated Data Infrastructures. Such links and the use of integration through unique identifiers needs to happen while protecting the privacy and consent of users.

Besides privacy and consent, it is essential for organisations to make transparent and accountable the way new technologies are being used, the way data are being treated, and how they contribute to policy-making due to the fast development of these areas. For this, governments develop frameworks or principles to advise public servants on their behaviour or best practices when feeding data into algorithmic systems.

As mentioned above, some countries have established their own principles to ease information flow and set standards that enable data to be collected, analysed, and stored in the same way. Data principles are aimed to guarantee that citizens' data is treated responsibly, which increases evidence reliability, encourages more accurate policy-making and builds public trust. The OECD has also developed a set of

principles for Artificial Intelligence, which are practical and flexible enough to stand the test of time in a rapidly evolving field (Box 2.12). They complement existing OECD standards in areas such as privacy, digital security, risk management, and responsible business conduct (OECD, 2019[26]). These principles, serve not only to promote international standards as ways to develop and use AI systems (which lead to standardised methods for the generation of evidence) but also strengthen the trustworthiness of AI outcomes and policy decisions.

Box 2.12. The OECD Principles for Responsible Stewardship of Trustworthy AI

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

1. AI should benefit people and the planet by driving inclusive growth, sustainable development, and well-being.
2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
4. AI systems must function in a robust, secure, and safe way throughout their life cycles, and potential risks should be continually assessed and managed.
5. Governments, organisations, and individuals developing, deploying, or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Source: (OECD, 2019[26]).

In addition to data and AI principles, it is equally important to equip public servants with guidelines for handling data in flexible and agile ways while meeting high ethical standards. Assuring citizens that the data about them are being handled by public servants who act responsibly and with accountability according to published guidelines (Box 2.13) will increase public trust. Therefore the OECD has also been developing draft Ethics Good Practice Principles for Data Ethics in the Public Sector (Box 2.13).

Aiming at policymakers, statisticians, analysts, data scientists, and any public officers handling data, these guidelines provide public servants with a framework for the appropriate processing of data. The proposed ethics guidelines in the box below are intended to promote the ethical behavior of public servants as well as the rights of data subjects (citizens). The standardisation of behaviours and consistency of conducts that the guidelines below ensure enable public trust to be maintained.

Box 2.13. Good Practice Principles for Data Ethics in the Public Sector

Draft Good Practice Principles for Data Ethics in the Public Sector have been developed by the OECD Working Party on Senior Digital Government Officials in the context of the OECD work on digital government. As part of broader considerations, the current draft suggests that :

- Data use by Governments should serve the public interest.
- Data use by governments should deliver public good

There is a need to explore the collective and community nature of data governance, the environmental implications of data infrastructure, and the risks for abuse in the use of data, where risks go beyond the public sphere.

The Good Practice Principles invite to:

- **Use data with integrity**. Government should not abuse its position, the data at its disposal, or the trust of the public.
- **Be aware of and observe relevant arrangements for trustworthy data access, sharing and use.** Public officials should build knowledge on the specific governance arrangements framing data access, sharing and use, to secure they are respected and applied.
- **Incorporate data ethical considerations into governmental, organisational and public sector decision-making processes.** Public officials might consider the incorporation of data ethical considerations for the generation of public sector data and decisions on data collection, funding data projects, and use of data by third-parties.
- **Safeguard the agency of end-users of AI systems to make the final determination on the action taken following a machine-based recommendation.** Public officials should retain control over the data they access, share and use, including to help inform the development and training of those systems.
- **Be specific about the purpose of data use, especially in the case of personal data.** Make sure that data use has a clearly articulated purpose that explains the reason why data are being used and that addresses the concerns of different stakeholders.
- **Define boundaries for data collection, access, sharing and use.** Make sure that your design considers balanced data use by weighing relevant societal costs and benefits, with data minimization as the norm when it comes to personal data. This ensures the quality of design and the ability to explain how data are being used.
- **Be clear, inclusive and open.** The applied use of data should recognize, and mitigate, any potential bias in order that it never leads to discrimination with people in similar cases to ensure they always treated equally.
- **Broaden individuals' and collectives' control over their data.** Citizens are empowered and have action perspective to make decisions regarding the sharing of their personal data within, or external to, government.
- **Be accountable and proactive in managing risks.** Governments design mechanisms for giving citizens' insights into, and consent for the use of their personal data, by organizing internal and external accountability. Stakeholders should know where to address questions, remarks, or mistakes and governments should be responsive to the input of citizens.

Source: Adapted from OECD, (Forthcoming[87]), Good Practice Principles for Data Ethics in the Public Sector.

Principles on the standardisation of data use, AI systems, and public servants' behaviours ensure the quality of commitment made to be evaluated. Indeed, well-designed and responsible principles enable further collaborations, easier data and information collection, and the sharing of data, as well as more effective and reliable analyses for developing evidence-based contributions to policymaking.

Key questions

- How can evidence be collected following national and/or international principles coordinating the key dimensions of ethics, privacy, transparency, and security in the use and access to data?
- How to ensure that citizens are given an opportunity to confirm consent to the ethical use of the data to inform evidence as a way to foster trust in the results?
- What is the role of standards in ensuring that evidence drawn from AI processes be generated in ethical ways that foster public trust?
- What are the respective roles of ethical principles versus regulation to protect the core elements in the use of, and access to, data?

References

Allcott, H. and M. Gentzkow (2017), "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives*, Vol. 31/2, pp. 211-236, http://dx.doi.org/10.1257/jep.31.2.211. [45]

Barberá, P. et al. (2015), "Tweeting From Left to Right", *Psychological Science*, Vol. 26/10, pp. 1531-1542, http://dx.doi.org/10.1177/0956797615594620. [46]

Barlow, J. et al. (2016), "Questioning the outcome of the Building Blocks trial", *The Lancet*, Vol. 387/10028, pp. 1615-1616, http://dx.doi.org/10.1016/S0140-6736(16)30201-X. [15]

Bond (2018), *An introduction to the principles for assessing the quality of evidence*, Bond. [62]

Cabinet Social Policy Committee (2017), "Implementing Social Investment: Report Back", Office of the Minister of Finance Office of the Minister of State Services, Wellington, http://dx.doi.org/12345. [86]

Campbell Collaboration (2019), *Campbell Collaboration Guidance for establishing and managing a Stakeholder Advisor Review Group 1*, https://campbellcollaboration.org/guidance-on-establishing-managing-stakeholder-advisory-review-groups.html. [42]

Carson, L. (2017), *NewDemocracy Research and Development Note: Deliberation*, newDemocracy, http://www.newdemocracy.com.au/research-note-deliberation (accessed on 2 May 2019). [64]

Chwalisz, C. (2017), *The People's Verdict : adding informed citizen voices to public decision-making*. [63]

Conseil d'Etat (2020), *Conduire et partager l'évaluation des politiques publiques*. [49]

Council of the European Union (2015), "Council conclusions on the implementation of the EU Action Plan on Drugs 2013-2016 regarding minimum quality standards in drug demand reduction in the European Union", https://www.emcdda.europa.eu/system/files/attachments/8043/INT19_EU%20Min%20Quality%20Standards_ST11985.EN15.pdf. [22]

Coyle D., W. (2020), ""Explaining" machine learning reveals policy challenges", *Science,*, Vol. 368/6498, pp. 1433-1434, http://dx.doi.org/10.1126/science.aba9647. [85]

Data.gov.nz (2019), *Algorithm review underway to increase transparency and accountability*, https://www.data.govt.nz/blog/algorithm-review-underway-to-increase-transparency-and-accountability/. [84]

Dreyfus (2019), *France: Public service and processing of personal data*, https://dreyfus.fr/en/2019/08/05/public-service-and-processing-of-personal-data/. [81]

Early Intervention Foundation (2017), *Getting your programme assessed | EIF Guidebook*, http://dx.doi.org/12345. [50]

Eckenrode, J. et al. (2010), "Long-term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths", *Archives of Pediatrics & Adolescent Medicine*, Vol. 164/1, pp. 9-15, http://dx.doi.org/10.1001/archpediatrics.2009.240. [11]

European Commission (2002), *Communication from the comission on the collection and use of expertise by the commission: principles and guidelines*, Commission of the European Communities, http://ec.europa.eu/governance/docs/comm_expertise_en.pdf (accessed on 22 March 2019). [43]

European Food Safety Authority (2014), "Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment", *EFSA Journal*, Vol. 12/6, p. 3734, http://dx.doi.org/10.2903/j.efsa.2014.3734. [44]

European Food Safety Authority (2010), "Application of systematic review methodology to food and feed safety assessments to support decision making", *EFSA Journal*, Vol. 8/6, p. 1637, http://dx.doi.org/10.2903/j.efsa.2010.1637. [70]

EuroScientist (2017), *The Brussels declaration on ethics and principles for science and society policy-making*, EuroScientist, http://www.euroscientist.com/policy-making-manifesto-squaring-science-human-factor (accessed on 22 March 2019). [59]

Ferri, M. et al. (2015), "What is needed in future drug treatment research? A systematic approach to identify gaps on effectiveness of drug treatment from the EMCDDA", *Drugs: Education, Prevention and Policy*, Vol. 22/1, pp. 86-92, http://dx.doi.org/10.3109/09687637.2014.954988. [58]

Figueira, Á. and L. Oliveira (2017), "The current state of fake news: challenges and opportunities", *Procedia Computer Science*, Vol. 121, pp. 817-825, http://dx.doi.org/10.1016/j.procs.2017.11.106. [47]

France Stratégie (2019), *Expertise and democracy: Coping with mistrust*, https://www.strategie.gouv.fr/english-articles/expertise-and-democracy-coping-mistrust (accessed on 25 October 2019). [48]

Gluckman, P. (2014), "Policy: The art of science advice to government", *Nature*, Vol. 507/7491, pp. 163-165, http://dx.doi.org/10.1038/507163a. [25]

GOV.UK (2019), *Guidance Data Ethics Framework*, https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework. [80]

Government of Canada (2019), *Algorithmic Impact Assessment (AIA)*, https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html. [34]

Government of Ireland (2019), *National Research Ethics Committees Bill*. [82]

Government Office for Science (2010), *Principles of scientific advice to government - GOV.UK*, https://www.gov.uk/government/publications/scientific-advice-to-government-principles/principles-of-scientific-advice-to-government (accessed on 2 April 2019). [36]

Government office for Science (2010), *The Government Chief Scientific Adviser's Guidelines on the Use of Scientific and Engineering Advice in Policy Making*, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/293037/10-669-gcsa-guidelines-scientific-engineering-advice-policy-making.pdf. [6]

Hawkins, B. and J. Parkhurst (2016), "The 'good governance' of evidence in health policy", *Evidence & Policy: A Journal of Research, Debate and Practice*, Vol. 12/4, pp. 575-592, http://dx.doi.org/10.1332/174426415X14430058455412. [41]

Health Evidence (2018), *Quality Assessment Tool*, https://www.healthevidence.org/documents/our-appraisal-tools/quality-assessment-tool-dictionary-en.pdf (accessed on 8 March 2019). [8]

Healy, D. (2019), "The crisis in Cochrane: Evidence Debased Medicine", *Indian Journal of Medical Ethics*, Vol. IV/1, pp. 52-54, http://dx.doi.org/10.20529/ijme.2018.091. [32]

Home Visiting Evidence of Effectiveness (2018), *Assessing Evidence of Effectiveness*, https://homvee.acf.hhs.gov/Review-Process/4/Assessing-Evidence-of-Effectiveness/19/7 (accessed on 19 February 2019). [54]

INTOSAI (2016), *Guidelines on the Evaluation of Public Policies*, INTOSAI. [51]

James Lind Alliance (2019), *About the James Lind Alliance*, http://www.jla.nihr.ac.uk/ (accessed on 22 April 2020). [61]

Langeveld, K., K. Stronks and J. Harting (2016), "Use of a knowledge broker to establish healthy public policies in a city district: a developmental evaluation", *BMC Public Health*, Vol. 16/1, p. 271, http://dx.doi.org/10.1186/s12889-016-2832-4. [7]

Lavis, J. et al. (2010), *SUPPORT Tools for evidence-informed health Policymaking (STP) 9: Assesing the applicability of the findings of a systematic review*, BioMed Central, http://dx.doi.org/10.1186/1478-4505-7-S1-S9. [20]

Leadbeater, B. et al. (2018), "Ethical Challenges in Promoting the Implementation of Preventive Interventions: Report of the SPR Task Force", *Prevention Science*, pp. 1-13, http://dx.doi.org/10.1007/s11121-018-0912-7. [31]

Leviton, L. and M. Trujillo (2017), "Interaction of Theory and Practice to Assess External Validity", *Evaluation Review*, Vol. 41/5, pp. 436-471, http://dx.doi.org/10.1177/0193841X15625289. [10]

Love, P. and J. Stockdale-Otárola (eds.) (2017), *Debate the Issues: Complexity and Policy making*, OECD Insights, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264271531-en. [4]

Mair, D. et al. (2019), *Understanding our political nature: How to put knowledge and reason at the heart of political decision-making*, Publications Office of the European Union, http://dx.doi.org/10.2760/374191. [30]

McDonald, R. (2010), *The Government Chief Scientific Adviser's Guidelines on the Use of Scientific and Engineering Advice in Policy Making*, Government Office for Science, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/293037/10-669-gcsa-guidelines-scientific-engineering-advice-policy-making.pdf (accessed on 22 March 2019). [67]

Mejdoubi, J. and S. Midwifery (2014), "Effects of nurse home visitation on cigarette smoking, pregnancy outcomes and breastfeeding: a randomized controlled trial", *nursingplus.com*, http://www.nursingplus.com/article/S0266-6138(13)00243-X/abstract (accessed on 31 January 2018). [13]

Moberg, J., P. Alonso-Coello and A. Oxman (2015), *GRADE Evidence to Decision (EtD) Frameworks Guidance. Version 1.1*, The GRADE Working Group, https://ietd.epistemonikos.org/#/help/guidance (accessed on 21 April 2020). [21]

Munthe-Kaas, H., H. Nøkleby and L. Nguyen (2019), "Systematic mapping of checklists for assessing transferability", *Systematic Reviews*, Vol. 8/1, http://dx.doi.org/10.1186/s13643-018-0893-4. [18]

National Institute for Health and Care Excellence (2013), *How NICE measures value for money in relation to public health interventions*, https://www.nice.org.uk/Media/Default/guidance/LGB10-Briefing-20150126.pdf (accessed on 1 May 2019). [39]

New Zealand Government (2020), *Government Chief Data Steward*, https://www.data.govt.nz/about/government-chief-data-steward-gcds/. [83]

Nuffield Council on Bioethics (2012), *Emerging biotechnologies: technology, choice and the public good*, Nuffield Council on Bioethics, http://www.espcolour.co.uk (accessed on 2 April 2019). [37]

OECD (2020), *Delivering evidence based services for all vulnerable families*, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DELSA/ELSA/WD/SEM%282020%298&docLanguage=En. [16]

OECD (2020), *OECD Digital Economy Outlook 2020*, OECD Publishing, https://doi.org/10.1787/bb167041-en. [75]

OECD (2019), *OECD Best Practice Principles for Regulatory Policy: Regulatory Impact Assessment*, OECD, Paris, http://dx.doi.org/123. [52]

OECD (2019), *OECD Principles on Artificial Intelligence*, https://www.oecd.org/going-digital/ai/principles/ (accessed on 13 January 2020). [26]

OECD (2019), *Openness and Transparency - Pillars for Democracy, Trust and Progress*, https://www.oecd.org/fr/corruption/opennessandtransparency-pillarsfordemocracytrustandprogress.htm (accessed on 2 May 2019). [68]

OECD (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/059814a7-en. [76]

OECD (2018), *OECD Regulatory Policy Outlook 2018*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264303072-en. [53]

OECD (2017), "Policy Advisory Systems: Supporting Good Governance and Sound Public Decision Making", https://www.oecd-ilibrary.org/docserver/9789264283664-en.pdf?expires=1556887846&id=id&accname=ocid84004878&checksum=7A3FD77BBF5B7A94E579CD379493A5CA (accessed on 3 May 2019). [40]

OECD (2017), *Recommendation of the Council on Open Government*, http://www.oecd.org/gov/Recommendation-Open-Government-Approved-Council-141217.pdf. [35]

OECD (2017), *Recommendation of the OECD Council on Health Data Governance*, OECD, http://www.oecd.org/health/health-systems/Recommendation-of-OECD-Council-on-Health-Data-Governance-Booklet.pdf (accessed on 2 April 2019). [56]

OECD (2016), *Open Government: The Global Context and the Way Forward*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264268104-en. [55]

OECD (2015), "Scientific Advice for Policy Making: The Role and Responsibility of Expert Bodies and Individual Scientists", *OECD Science, Technology and Industry Policy Papers*, No. 21, OECD Publishing, Paris, https://dx.doi.org/10.1787/5js33l1jcpwb-en. [73]

OECD (2014), "Recommendation of the Council on Digital Government Strategies". [77]

OECD (2011), *Ministerial Advisors: Role, Influence and Management*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264124936-en. [38]

OECD (2008), *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*. [88]

OECD (2003), *Recommendation of the council on guidelines for managing conflict of interest in the public service*, http://www.oecd.org/governance/ethics/2957360.pdf (accessed on 3 October 2018). [28]

OECD (2000), *Trust in Government: Ethics Measures in OECD Countries*, http://www.oecd.org/puma (accessed on 3 October 2018). [29]

OECD (Forthcoming), *Good Practice Principles for Data Ethics in the Public Sector*. [87]

Office of Management and Budget (2018), *Building Capacity to Produce and Use Evidence*, Office of Management and Budget, http://www.whitehouse.gov/omb/evidence. (accessed on 2 April 2019). [72]

Office of Management and Budget (2017), *Analytical Perspectives, Budget of the United States Government: Building the Capacity to Produce and Use Evidence*, Office of Management and Budget, Washington, https://obamawhitehouse.archives.gov/omb/budget/Analytical_Perspectives (accessed on 28 May 2019). [5]

Olds, D. et al. (2003), "Taking preventive intervention to scale: The nurse-family partnership", *Cognitive and Behavioral Practice*, Vol. 10/4, pp. 278-290, http://dx.doi.org/10.1016/S1077-7229(03)80046-9. [12]

Oliver, K. and W. Pearce (2017), "Three lessons from evidence-based medicine and policy: increase transparency, balance inputs and understand power", *Palgrave Communications*, http://dx.doi.org/10.1057/s41599-017-0045-9. [57]

Oliver, S. et al. (2018), *Stakeholder Engagement for Development Impact Evaluation and Evidence Synthesis*. [65]

Parkhurst, J. (2017), *The politics of evidence : from evidence-based policy to the good governance of evidence*, Routledge, London, http://researchonline.lshtm.ac.uk/3298900/ (accessed on 23 November 2018). [1]

Parkhurst, J. and S. Abeysinghe (2016), "What Constitutes "Good" Evidence for Public Health and Social Policy-making? From Hierarchies to Appropriateness", *Social Epistemology*, Vol. 30/5-6, pp. 665-679, http://dx.doi.org/10.1080/02691728.2016.1172365. [2]

Pollock, A. et al. (2018), "Stakeholder involvement in systematic reviews: a scoping review", http://dx.doi.org/10.1186/s40900-017-0060-4. [66]

Rantala, L. et al. (2017), "How to Earn the Status of Honest Broker? Scientists' Roles Facilitating the Political Water Supply Decision-Making Process", *Society & Natural Resources*, Vol. 30/10, pp. 1288-1298, http://dx.doi.org/10.1080/08941920.2017.1331484. [23]

Robling, M. et al. (2016), "Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): a pragmatic randomised controlled trial", *The Lancet*, Vol. 387/10014, pp. 146-155, http://dx.doi.org/10.1016/S0140-6736(15)00392-X. [14]

Saidot (2019), *A Consortium of Finnish organisations seeks for a shared way to proactively inform citizens on AI use*, https://www.saidot.ai/post/a-consortium-of-finnish-organisations-seeks-for-a-shared-way-to-proactively-inform-citizens-on-ai-use. [78]

Shaxson, L. (2019), "Uncovering the practices of evidence-informed policy-making", *Public Money & Management*, Vol. 39/1, pp. 46-55, http://dx.doi.org/10.1080/09540962.2019.1537705. [3]

Sturmberg, J. (2019), "Evidence-based medicine—Not a panacea for the problems of a complex adaptive world", *Journal of Evaluation in Clinical Practice*, http://dx.doi.org/10.1111/jep.13122. [33]

The California Evidence-Based Clearinghouse for Child Welfare (2018), *Scientific Rating Scale*, http://www.cebc4cw.org/ratings/scientific-rating-scale/ (accessed on 25 January 2019). [17]

The GRADE Working Group (2015), *Key DECIDE tools*, http://www.decide-collaboration.eu/ (accessed on 21 April 2020). [19]

The Royal Society (2018), *Evidence synthesis for policy a statement of principles*, https://royalsociety.org/-/media/policy/projects/evidence-synthesis/evidence-synthesis-statement-principles.pdf (accessed on 2 April 2019). [69]

Ubaldi, B. et al. (2019), "State of the art in the use of emerging technologies in the public sector", *OECD Working Papers on Public Governance*, No. 31, OECD, Paris. [79]

UK Committee on Standards in Public Life (2019), *The 7 principles of public life - GOV.UK*, http://dx.doi.org/1234. [27]

van Ooijen, C., B. Ubaldi and B. Welby (2019), "A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance", *OECD Working Papers on Public Governance*, No. 33, OECD Publishing, Paris, https://dx.doi.org/10.1787/09ab162c-en. [71]

Wagner, B. (2018), *Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?*, https://www.privacylab.at/wp-content/uploads/2018/07/Ben_Wagner_Ethics-as-an-Escape-from-Regulation_2018_BW9.pdf. [74]

Wandersman, A. et al. (2000), "Getting to outcomes: a results-based approach to accountability", *Evaluation and Program Planning*, Vol. 23/3, pp. 389-395, http://dx.doi.org/10.1016/S0149-7189(00)00028-8. [9]

What Works Clearinghouse (2020), *Standards Handbook (Version 4.1)*, https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf (accessed on 5 February 2019). [60]

Wilsdon, J., M. Saner and P. Gluckman (2018), "INGSA Manifesto for 2030: Science Advice for Global Goals", INGSA, http://dx.doi.org/10.1057/palcomms.2016.77. [24]

Notes

[1] See French Law 2013-907, that complements the law 83-634 from 1983 that initially concerned only senior civil servants, and that extends the field of control to the members of governments and heads of many agencies and public organisation.

[2] Issues around uncertainty create particular challenges for conducting economic evaluation, which is discussed in section 7.

[3] See www.ecologie.gouv.fr/sites/default/files/CDDEP_Guide%20du%20dialogue%20avec%20les%20parties%20prenantes.pdf.

[4] See www.conventioncitoyennepourleclimat.fr.

[5] See https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0362.

[6] The 2020 OECD Global Conference on Governance Innovation addressed the need for agile regulatory frameworks in the context of the 4th industrial revolution, to promote outcome focused, anticipatory approaches, and enhance accountability. See www.oecd.org/fr/reformereg/politique-reglementaire/oecd-global-conference-on-governance-innovation.htm.

3 Standards of Evidence: Mapping the experience in OECD countries

While many experts in governments recognise the importance of using evidence in the policy-making process, figuring out which evidence is robust and translating it to decision makers is challenging. This chapter reviews the existing standards used to strengthen the quality of evidence supporting the design of public interventions, policies and programmes, and promoting its uptake in policy-making. The chapter covers seven issues: evidence synthesis; theory of change and logic underpinning the programme; design and development of policies and programmes; efficacy; effectiveness; cost (effectiveness); and implementation and scale up of intervention. Each subsection offers details of why each standard is important and a summary of existing approaches used in OECD countries.

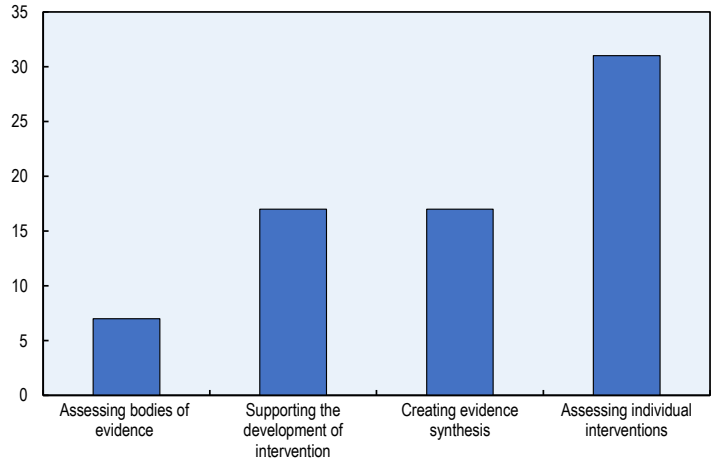
Introduction

The approaches to evidence standards found across OECD countries are focused on a variety of forms of evidence and cover seven standards of evidence. This chapter is structured as follows: i) the four main functions of the standards of evidence; ii) distribution of standards of evidence across the OECD countries; iii) forms of evidence that are addressed by the standards; iv) and, an introduction to the seven standards of evidence reviewed in this report.

What are standards of evidence for?

The standards of evidence reviewed in this report vary in terms of the 'unit of analysis' they focus on. Some standards focus on the entirety of the existing evidence base (evidence synthesis). This includes standards focused on assessing the quality of existing evidence syntheses and standards focused on the generation of new evidence syntheses (see Figure 3.1).

Figure 3.1. Focus of the standards of evidence



Source: Source: Mapping Exercise for this report.

Other standards are focused on generation of new evaluation evidence. This includes standards for assessing the evidence for an individual intervention as well as standards for supporting the development of an intervention. The approaches also vary in whether they address one key function or whether they address multiple functions. More than half of the approaches have one key function, versus the rest with multiple functions.

Assessing the strength of evidence of an intervention

Assessing the strength of evidence of an intervention refers to standards that examine the quality design and the robustness of findings of single studies in order to determine the strength of evidence for individual interventions. This assessment also involves an analysis of the findings and impacts in the study/studies. For instance, What Works for Kids (Australia) rates the evidence according to the evaluation(s) that has been conducted on each programme (Nest What works for kids, 2012[1]). Another approach that focuses on individual intervention is Social Programmes That Work (USA) , that seeks to identify those social programmes shown in rigorous studies to produce sizable, sustained benefits to participants (Coalition for Evidence-Based Policy, 2010[2]).

Assessing bodies of evidence

Assessing bodies of evidence refers to standards for appraising the totality of evidence included in a review of an evidence base. This includes: (a) the nature of the totality of evidence; (b) the extent and distribution of that evidence; and (c) the methods for undertaking a review (Gough and White, 2018[3]). Health Evidence provides a Quality Assessment Tool to evaluate systematic reviews or bodies of evidence (Health Evidence, 2018[4]). Another example is the Practice Scoring Instrument from Crime Solutions (USA). They present guidelines to identify a body of evidence (e.g. what qualifies as an eligible meta-analysis?) and evaluate it: eligibility criteria; comprehensive literature search; methodological quality; publication bias (Crime Solutions, 2013[5]).

Reviewing the evidence base for an intervention

Reviewing the evidence base for an intervention refers to standards for creating evidence synthesis. This can include methods to identify, select, appraise, and synthesize high quality research. For instance, Clearinghouse for Labor Evaluation and Research (CLEAR), reviews the evidence base for interventions, providing information related to the selection process, key features of all the relevant research identified for a given topic area, and reference documents of the review process (Clearinghouse for Labor Evaluation and Research, 2017[6]). An additional example is The Community Guide approach, which has a guide for the execution of systematic reviews (Zaza et al., 2000[7]).

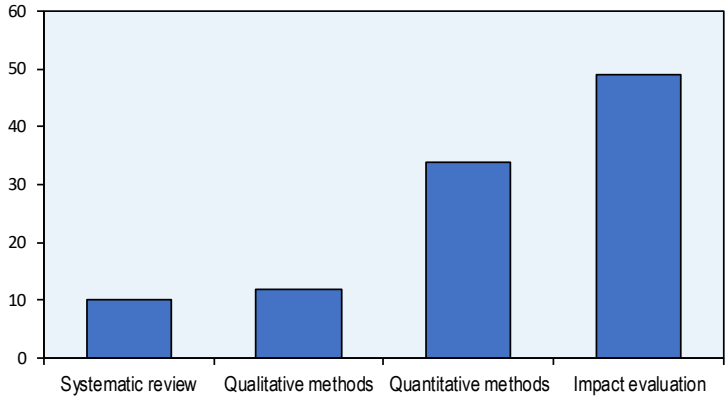
Supporting the development of an intervention

Supporting the development of an intervention refers to standards that focus on creating guidelines for the use of entities to better understand how an intervention fits into an implementing site's existing work and context (National Implementation Research Network, 2018[8]). The European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) has developed the *European drug prevention quality standards,* which outline the steps to be taken when planning, conducting, or evaluating programmes. These standards inform the development of interventions and serve as a reference framework for professional development (2011[9]).

What forms of evidence are covered by standards of evidence?

The mapped approaches cover a variety of different types of evidence, including quantitative research, impact evaluation, systematic review, and qualitative research (See Figure 3.2). Approaches can focus specifically on one type of evidence or may focus on more than one type. A small number only cover one kind of evidence, versus the majority with multiple use of these types of evidence. *Education counts* describes the four types (Alton-Lee, 2004[10]).

Figure 3.2. Type of evidence



Source: Mapping Exercise for this report.

- Most of the approaches assess evidence **impact evaluations**. This can include Randomized Control Trials (RCTs) and Quasi-Experimental Designs (QEDs).
- Thirty-four approaches concern **quantitative methods**. This includes approaches that assess correlational analyses, single case studies and pre-post studies without control groups.
- Twelve approaches evaluate **qualitative methods**. This includes approaches that include interviews, focus groups, panels of experts, or ethnographies. Qualitative research is often concerned with the implementation process
- Ten approaches assess **systematic review or meta-analysis**.

What are the seven issues covered in the standards of evidence?

The seven standards of evidence are: Evidence synthesis; Theory of Change and Logic underpinning the Programme; Design and Development of Policies and Programmes; Efficacy; Effectiveness; Cost (effectiveness); and, Implementation and scale up of intervention.

Evidence Synthesis

Evidence synthesis informs policy makers of what is known from research, making it fundamental for informing policy decisions and for promoting the uptake and use of evidence from evaluations and other evidence (Oliver et al., 2018[11]; Shemilt et al., 2010[12]). Evidence syntheses come in a variety of forms and of varying quality (as with primary studies), so standards to enable readers to appraise the quality of evidence synthesis are critical.

Evidence synthesis is an important tool for good knowledge management. Given the breadth of literature, including impact evaluations and RCTs, being published each year, knowledge management is essential as it becomes more difficult for policy makers and practitioners to keep abreast of the literature. Furthermore, policies should ideally be based on assessing the full body of evidence, not single studies, which may not provide a full picture of the effectiveness of a policy or programme.

Why is it important?

Evidence syntheses provide a vital tool for policy makers and practitioners to find what works, how it works – and what might do harm. Evidence syntheses are also critical in informing what is not known from previous research. As with primary studies, readers can (and should) appraise the quality and relevance of evidence synthesis (Gough, Thomas and Oliver, 2019[13]).

Since the early 2000s, across many sectors and countries, there has been an increase in the number of impact evaluations, including Randomization Control Trials (RCTs), being published each year (White, 2019[14]). For example, in education around ten RCTs were published each year in the early 2000s, growing to over 100 a year by 2012. As the number of studies increases, it becomes more difficult for policy-makers and practitioners to keep abreast of the literature. Furthermore, evidence synthesis allows for the amalgamation of findings and easier navigation of bodies of literature (Gough, Thomas and Oliver, 2019[13]), not single studies—which may not provide a full picture of the effectiveness of a policy or programme.

Evidence synthesis can come in a variety of forms, depending on the research questions and resources available, such as:

- *Map of maps*: provide reports from other evidence and gap maps in that policy space and by doing so act as a navigation tool (Gough, Thomas and Oliver, 2019[13]);
- *Mega-maps:* show other maps and reviews, but not primary studies;
- *Evidence and gap maps*: are even broader in scope but report a far more limited range of information about the reviews and primary studies they include (Saran and White, 2018[15]);
- *Review of reviews*: may be broader in scope but may be more restricted in the depth of analysis. This method only includes existing reviews, preferably systematic, rather than primary studies (Saran and White, 2018[15]);
- *Systematic reviews:* are narrow in scope but provide in-depth analysis (Saran and White, 2018[15]). This is the most robust method for reviewing, synthesising and mapping existing evidence on a particular policy topic. It is more resource-intensive, as it can take up to 8 to 12 months minimum and requires a researcher team (The UK Civil Service, 2014[16]). Systematic reviews have a number of stages, including: defining the review question; conceptual framework; inclusion criteria; search strategy; screening; coding of information from each study; quality and relevance appraisal; and, synthesis of study findings to answer the review question (Gough, Thomas and Oliver, 2019[13]).
- *Meta-analysis:* It refers to the use of statistical methods to summarise the results from individual programme evaluations on a given topic. A meta-analysis produces a weight-of-the-evidence summary to achieve a specific outcome or the relationship between one outcome and another; and

therefore, to draw an overall conclusion about the average effectiveness of a programme (Washington State Institute for Public Policy Benefit, 2017[17]).

- *Rapid Evidence Assessment (REA):* It is a quick overview of existing research on a (constrained) topic and a synthesis of the evidence provided by these studies to answer a specific policy issue or research question. REAs tend to be rigorous and explicit in method and thus systematic but make concessions on the depth of the process by limiting particular aspects of the systematic review process such as the screening stage (e.g. only electronically available texts) or considering using less developed search strings (The UK Civil Service, 2014[16]).
- *Quick Scoping Review:* It consists of a quick overview of the available research— from accessible, electronic and key resources, going up to two bibliographical references on a specific topic —to determine the range of existing studies on the topic. This non-systematic method can take from 1 week to 2 months strings (The UK Civil Service, 2014[16]).

Evidence synthesis also has a critical role to play in evidence-informed recommendations and guidance. In a number of policy areas, notably in health, formal processes have been developed for interpreting research evidence in order to develop and make recommendations (Ferri and Griffiths, 2015[18]) (see Box 3.1). At the European level, the EMCDDA (2020[19]) with its experience in monitoring and disseminating best practice promotes and supports guideline adaptation. An inventory of national guidelines and standards in treatment, prevention and harm reduction functions as a tool for ensuring that there are processes for translating the evidence base into appropriate recommendations and guidelines. At a global level the WHO produces guidelines that are underpinned by evidence synthesis (Oxman, Lavis and Fretheim, 2007[20]).

Box 3.1. Evidence Syntheses as a tool to support decision making

- Systematic reviews that provide accurate estimates of the effects of a policy or intervention can have a significant influence on decisions about whether to implement or disinvest in an intervention.
- Several global organisations produce evidence informed guidelines based on systematic reviews. The WHO produces guidelines that are underpinned by systematic reviews of interventions and health technologies that aim to follow a transparent and evidence-informed process.
- The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group has taken a leading role in developing guidance and methods for using research evidence to inform decision making. GRADE offers an explicit and transparent system for rating certainty in the evidence that underpins conclusions in a systematic review, a Health Technology Assessment or a guideline.
- Many of the approaches mapped by this report either use or are based-on GRADE. GRADE is used by over 100 organisations worldwide, including the WHO, the Cochrane Collaboration and the Campbell Collaboration. The Campbell Collaboration aims to promote positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-informed policy and practice.

Source: Adapted from Montgomery, Movsisyan, Grant, Macdonald and Rehfuess (2019) Saran and White (2018[15]).

Many OECD countries have a strong focus on producing systematic reviews and evidence-informed recommendations and guidance, following the long established practice of the Cochrane centres in the health area. The Danish government funds 'Cochrane Denmark', which supports synthesizing and dissemination of the best available evidence for health professionals, researchers and decision-makers (The Cochrane Collaboration, 2021[21]). The Norwegian Institute of Public Health, a government agency under the Ministry of Health and Care Services, has a strong focus on producing evidence synthesis to support decision making. Recent reviews include a live map of COVID-19 evidence (Norwegian Institute of Public Health, 2020[22]) and a review of weight reduction strategies among adults with obesity (Norwegian Institute of Public Health, 2021[23]).The Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU) is an independent national agency, tasked by the government with assessing health care and social service interventions, covering medical, economic, ethical and social aspects. SBU conducts health technology assessments and systematic reviews of published research to support key decisions in health, medical care and social services. These approaches are now being extended to other areas, beyond health, such as social policy, The SBU is currently pioneering a new international initiative in this area.

Summary of the mapping of existing approaches

Of the approaches included in the mapping, around half concern evidence synthesis. These could be divided into two broad categories, standards for assessing the quantity and quality of existing evidence syntheses and standards for executing evidence synthesis.

Standards for assessing the quantity and quality of existing reviews.

Around half of the approaches concerned with evidence synthesis were primarily concerned with providing standards for the quantity and quality of existing syntheses.

Some approaches were primarily concerned with completing a review of existing reviews and translating this into conclusions about the strength of the evidence base for a policy or programme. These approaches include the Education Endowment Foundation's Teaching and Learning Toolkit (UK), What Works for Health (USA) and the European Monitoring Centre for Drugs and Drug Addiction. For example, What Works for Health (2010[24]) has a rating of 'Scientifically Supported' which is awarded to interventions that have one or more systematic review(s). The Education Endowment Foundation has developed a 'padlock' rating system to rank the practices within the Teaching and Learning Toolkit (see Box 3.2).

Box 3.2. The Education Endowment Foundation's Padlock Rating System

This rating is designed to determine the strength of the causal inference for impact on learning outcomes in schools, the quantity and consistency of the findings (both the overall pooled effect and the pattern of effects relating to moderator variables) and the ecological validity of the studies (where studies took place in schools with interventions managed by teachers rather than researchers). The security ratings are allocated as following:

1. **Very limited: One padlock:** Single studies with quantitative evidence of impact, including effect size data reported or calculable (e.g. from randomised controlled trials, well-matched experimental designs, regression discontinuity designs, natural experiments with appropriate analysis); and/or observational studies with correlational estimates of effect related to the intervention or approach; but no publically available meta-analyses.
2. **Limited:** Two padlocks: At least one publically available meta-analysis
3. **Moderate:** Three padlocks: Two or more publically available meta-analyses, which meet the following criteria: they have explicit inclusion and search criteria, risk of bias is discussed, and tests for heterogeneity are reported. They include some exploration of methodological features such as research design effects or sample size.
4. **Extensive:** Four padlocks: Three or more meta-analyses, which meet the following criteria: they have explicit inclusion and search criteria, risk of bias is discussed, and tests for heterogeneity are reported. They include some exploration of the influence of methodological features such as research design effects or sample size on effect size. The majority of included studies should be from school or other usual settings.
5. **Very Extensive:** Five padlocks: Three or more meta-analyses, which meet the following criteria: they have explicit inclusion and search criteria, risk of bias is discussed, and tests for heterogeneity are reported. They include some exploration of the influence of methodological features such as research design effects or sample size on effect size. The majority of included studies should be from school or other usual settings.

Source: Adapted from Education Endowment Foundation (2018[25]).

A small number of approaches go further in providing tools that can be used to rate the quality of existing evidence syntheses in order to reach conclusions about the strength of evidence of the body of evidence underpinning a policy or programme. These approaches include ROBIS, Crime Solutions and the EMMIE framework used by the What Works Centre for Crime Reduction and the What Works Centre for Children's Social Care. Some of these tools originate in the academic literature but have not yet been used by international clearinghouses. ROBIS is one of the most comprehensive and is described in detail in Box 3.3.

Box 3.3. ROBIS - tool to assess risk of bias in systematic reviews

The ROBIS tool aims to assess the risk of bias in systematic reviews (rather than in primary studies), and the relevance of a review to the research question. The development of this tool was based on a four stage approach: define the scope; review the evidence base; hold a face to face meeting; and refine the tool through piloting. The evaluation (or tool) is completed in three (3) phases:

1. **Assessing relevance (optional):** 1) Assessors report the question that they are trying to answer (target question) in terms of the PICO (participants, interventions, comparisons, and outcomes) or equivalent. 2) Assessors complete the relevant PICO for the systematic review. 3) Assessors are asked whether the two questions match. If a review is being assessed in isolation and there is not target question, then this phase can be omitted.
2. **Identifying concerns with the review process**: this phase involves the assessment of four domains covering the key review process:
    a. Study eligibility criteria: aims to assess whether primary study eligibility criteria were pre-specified (rather than on existing knowledge of the studies themselves), clear, and appropriate for the review question.
    b. Identification and selection of studies: aims to assess whether any primary studies that would have met the inclusion criteria were not included in the review (i.e. review of databases, range of terms in the searching, screening titles and abstracts, and assessing full text- studies for inclusion).
    c. Data collection and study appraisal: aims to assess whether bias may have been introduced through the data collection or to establish the risk of the bias assessment process (e.g. transcribing data or failing to collect relevant information). Therefore, the risk of bias assessment should involve a minimum of two reviewers working independently.
    d. Synthesis and findings: aims to assess whether, given the decision has been made to combine data from the primary studies, the reviewers have used appropriate methods to do so (i.e. whether between-study variation (heterogeneity) variation is taken into account, publication bias, or interpreting standards errors as standards deviations (in meta-analysis)).
3. **Judging risk of bias:** This phase considers whether the systematic review, as a whole, is at risk of bias. This assessment uses the same process as in phase 2, but the judgment regarding the bias is replaced with an overall judgment of risk of bias.

Source: Adapted from Whiting et al (2016[26]).

Several clearinghouses and What Works centres have also developed their own frameworks to assess existing evidence syntheses. The EMMIE framework score focuses on five dimensions which should be covered in any systematic reviews intended to inform crime prevention (Johnson, Tilley and Bowers, 2015[27]). These are the *Effect* of intervention, the identification of the causal *Mechanism**(s)*** through which interventions are intended to work, the factors that *Moderate* their impact, the articulation of practical *Implementation* issues, and the *Economic* costs of intervention. In the US, the Crime Solutions clearing house has also developed a detailed scoring system that is applied to existing systematic reviews (See more in Box 3.4).

Box 3.4. Crime Solutions – Practice Scoring Instrument

The National Institute of Justice's CrimeSolutions.gov from USA comprises two components — a web-based clearinghouse of programmes and practices, and a process for identifying and rating those programmes and practices. From its second component, Crime solution provides a scoring instrument to assess the quality, strength, and extent of a meta-analysis. The quality rating section analyses each meta-analysis, defined as eligible for consideration, under the following criteria:

- *Eligibility Criteria* - rates the degree to which the meta-analysis provides a clear, detailed statement of the inclusion and exclusion criteria used to determine whether primary studies were eligible for inclusion in the final meta-analysis. Inclusion and exclusion criteria must be outlined.
- *Comprehensive Literature Search* - rates the degree to which the meta-analysis conducted an exhaustive, comprehensive review of the literature in an attempt to identify all eligible studies.
- *Grey Literature Coverage* - assesses the extent to which a meta-analysis includes results from unpublished or "grey" literature sources.
- *Coder Reliability* - assesses how the authors of the meta-analysis handled reliability of data extraction from the primary research reports (ideally, two or more coders).
- *Methodological Quality* - assesses the extent to which the authors of the meta-analysis were aware of and attentive to the methodological quality of the studies included in the meta-analysis.
- *Outlier Analysis* - assesses whether the meta-analysis checks for effect size outliers in the data.
- *Handling Dependent Effect Sizes* - rates a meta-analysis based on its appropriate analysis of effect sizes (e.g. appropriate statistical procedures were used to handle dependent effect size estimates).
- *Effect Size Reporting* - assesses whether the meta-analysis reported an aggregate mean effect size that synthesized (averaged) effect sizes across one or more sets of multiple studies, and whether the meta-analysis provided estimates of precision around the point estimate(s).
- *Weighting of Results* - assesses whether the meta-analysis uses appropriate weighting schemes when estimating mean effect sizes and in other analyses in order to give greater weight to the effect sizes estimated with more precision (e.g., based on larger samples).
- *Analysis Model* - rates a meta-analysis based on whether the authors recognized and addressed the issue of random effects versus fixed effect analysis models (Preference is given to random effects models because they provide results that are generalizable)
- *Heterogeneity Attentiveness* - rates a meta-analysis on whether the authors were aware of and attentive to heterogeneity (i.e. variability) in the effect size estimates from the studies in the meta-analysis.
- *Publication Bias* - refers to whether the meta-analysis descriptively or statistically assessed the possibility of publication bias in the results (e.g. funnel plot graphs or contour-enhanced funnel plot graphs)

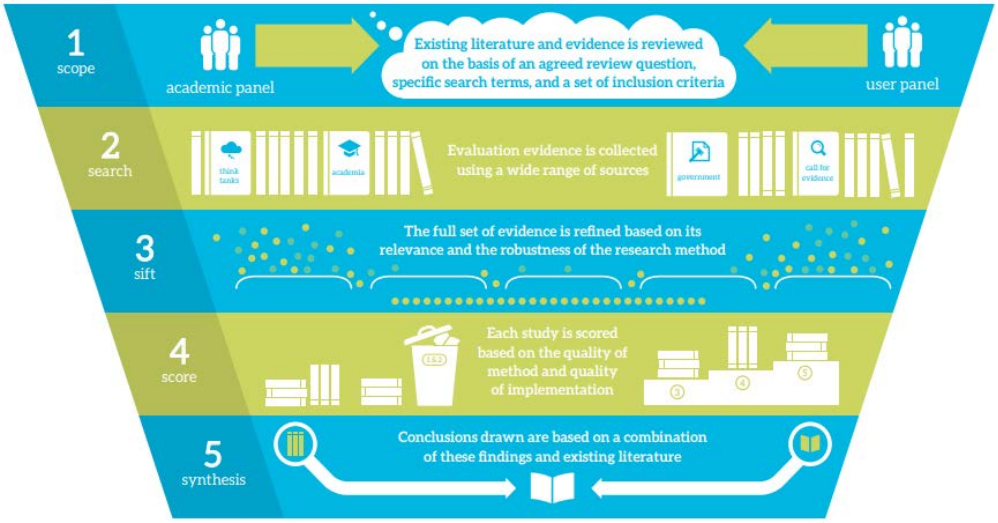
Source: Adapted from Crime solutions: Practices Scoring Instrument (2013[5]).

Standards for executing evidence synthesis

Seventeen of the approaches focus on standards for executing evidence synthesis. Figure 3.3 provides an example of the general stage for conducting reviews. First, standards for setting up and scoping a review (*Stage 1*), followed by standards for searching research (*Stage 2*), and standards rating the quality

of evidence and strength of recommendations (*Stage 4*). Details of the standards in *stage 1, 2* and *4* will be presented below. For example, The Equator Network contains a comprehensive searchable database of reporting guidelines and also links to other resources relevant to research reporting, this includes guidelines for systematic reviews from deciding the scope and the title of the review, to drawing conclusions (The EQUATOR Network, 2020[28]).

Figure 3.3. Methodology for conducting a review



Source: What Works for centre for Local Economic Growth (2015[29]).

Standards for setting up and scoping a review

Several approaches recommend the development of a protocol to define the conceptual framework for the review, the main review question, the inclusion and exclusion criteria, the review methods, and its documentation. These approaches include the Campbell Collaboration (2019[30]), What Works for Wellbeing (UK) (2017[31]), and EIF (UK) (2018[32]).

Other approaches also stipulate the use of the PICOS (Participants, interventions, comparisons, outcomes, and study design) framework to determine the inclusion and exclusion criteria of the review. For example, Campbell Reviews stipulate that the inclusion criteria should be stated specifically enough, with key terms clearly defined, to be applied with consistent results by anyone screening studies.

Standards for searching research

Some approaches stipulate that a search process should be transparent, comprehensive, and replicable. For instance, Education counts (NZ) (2004[10]) emphasises the importance of being transparent in approach, including the use of language when making claims as a fundamental tool to support both rigour and effective communication through each synthesis report.

Several approaches go further and provide guidelines to develop the search protocol, the search strategy, and how to document the search process. For instance, What Works for Wellbeing (UK) (2017[31]) specifies what to include in the search protocol (e.g. electronic sources to be searched, and restrictions), and refers to the need to balance sensitivity (ability to identify relevant information) and precision (ability to exclude irrelevant documents).

Other approaches also stipulate considering grey literature in the review to avoid publication bias. For instance, Evidence Based Teen Pregnancy Programmes (USA) (2016[33]) identifies new studies through

public calls for new and unpublished research. Early Childhood Foundation (2018[32]), on the other hand, provides methodologies in practice to measure the publication bias in their systematic reviews.

Standards for rating the quality of evidence and strength of recommendations

Most of the approaches agree that quality assessment is a critical stage of the evidence review process. Some approaches provide checklists to evaluate the evidence, including the issues of efficacy and effectiveness addressed in the following sections (Efficacy of an Intervention and Effectiveness of Interventions), including evidence for ESSA (USA) (2019[34]), the Strengthening Families Evidence Review (USA) (2018[35]), and the Community Guide (USA) (2018[36]). Another example is the European Food Safety Agency (2010[37]) which has produced guidance on the application of systematic review methodology to food safety assessments to support decision-making, which includes a number of key conclusions concerning the importance of methodological quality assessment:

- In a systematic review, each study should undergo a standardised assessment, checking whether it meets or not a predefined list of methodological characteristics, to assess the degree to which it is susceptible to bias.
- There are many stages of the review at which the validity of the individual studies is considered.
- Common types of bias that can occur in many different study designs are often classified as selection, performance, detection, attrition and reporting biases.
- Assessment of methodological quality involves using tools (e.g. Checklists) to identify those aspects of study design, execution, or analysis, which induce a possible risk of bias.
- It is important to distinguish between the quality of a study and the quality of reporting the study, although both may be correlated.

Other approaches not only stipulate a rating for the quality of evidence, but also provide information about the overall impact regarding the multiple outcomes in an intervention or policy. For instance, GRADE has developed a method for creating a clear separation between quality of evidence and strength of recommendations and presents a rating for each of these categories (See below Box 3.5).

Box 3.5. Grading of Recommendations Assessment, Development and Evaluation (GRADE)

The GRADE system provides an explicit, comprehensive criterion for downgrading and upgrading the quality of evidence ratings, and supporting the clarity of the process to translate evidence into recommendations for clinicians, patients, and policy makers.

How does the GRADE system classify quality of evidence?

- The GRADE system classifies the quality of evidence at one of the four levels—*high, moderate, low, and very low*, according to study limitations, directness of evidence, consistency of the results, precision, and reporting bias.

How does the GRADE system consider strength of recommendations?

- Strong - refers to when the desirable effects of an intervention clearly outweigh the undesirable effects, or clearly do not
- Weak - refers to when the trade-offs are less certain—either because of low quality evidence or because evidence suggests that desirable and undesirable effects are closely balanced—weak recommendations become mandatory.

Source: Adapted from Guyatt et al. (2008[38]).

Key questions

Standards for assessing the quantity and quality of existing reviews.

- Does a review include a systematic search for unpublished reports, to avoid publication bias?
- Is a protocol developed (project plan) for the review in advance?
- Are study inclusion and coding decisions carried out by at least two reviewers who work independently and compare results?
- Is study quality assessed?
- Does a review undergo peer review and editorial review?
- Does a review use rigorous methods to synthesize evidence, including, where appropriate, statistical meta-analysis of quantitative evidence and theory-based analysis of qualitative evidence?

Standards for setting up and scoping a review

- Does a review use the PICO (Participants, interventions, comparisons, outcomes, and study design) framework to determine the inclusion and exclusion criteria?

Standards for searching research

- Does the review process provide guidelines to develop the search protocol, the search strategy, and how to document the search process?
- Does the review process include grey literature?

Standards for rating the quality of evidence and strength of recommendations

- Does the review process distinguish between evidence quality and strength of recommendation?
- Does the evidence review process provide an independent checklist to evaluate the evidence regarding the issues of efficacy and effectiveness?
- Does the evidence review process provide an independent system to evaluate the strength of recommendation according to the effects of an intervention or programme and the quality of evidence?

Theory of Change and Logic Model underpinning the programme

Whereas the previous section of this chapter focused on the synthesis of existing evidence, the remaining sections focus on standards for various aspects of primary evidence generation using a monitoring and evaluation framework. This section focuses on the theory of change and logic model underpinning a programme.

A theory of change can be defined as a set of interrelated assumptions explaining how and why an intervention is likely to produce outcomes in the target population (European Monitoring Centre for Drugs and Drug Addiction, 2011[9]). A logic model sets out the conceptual connections between concepts in the theory of change to show what intervention, at what intensity, delivered to whom and at what intervals would likely produce specified short term, intermediate and long term outcomes (Axford et al., 2005[39]; Epstein and Klerman, 2012[40]). In some cases, a single theory of change might be difficult to identify due to multiple and complex interactions, it might be difficult to identify a unique course of action and the

underlying policy goals could be multiple and conflicting; this should not impede the activation of evidence processes.

Why is it important?

. Engaging in the process of developing a theory of change leads to better policy planning and implementation, because the policy or programme activities are linked to a detailed and plausible understanding of how change happens; while a logic model is a critical tool to allow detailed coherent and realistic policy planning.

Although a theory of change and a logic model are often expected to be developed during the planning stage, putting them in practice can be tied to time settings, political context, etc. However, they can also be useful in the monitoring and evaluation stage. For instance, to identify key indicators for monitoring or gaps in available data (Better Evaluation, 2012[41]). A full list of the benefits of developing both a theory of change and logic model is reproduced in Box 3.6.

Box 3.6. The benefits of developing an intervention theory of change and logic model

- The evaluability of the programme—both for implementation and for outcomes—is facilitated, by signposting appropriate metrics.
- The original intentions of the programme developers are clearly set out, and are explicit and open to critique.
- The underlying logic of the assumptions made in the theory, for example, that undertaking a certain activity will lead to a particular outcome, can be scrutinised.
- The realism of the assumptions made by the programme developers can be checked against wider evidence of 'what works', to assess the likelihood of the programme being successful.
- Commissioners can check the programme meets their needs; and providers and practitioners delivering the programme can check their own assumptions and the alignment of their expectations against the original intentions of the programme developers.
- The key parameters or boundaries (e.g., who is the programme for, and under what specific circumstances) can be set out, reducing the likelihood that the programme is used inappropriately or ineffectively.
- Core components (of content, or of implementation, or both) that are believed to be essential to the programme's effectiveness can be identified.
- Activity traps can be identified and avoided.
- The most important features of the implementation model of the programme can be captured, enabling delivery that adheres to the original model and helps to prevent programme drift during maturation and scaling.

Source: Adapted from Ghate (2018[42]).

Although this report focusses on standards of evidence as they apply to discrete interventions, many of the concepts relevant to theory of change are also relevant to the discussions about policy evaluation, and around results-oriented policies, such as the importance of clearly distinguishing between concerns of input, output, outcome/result and impact (European Commission, 2011[43]; Gaffey, 2013[44]) For example, the EU Cohesion Policy (European Commission, 2018[45]) sets out several important changes in the

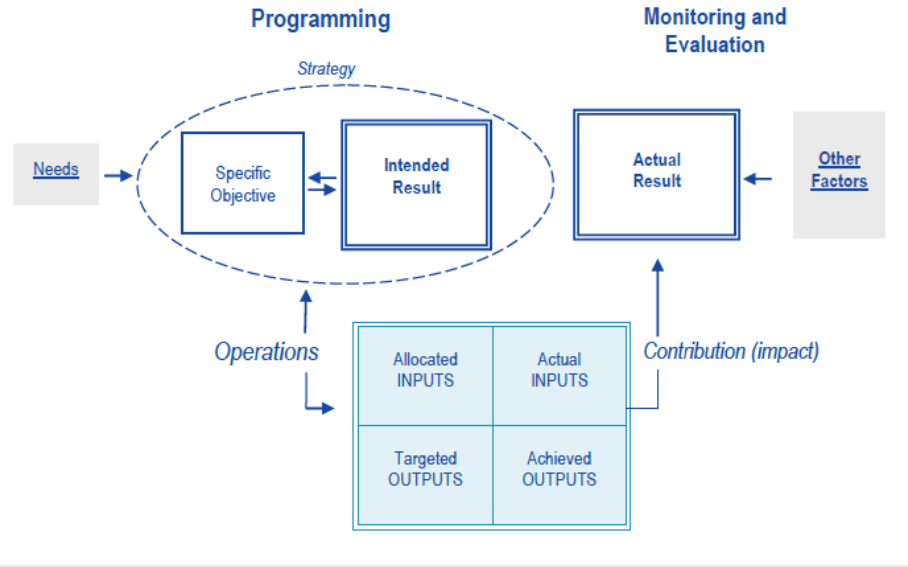
understanding and organisation of monitoring and evaluation, notably the emphasis on a clearer articulation of policy objectives (see Box 3.7).

Box 3.7. Guidance Document on Monitoring and Evaluation – European Commission

The European commission under the Directorate General for Regional and Urban Policy presented the *Guidance Document on Monitoring and Evaluation* to facilitate the implementation of programmes and to encourage good practices. Particularly, the first section: *Intervention logic of a programme as starting point. Results and result indicators* highlights key concepts and terms of programming, monitoring and evaluation to form the basis of practical applications in the policy cycle. (European Commission, 2018[45]).

The following graph (see Figure 3.4 illustrates their framework. As a starting point, a problem is addressed *(need*); following by the *intended result* that motivates policy action (dimension of well-being and progress for people). Once an *objective* of a policy has been chosen, they propose appropriate measures (*result indicators: responsive to policy, normative, robust, and timely collection of data*) to facilitate a later judgement of the policy. The specific activity of programmes leads to *outputs* (direct products of programmes), which are intended to contribute to results. Finally, the approach suggests identifying *other factors* that can drive the intended result towards or away from the desired result, and by which policymakers should be aware.

Figure 3.4. Monitoring and Evaluation framework



Source: European Commission (2018[45])

Finally, the guideline suggests that a representation of a programme should reflect that an intervention can lead to several results and that several outputs can lead to those changes. Similarly, it recommends differentiating the result(s) by affected groups and time horizons. The guideline also offers example for the use of result indicators and a list of common indicators used for Regional and Urban Policies.

Sources: Adapted from the European Commission (2018[45]), Guidance document on Monitoring and Evaluation.

Summary of the mapping of existing approaches

Of the approaches included in the mapping, around half include some coverage of either intervention theory of change or logic model.

Standards concerning theory of change

All of the approaches stipulated that the intervention should be underpinned by a theory of change, but the approaches vary in terms of how rigorous the theory of change must be to meet the required standard. For example, the Level 1 standard from Project Oracle requires a theory of change and an evaluation plan (2018[46]). Similarly, the Level 1 standards from Nesta (2013[47]) stipulates that the intervention should specify what it does and why it matters in a logical, coherent and convincing way. The Nesta standards identify that this standard represents a low threshold, appropriate for early stage innovations, which may still be at the idea stage.

Around half of the approaches go further in stipulating that the theory of change needs to be explicitly based on scientific theory and/or evidence. These approaches include the Canadian Best Practices Portal, the EU-Compass for Action on Mental Health and Well-being, Blueprints and SUPERU. Blueprints, for example stipulates that for an intervention to meet the 'Promising Programs' category it must clearly identify the outcome the programme is designed to change and the specific risk and/or protective factors targeted to produce this change in outcome (Blueprints for Health Youth Development, 2015[48]). The EU-Compass for Action on Mental Health and Well-being (2017[49]) stipulates that for an intervention to be considered 'evidence and theory based' it must be built on a well-founded programme theory which is evidence based, with the effective elements in the intervention stated and justified.

A number of standards go further in providing detailed criteria against which an intervention's theory of change could be rated. These criteria also facilitate comparisons between different interventions according to the quality of their theory of change. These approaches include the Early Intervention Foundation (2018[32]), the Green List Prevention (2011[50]), and the EMCDDA's European drug prevention quality standards (2011[9]). For example the Green List Prevention has a number of criteria for Conceptual Quality described in Box 3.8.

Box 3.8. The Green List Prevention – Conceptual Quality Criteria

- There is a theoretically well-defined model of the programme's effectiveness, the assumed underlying mechanisms have been defined clearly (based on scientifically recognized theoretical models).
- The methods and instruments applied are theoretically well grounded.
- There is a strong logical relationship between analyses of the problem – malleable factors – goals – target groups – used method.
- The programme is targeted at research-based risk- and protective factors
- The target group(s) are described comprehensively and precisely.
- Instructions for implementation and manuals are clearly deduced from the model.
- Goals are defined explicitly and are measurable
- Unless the programme was developed in Germany, the original context and the adaptations made are described.

Source: Adapted from Green List Prevention (2011[50])

A small number of approaches turn these criteria into a numerical scale against which a theory of change is assessed. There are differences in the approaches taken according to whether the approach looks at the evidence underpinning a discrete intervention, such as the Office of Juvenile Justice and Delinquency Prevention Model Programs Guide or whether they assess systematic review evidence underpinning a practice such as the What Works Centre for Crime Prevention (2017[51]). Further details of these two contrasting approaches are in Box 3.9.

Box 3.9. Two approaches to grading the quality of a theory of change

Office of Juvenile Justice and Delinquency Prevention Model Programs Guide

The Model Programs Guide includes a category 'theoretical base' which measures the degree to which the programme is based on a well-articulated, conceptually sound programme theory–it should explain why the programme should effect change. Acceptable programme theory may be articulated or implicit. The programme should provide an explanation of why and how it is expected to achieve its intended results. It uses the following numerical scale:

- 3 - A well-articulated programme theory is clearly defined and sound; previous empirical work related to the theory is described; and there is an explanation about how the theory relates to the specific programme components and how this should result in change for the participants.
- 2 - A programme that defines and describes an empirically supported theory, but does not necessarily connect the theory with the specific components of the programme or provide a theory of change.
- 1 - A programme that provides very little information on programme theory–perhaps referring to a theory but not describing it, referencing prior empirical support, describing the theory of change, or tying it into programme components.
- 0 - A programme that does not mention a theory; or a programme based on a theory known to be unsound.

Crime Reduction Toolkit

The Crime Reduction Toolkit is based on assessments *of systematic reviews* using the *EMMIE framework*. One of the dimensions of the framework is '**Mechanism/Theory of change** *dimension* focuses on the following questions:

- How is the intervention *presumed* to work and *impact* on the outcome?
- What is *needed* to make it work?
- What are the *anticipated impacts* on crime or on behaviour?

To distinguish its *quality*, the dimension *(MechanismQSCORE)* is scored using the following scale:

- ***0 (No information)*** - No reference to theory - simple black box
- ***1 (Limited quality)*** - General statement of assumed theory
- ***2 (Moderate quality)*** - Detailed description of theory - drawn from prior work
- ***3 (Strong quality)*** - Full description of the theory of change and testable predictions generated from it.
- ***4 (Very strong quality)*** - Full description of the theory of change and robust analysis of whether this is operating as expected.

Source: Adapted from Green list prevention (2011[50]) and Crime Reduction Toolkit (2017[51]).

Standards concerning the logic model underpinning an intervention

Only some of the approaches stipulate that the theory of change should be accompanied by a logic model. The majority stipulate that a logic model is necessary but do not provide detailed guidance on what it should contain. For example, SUPERU (2017[52]) has a category for pilot initiatives which have '*a plausible and evidence-based logic model or theory of change that describes what the intervention is, what change it hopes to achieve and for whom, and how the intervention is supposed to work (how its activities will cause change)'.*

A small number of standards go further in providing detailed criteria against which a logic model could be assessed. These include the Society of Prevention Research Standards (2015[53]) which stipulate that they must be described at a level that would allow others to implement/replicate it, including the content of the intervention, the characteristics and training of the providers, characteristics and methods for engagement of participants, and the organisational system that delivered the intervention. The EMCDDA's European drug prevention quality standards also provide very detailed criteria concerning logic models and the description of the intervention described in Box 3.10.

Box 3.10. EMCDDA's European drug prevention quality standards – intervention description

Defining aims, goals and objectives

1. It is specified what exactly is being 'prevented'.
2. The programme's aims, goals, and objectives are specified.
3. Aims, goals, and objectives are dependent on each other and form a logical progression.
4. Aims, goals, and objectives are:
    a. informed by the needs assessment;
    b. useful for the target population;
    c. clear, understandable, and easy to identify;
    d. formulated in terms of expected change in participants ('outcomes');
    e. in accordance with professional and good ethical principles.
5. Goals and objectives are specific and realistic.
6. Specific and operational objectives are distinguished.

Defining the setting

1. The setting(s) for the activities is (are) described.

The setting:

a. is most likely to produce the desired change; i.e. it is relevant to the target population.
b. matches the programme aims, goals, and objectives;
c. matches the available resources.

2. Necessary collaborations in this setting are identified.

The setting:

a. matches identified risk and protective factors;
b. matches likely activities with the target population;
c. makes participants feel comfortable.

Source: An abbreviated list of the European drug prevention quality standards (2011[9])

The Office of Juvenile Justice and Delinquency Prevention Model Programs Guide was unique amongst the approaches in turning the detailed criteria into a numerical scale against which the programme logic model could be assessed as described in Box 3.11.

Box 3.11. Office of Juvenile Justice and Delinquency Prevention Model Programs Guide

Programme Description criteria

The Programme Description rates the degree to which the programme details are described. A full and thorough description should serve as a guide for the implementation of the programme. It should include the following information:

- the logic of the programme,
- the details of all key components,
- the frequency and duration of the programme activities,
- the targeted population,
- the targeted behaviour(s) (i.e., the intent of the programme),
- the setting.

The rating should reflect the degree to which the provided materials afford an adequate programme description and/or direct the reader to references containing such a description and is assessed using the following numerical scale.

a. 3= All programme details are specified (5-6 items are described).
b. 2= Most programme details are specified (3-4 items are described).
c. 1= Some programme details are specified (1-2 items are described).
d. 0= No programme details are specified.

Source: Adapted from Crime Solutions (2013[54]) Program Scoring Instrument Version 2.0

Key questions

- Is there a theory of change that explains how and why the intervention should work?
- Does the logic model clearly specify the assumptions underlying the intervention?
- Does the logic model specify the target population, intervention's resources and inputs?
- Does the logic model clearly describe the intervention's activities and outputs, as well as identify the short and long-term outcomes?

Design and Development of Programmes

Standards concerning the design and development of policies and programmes focus on evidence that tests the feasibility of delivering a policy in practice. At the design and development stage, analysts are often doing important work in testing theories of change and logic models, carrying out process evaluations and pre/post studies.

Why is it important?

Most of the approaches at this stage do not attempt to assess the casual impact of an intervention. Instead, standards concerning design and development aim to identify promising interventions that may be suitable or merit further investigation, at a later stage, for efficacy testing. Efficacy studies (discussed in the next chapter) are complex, time-consuming and expensive to carry out, especially where the collection of new data is required. Therefore, feasibility and pilot studies are an important way of providing information with which to make programme refinements and to inform the design of efficacy studies.

Summary of the mapping of existing approaches

Thirty approaches recognise a phase of design and development of policies and programmes. Most of the approaches categorize these interventions using descriptions such as “emerging”, “delivery and monitoring”, “exploration and development”, or “probable effectiveness”. The approaches can be divided into two broad categories, those establishing the feasibility of an intervention and those focused on piloting the outcomes of the intervention.

Standards for establishing the feasibility of an intervention

A feasibility study typically evaluates whether a range of activities in an intervention, or key components of an intervention’s logic model – including its resources, activities and population reach– are practical and achievable (See Box 3.12). This allows researchers to investigate whether an intervention can work by systematically testing the intervention’s progress towards its intended outputs as it is being implemented (Early Intervention Foundation, 2019[55]).

Feasibility studies can use a variety of quantitative methods (to determine whether the intervention is reaching its delivery and recruitment targets), and qualitative research (to understand the views of the intervention’s recipients and whether these views are consistent with the intervention).

Box 3.12. Feasibility: Can an intervention work?

The Early Intervention Foundation – EIF (UK) presents 10 steps for evaluation success. One of the first steps is to conduct a feasibility study in order to verify if an intervention can really work, and to know under what conditions and with what resources. Key elements include:

- Specifying the intervention's core activities and identify the factors that support or interfere with their successful delivery.
- The use of qualitative research methods to understand which factors contribute to the success of the intervention from the perspective of those delivering it.
- The use of qualitative methods to understand how those receiving the intervention perceive the intervention's benefits and whether these perceptions are consistent with the intervention's original theory of change.
- Building an understanding of how best to recruit and retain participants.
- Developing systems for monitoring participant reach and core delivery targets.
- Applying methods for verifying user satisfaction.
- Tracking and documenting intervention costs.

Source: Adapted from Early Intervention Foundation (2019[55]).

Some of the approaches that recognize qualitative methods at this stage are SUPERU (NZ), What Works for Health (USA), and the Agency for Healthcare Research and Quality (USA). For instance, SUPERU (2017[52]) includes personal experiences from individuals participating in the intervention, such as: interviews, case studies, and ethnographic research. What Works for Health (USA) recognizes studies that describe the intervention, and studies that ask respondents or experts about the intervention (e.g. descriptive, anecdotal, expert opinion). Finally, the Agency for Healthcare Research and Quality (USA) (2012[56]), includes non-comparative case studies or anecdotal reports in its "suggestive" category.

Standards for piloting the outcomes of an intervention

A pilot study is a preliminary and often small-scale investigation conducted to assess the feasibility of the methods to be used in a larger and more rigorous evaluation study. These studies may also focus on which measures are most appropriate for testing the target outcomes (Early Intervention Foundation, 2019[55])

Research and Evaluation design

A variety of different approaches are used to provide preliminary support for programme outcomes. These include administrative data, pre-post-test design and correlational analysis. Most of the approaches agree on implementing pre-post-test at this stage such as Project Oracle (UK) (2018[46]), the European Platform for Investing in Children (EU), and What Works for Health (USA). For instance, the European Platform for Investing in Children (2017[57]) recognizes evaluations using at the minimum pre/post design with appropriate statistical adjustments, and What Works for Health (2010[58]) considers studies comparing outcomes before and after an intervention, and with a statistical analysis. The Green List Prevention (Germany), includes benchmark or non-references-studies (Groeger-Roth and Hasenpusch, 2011[50]) which is described in detail in Box 3.13.

Box 3.13. Standards concerning Design and Development: Green List Prevention

The Green list prevention was established by the Crime Prevention Council of Lower Saxony (Germany) to provide an inclusive rating criterion for the Communities That Care Programme – Databank.

The Rating of the Prevention Programmes is divided in three levels:

- Level 1: Effectiveness theoretically well-grounded
- Level 2: Probable Effectiveness (promising)
- Level 3: Proven Effectiveness (effective)

Within levels, there is a Rating of the Evaluation Studies from "No statement on effectiveness possible" to "five stars". Within Threshold Level 2 to Threshold Level 3, the criteria focuses on conceptual, implementation and evaluation quality in the following order:

No Star

- Participant-satisfaction assessment
- Pre-post assessment without control-group
- Goal-attainment study
- Quality-assurance-study

One star

- Benchmark / Norm-reference-study
- Theory of Change – study

Two stars

- RCT or QED with or without follow-up (not in routine context)
- Pre-post assessment with control-group(s) in routine context

Finally, an Assessment of the Conclusiveness of Evaluation Results is given regarding the previews rating. If the evaluation design obtained:

- No star = no conclusiveness
- One to two stars = preliminary

Source: Adapted from Groeger-Roth and Hasenpusch (2011[50]).

Sample

Design and development standards consider several recommendations regarding the study sample, including its representativeness, the sample design, the sample size, and processes for dealing with study drop out.

Representativeness of the sample. Most of the approaches demand study samples that accurately represent the target population and will be relevant to the research question.

Sampling approach. Most approaches specify that the sampling approach should be well-defined and mention its restrictions. For instance, the Housing Associations' Charitable Trust (UK) (2016[59]) mentions

that the sampling design should include the setting and location where the data are planned to be collected; and a comprehensive description of the eligibility criteria used to select the study participants and the recruitment methods.

Sample size. Some of the approaches specify a minimum sample size threshold required for the research, but the threshold can be set at different sizes. For instance, the European Platform for Investing in Children (EU) (2017[57]) requires a sample size of at least 20 in each study group. Another example is Project Oracle (2018[46]), which considers that a reasonable sample is at least 30 individuals.

Study drop out. Most of the approaches recognise the relevance of study drop out but most of them do not specify any rate beyond which the strength of evidence is compromised. For example, the Clearinghouse for Labor Evaluation and Research (USA) asks if the researchers took steps to reduce study drop out to resolve these issues (2014[60]). Only a few approaches highlight acceptable rates of drop out. In the EIF (UK) (2018[32]) recommends that overall study attrition should not be higher than 40% (i.e. with at least 60% of the sample retained).

Measurement

Most design and development standards stipulate that an evaluation must use valid and reliable measurements. There are some differences between the specifications across standards about how to specify the validity, reliability, and the independence of measurement. See further information in Box 3.14.

Box 3.14. Definitions: Reliability and validity

Reliability refers to the stability of measurement over a variety of conditions—such as across two measure time point—in which the same results should be obtained. Measurement tools must have established test-retest reliability to provide an accurate estimate of pre-and post-intervention change. Issues with reliability occur when measurement items are ambiguous or are subject to fleeting changes in study participants' moods.

Validity is concerned with the extent to which measurement accurately corresponds to the real world. This means that a measurement tool measures what it claims to measure. Establishing measurement validity often involves psychometrically testing to verify the internal, structural validity of measures as well the consistency with other measures.

Source: Adapted from Scholtes, Terwee and Poolman (2010[61]), Drost (2011[62]), and Early Intervention Foundation (2019[55]).

Reliability and validity. Most of the approaches stipulate that measurements should be valid and reliable measures of an outcome. For instance, SUPERU (2017[52]) specifies that the evaluation should use valid and reliable methods and measurement tools that are appropriate for participants and relevant to what the intervention is trying to achieve (See Box 3.15). Project Oracle (2018[46]) also stipulates that valid and reliable measurement tools have been used that are appropriate for the participants in the research.

Other standards provide further specification on the technical requirements that the measurements should meet, such as the European Drug Prevention Standards (2011[9]), which requires that measures should demonstrate internal consistency, test-retest, inter-rater reliability; and construct validity.

Box 3.15. Standards concerning Design and Development: Social Policy Evaluation and Research Unit- Superu

Superu was a government agency from New Zealand focused on what works to improve the lives of families, children and whanau[1]. Part of its goals was to provide independent *standards of evidence* for evaluation guidance, funding support, and to promote the development and uptake of evidence-based interventions.

In that direction, *Superu* presented a **Strength of evidence scale** that consists in an ascending ranking for *quality* of evidence, and brings outlooks about the *type of evidence* that can and should be generated about an intervention as it *matures and grows*.

The scale comprises five levels:

- Pilot initiative
- Early stage, good in theory
- Progressing, some evidence
- Good evidence, sufficient for most interventions
- Extra evidence for large or high risk interventions

By Level 2, an intervention should be in operation for around 1 to 3 years. It has met all level 1 criteria, and been evaluated at least once, but it may not yet be possible to directly attribute outcomes to it.

The intervention is evaluated by the following criteria:

Effectiveness:

- How efficient was the intervention (the delivery of outputs in relation to inputs)?
- The evaluation used a convincing method to measure change: pre- and post-analysis, or a recognised qualitative method. Not necessary with control group.
- The evaluation used valid and reliable methods to measure what the intervention is trying to achieve.
- Was data was used properly?

Intervention consistency and documentation

- There is clarity and documentation about what the intervention comprises.
- Manuals and staff training processes.
- Resources (money and people) required to deliver the intervention.

Sources: (SUPERU, 2017[52]).

Independence from the intervention. Some approaches request independency of a measurement from participants and data collecters. For instance, Clearinghouse for Labor Evaluation and Research (USA) (2014[63]) indicates that data collection must reflect methods that produce unbiased results such as independency and objectivity of the outcomes from the research team. The European Drug Prevention Standards (EU) (2011[9]) also agrees that measures must produce results independently of who uses the instrument.

Approach to analysis

Design and development standards highlight the importance of well executed and described analysis, which covers: the data collection, hypothesis testing, and methods of address missing data or other sources of bias.

Data collected. Most approaches stipulate that a complete report should be able to explain and justify why and how the analysis was conducted. For instance, Housing Associations' Charitable Trust (2016[59]) requires the study protocol, recording of any deviations, and a structured report of findings.

Hypothesis testing. Most of the approaches stipulate a clear description of the analysis methods selected to test the research question. For example, Clearinghouse for Labor Evaluation and Research (USA) (2014[63]) demands analysis methods that are very well-described, relevant to the research question, sufficiently rigorous, and correctly executed.

Missing data. Many of the approaches consider issues of missing data, with most requesting that the analysis specifies how these issues were managed, and how this could affect the interpretation of the findings. For instance, Project Oracle (2018[46]) asks if the research provides all the details concerning the data analysis, or any weaknesses of the design, and their impact on the results. Another example is European Drug Prevention Standards (EU) (2011[9]), which requests reporting and appropriate handling of missing data.

Findings and conclusions

Design and development standards request coherence between the programme's theory of change, the data analysis, and findings. Some of the approaches go further and specify that findings should be statistically significant on at least one of the outcomes; and not have harmful effect. Other approaches define the findings in this stage as unclear/undetermined effects.

Statistical significance. Among the approaches dealing with quantitative research, there are variations between the information required about statistical significance. Some approaches only recommend that the results were tested for statistical significance, such as Project Oracle (UK) (2018[46]). Other approaches stipulate that the findings must be significant. For instance, What Works for Health (USA) (2010[58]) scores pre/post studies with statistically significant favourable findings higher, and Evidence for ESSA (USA) (Every Student Succeeds Act - ESSA, 2019[34]), which requires findings of a statistically significant effect for correlational studies. Some other approaches require a specific level of significance, such as European Platform for Investing in Children (EU) (2017[57]), which asks for positive results at 10 % of significance.

No Harmful effects. Many of the approaches expect that the intervention does not constitute a risk of harm. For example, The Centers for Disease Control and Prevention suggest that studies should indicate any negative effect. (Puddy and Wilkins, 2011[64]).

Unclear/undetermined effects. Most of the approaches accept unclear or undetermined effects given the type of evidence, and rigorous on the study design. For example, SUPERU (NZ) (2017[52]) mentions that at this stage an evaluation (pre/post study) indicates some effect, but it may not yet be possible to directly attribute outcomes to it. Another example is the Housing Associations' Charitable Trust (UK) (2016[59]), where the lack of a good design limits any conclusion of causality.

Subgroup analysis[2]. Only a few approaches discuss subgroup analysis to verify for whom the effects are claimed. For example EIF (UK) (2018[32]) stipulates that subgroup analysis is used to verify for whom the intervention is effective and under what conditions. The Clearinghouse for Labor Evaluation and Research (USA) (2014[60]) discusses if the sample analysis allows generalizing the results to a wider population, or if it is presented the limitations of this inference.

Key questions

Standards concerning the feasibility of the intervention

- Has a feasibility study been conducted to test whether the intervention can achieve its intended outputs?
- Has the feasibility study led to an understanding of which factors contribute to the success of the intervention from the perspective of those who are implementing it?
- Has the feasibility study led to an understanding of how those receiving the intervention perceive the intervention's benefits and whether they are satisfied with the intervention?

Standards concerning pilot studies

- Has a pilot study been carried out to investigate an intervention's potential for improving its intended outcomes?
- Does the study have a minimum of pre and post measurement to establish a correlation between programme participation and outcomes?

Sample

- Is the study sample representative of the intervention's target population?
- Is the sample design clearly described?
- Does the study give information about dropouts and dose?
- Does the study present information regarding attrition?

Measurements

- Are the measurements used reliable and valid? Does the study report a specific method to do so?
- Are the measurement independent on any measures or information given as part of the intervention?
- Has the evaluation used valid and reliable methods and measurement tools that are appropriate for participants and relevant to what the intervention is trying to achieve?

Approach to analysis

- Does the study stipulate the analysis methods used and are these appropriate for the research question?
- Does the study have procedures for managing issues of missing data?
- *Finding and conclusions*
- Does the study find statistically significant positive results and are there no harmful effects on at least one of the target outcomes of the study?

Efficacy of an Intervention

Once an intervention has been identified as 'promising' in preliminary research, many standards of evidence emphasise the need for rigorous efficacy testing. Efficacy studies typically privilege internal validity, which pertains to inferences about whether the observed correlation between the intervention and outcomes reflect and underlie causal relationship (Society for Prevention Research Standards of Evidence,

2015[53]). In order to maintain high internal validity, efficacy trials often test an intervention under 'ideal' circumstances. This can include a high degree of support from the intervention developer and strict eligibility criteria thus limiting the study to a single population of interest.

Why is it important?

A critical goal of standards of evidence is to facilitate the communication of which policies and programmes are efficacious. A statement of efficacy should be of the form that Intervention X is efficacious for producing Y outcomes for Z population at time T in setting S (Society for Prevention Research Standards of Evidence, 2015[53]). In order to maintain high internal validity, efficacy trials often test an intervention under 'ideal' circumstances, and tell us little about the impact of an intervention in 'real world conditions", because the evaluation is often overseen by the developer of the policy or programme, with a carefully selected example. This can include a high degree of support from the intervention developer and strict eligibility criteria thus limiting the study to a single population of interest.

Therefore, standards of evidence often stipulate that a policy or programme demonstrates effectiveness, in studies where no more support is given than would be typical in 'real world' situations. This requires flexibility in evaluation design to address cultural, ethical and practice challenges. Systematic reviews, observational studies and participatory evaluations which gather attitudinal and experiential considerations from the main beneficiaries can still be considered useful evidence and guide improvements in the design or implementation of the intervention.

Summary of the mapping of existing approaches

Determining the efficacy of an intervention is a complex process, involving considerations on the evaluation design, sample, measurements, methods of analysis, and findings. There is wide variety of specification standards that an evaluation must meet for an intervention to be deemed efficacious.

Evaluation design

All the standards consider Randomized Control Trials (RCTs) as an appropriate study design to generate a counterfactual as the basis for making efficacy claims. However, there is wide variation across standards regarding Quasi-Experimental Design (QEDs). Some approaches consider that QEDs can be used to generate comparable samples as RCTs, whereas other standards only recognise that QEDs are better than pre/post studies.

Sixteen of the approaches privilege the use of RCTs over QEDs. For example, Nest What Works for Kids (Australia) (2012[1]) ranks programmes or policies with well-implemented RCTs in the highest levels. Evidence for Every Student Succeeds Act (USA) (Every Student Succeeds Act - ESSA, 2019[34]) defines a programme or policy as *Strong evidence* when it has at least one well-designed and well-implemented RCT.

Thirty of the approaches consider both RCTs and QEDs as robust evaluation designs to support causal inference. For instance, the European Platform for Investing in Children (EU) (2017[57]) defines both evaluation designs as methodologies that can be used to construct convincing comparison groups to identify policy impacts. Other approaches that treat suitably designed RCTs and QEDs as equivalent are SUPERU (NZ) (2017[52]) and the Green List Prevention (Germany) (Groeger-Roth and Hasenpusch, 2011[50]).

Among the approaches that accept QEDs, some of them distinguish between how rigorous different type of designs are, such as: difference in difference (DD); propensity score matching (PSM); and Regression Discontinuity Designs (RDD). A few approaches also provide a score according to the rigour and limitations of QEDs. For example, What Works Centre for Local Economic Growth (UK) (2016[65]) presents a guide

scoring evidence using the Maryland Scientific Methods Scale to evaluate the different type of designs, from PSM, Panel Methods, DD, RDD to Instrumental variables (IV), see Box 3.16.

Many approaches recognise that, whilst RCTs might in theory present the 'gold standard' in reducing threats to internally validity, in practice randomisation might not be practicable for a range of policy challenges, including ethical concerns. In the health policy area, the famous "Rand experiment", which allowed for computing the price elasticity of the demand for health, could probably not be replicated today. (Newhouse J.P., 1993[66]) . For example, in OECD's work on Regional Development Policy (OECD, 2017[67]), it is recognised that randomisation is not always possible, and quasi-experimental designs can be used as an alternative method to identifying causal effects. In addition, the development of econometric methods with the use of *Difference in Differences* with instrumental variables in econometrics, has helped to diffuse the use of alternative quasi experimental approaches to producing reliable estimates.

Box 3.16. Standards concerning Efficacy: What Works Centre for Local Economic Growth

The *What Works Centre for Local Economic Growth (WWG)* is an independent organisation from the UK, mainly focused on producing systematic reviews of the evidence on a broad range of policies in the area of local economic growth.

WWG assessment is based on the Maryland Scientific Methods Scale (SMS), which ranks policy evaluations from 1 (least robust: studies based on simple cross sectional correlations) to 5 (most robust: Randomised Control Trials.). The ranking aims to present the extent to which the methods deal with the selection biases inherent to policy evaluations (robustness), and the quality of its implementation to achieve efficacy, as following:

- **Level 2:** Use of adequate control variables and either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group.
- **Level 3:** Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. *difference in difference*). Techniques such as regression and *propensity score matching* may be used to adjust for difference between treated and untreated groups, but important unobserved differences are likely to remain.
- **Level 4:** *Quasi-randomness* in treatment is exploited, so that it can be credibly held that treatment and control groups differ only in their exposure to the random allocation of treatment. This often entails the use of an *instrument* or *discontinuity in treatment*, the suitability of which should be adequately demonstrated and defended.
- **Level 5:** Reserved for research designs with *Randomised Control Trials (RCTs)* providing the definitive example. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Additionally. Attention paid to problems of *selective attrition*, and there should be limited or, ideally, no occurrence of 'contamination' of the control group with the treatment.

Source: (What Works Centre for Local Economic Growth, 2016[65]).

Intention to treat (ITT). Although the importance of ITT in the academic field is well-established (Hollis and Campbell, 1999[68]), there is variation within the approaches concerning their treatment of ITT. Some

of them clearly request in their criteria that analysis must be based on ITT. For instance, What Works Centre for Children's Social Care (UK) (2018[69]) establishes that acceptable quality study must have an intent-to-treat design. Social Programmes that Work (USA) (2019[70]), which stipulate an ITT approach for the intervention group, and Child Trends (USA) (2018[71]) affirms that only results based on an intent-to-treat analysis can be reported.

Sample

Most standards present clear conditions regarding the nature of the sample required in providing an appropriate basis for the analysis. The standards specify issues concerning a baseline equivalence, attrition, and risks of contamination.

Baseline equivalence. Some of the standards focus on baseline characteristics of the treatment and comparison-groups before running a programme or policy. For instance, Nest What Works for Kids (AU) (2012[1]) requests clear analysis of baseline characteristics. And Social programmes that Work (USA) stipulates that the intervention and control groups must be highly similar in key characteristics prior to the intervention (Coalition for Evidence-Based Policy, 2010[2]).

Other standards treat baseline equivalence differently according to whether the study design is an RCT or QED. Evidence Based Teen Pregnancy Programs stipulates that an RCT must control for statistically significant baseline differences and QEDs must establish baseline equivalence of research groups and control for baseline outcome measures (Mathematica Policy Research, 2016[33]).

Attrition. Some of the standards recognise an attrition threshold. For instance, European Platform for Investing in Children (EU) (2017[57]), which states that attrition must be less than 25% or that it has been accounted for using an acceptable procedure. Another example is the Clearinghouse for Military Family Readiness (USA) (2012[72]), which stipulates an attrition at immediate post-test, of less than 10%.

Other standards also stipulate conditions for overall and differential attrition. For example, Darlington Service Design Lab requests no evidence of significant differential attrition (Graham Allen, 2011[73]). Other approaches go further in stipulating specific attrition thresholds. For instance, What Works for Clearinghouse (USA) (2020[74]) defines that for studies with a relatively low overall attrition rate of 10%, a rate of differential attrition up to approximately 6% is acceptable. For studies with a higher overall attrition rate of 30%, a lower rate of differential attrition, at approximately 4% is acceptable.

Risk of contamination. Only few standards highlighted the issues around risk of contamination. For example, What Works Centre for Local Economic Growth (UK) (2016[65]) stipulates no occurrence of contamination of the control group for the treatment. Crime Solutions (2013[54]) assesses the degree to which internal validity is threated, within other aspects, by contamination.

Measurements

Some efficacy standards stipulate that evaluations must use valid and reliable measurements. In general, these standards tend to be broadly equivalent to those already discussed at the design and development phase. For example, Blueprints for Healthy Youth Development (2015[48]) demands use of valid and reliable measures, and California Evidence-Based Clearinghouse for Child Welfare (2019[75]) provides a measurement tools rating scale based on the level of psychometrics (e.g., sensitivity and specificity, reliability and validity) in peer review studies using QEDs or RCTs.

A few approaches go further and recommend the independency of the measurement with the participants of an intervention. For instance, EIF (UK) (2018[32]) requests that measurements are blind to group assignment if possible. The European Monitoring Centre for Drugs and Drug Addiction (EU) (2011[9]) specifies that an instrument is objective if it produces results independently of who uses the instrument to take measurements. The Dartington Service Design Lab (UK), stipulates that outcome measures must not

depend on the unique content of the intervention, and they are not rated solely by the person or people delivering the intervention (Graham Allen, 2011[73]).

Approach to analysis

A few of the approaches provide details on the appropriate analysis required in order to establish the efficacy of a policy. These standards focus on establishing baseline conditions, and the analysis of the effects at the correct level of assignment.

Baseline conditions. Most of the approaches require that evaluations use statistical models to control for baseline differences between treatment and control group. For instance, Strengthening Families Evidence Review (USA) (2019[76]) requests statistical adjustment when treatment and comparison groups are not equivalent. Another example is HomeVEE (USA) (2018[77]), which requests that the analysis should control for differences in baseline characteristics and baseline measures.

Level of analysis. Only a few of the standards demand that the analysis needs to be appropriate according to whether the assignment is at the individual or group (or cluster) level. For instance, Evidence Every Student Succeeds Act (USA) (Every Student Succeeds Act - ESSA, 2019[34]) stipulates that clustered designs must use Hierarchical Linear Modelling (HLM), or other methods accounting for clustering. A second example is the Society of Prevention Research (Society for Prevention Research Standards of Evidence, 2015[53]), which specifies that the analysis must assess the treatment effect at the level at which randomization took place.

Findings and impacts

Most of the standards focus on impact effects and their statistical significance whereas other standards that also request effect size measures.

Statistical significance. Across the standards, the majority demand information regarding whether effects are statistically significant. For instance, Clearinghouse for Military Family Readiness (2012[72]) in the Promising Programme Category requests specific conditions for significant Effects—Two-tailed tests of significance are preferable to one-tailed tests.

Impact effects. Most of the standards claim that an intervention is efficacious when the findings of an intervention are positive and significant, and there is no evidence of harmful effects. Other standards request reporting of mixed effects or null effects. The standards may present these criteria as part of a one ranking; or independently with a ranking solely focused on impact.

- Forty-one of the standards consider positive impact effects to claim efficacy. For instance, Be you (AU) (2020[78])— the new integrated national initiative of the Australian government to promote mental health from early years through evidence-based, flexible online professional learning, complemented by a range of tools and resources to turn learning into action (Early Childhood Australia, 2020[79])—requests that a programme have at least one research or evaluation study which demonstrates a positive impact on mental health outcomes for children or young people.
- Nineteen of the standards also request reporting on whether there are harmful effects. For example, EU-Compass for Action on Mental Health and Well-being (EU) (European Commission - Directorate-General for Health and Food Safety, 2017[49]) requires that the evaluation outcomes demonstrate beneficial impact, and that possible negative effects be identified and stated.
- Eighteen of the standards consider mixed effects or null effects. For example, EIF (UK) (2018[32]) has a Not effect level (NE). This level is reserved for programmes where there is evidence from a high-quality evaluation of the programme that did not provide significant benefits for children.

Some of the standards present an independent ranking or score to assess impact. For instance, Darlington Service Design Lab (UK) evaluates impact according to interventions with positive effect size, and no

harmful effects or negative side–effects of intervention (Graham Allen, 2011[73]). Another example is Evidence Based Teen Pregnancy Programs (USA), which classifies the programme evidence as positive, mixed, indeterminate or negative (Mathematica Policy Research, 2016[33]).

Magnitude of the findings. Some standards recognise the importance of reporting effect size. For example, Blueprints (2015[48]) stipulates that effect sizes should be reported, along with the significance levels of those differences, or that it should be possible to calculate the effect size from the data reported (means and standard deviations).

Other standards establish an effect size threshold. For instance, the European Platform for Investing in Children (EU) (2017[57]) stipulates an effect size of at least 0.1 of a standard deviation. A second example is Best Evidence Encyclopaedia (USA), which assesses an intervention by the sample size and effect size (Johns Hopkins University School of Education's Center for Data-Driven Reform in Education - CDDRE[80]), in the following order:

- *Moderate evidence* level requests specifically studies with weighted mean effect size of at least +0.20;
- *Limited evidence* level a study can meet the criteria except that the weighted mean effect size is +0.10 to +0.19; or the weighted mean effect size is at least +0.20, but the study is insufficient in number or sample size.

Key questions

Sample

- Are the treatment and comparison group assignment at the appropriate level (e.g. individual, family, school, community)?
- Is there baseline equivalence between the treatment and comparison-group participants on demographic variables before running the intervention?
- Does the study provide information about risks of contamination between intervention and control groups?
- Does the study provide information about overall and differential attrition?

Approach to analysis

- Does the evaluation use statistical models to control for baseline differences between treatment and control groups?
- Does the study use an appropriate analysis according to level of assignment (e.g. individual or group)?
- Finding and Impacts
- Does the study present statistically significant positive effects on at least one or more of the primary outcomes of the study and no harmful effects?
- Does the study report the magnitude of the effects (e.g. effect size)?

Effectiveness of Interventions

Efficacy trials often tell us little about the impact of an intervention in 'real world' conditions, because the evaluation is often overseen by the developer of the policy or programme, with a carefully selected sample.

What are the benefits or damages, independently from the policy goals? Therefore, standards of evidence often stipulate that a policy or programme demonstrates effectiveness, in studies where no more support is given than would be typical in 'real world' situations.

Why is it important?

Demonstrating effectiveness of a policy or programme in 'real world' situations requires flexibility in evaluation design to address cultural, ethical and practice challenges. During policy implementation, evidence is useful to understand for whom it works and for whom it does not work. Therefore it is important to learn how to maximize benefits and minimize damages, also within a no policy change scenario.

Summary of the mapping of existing approaches

For the majority of standards, in order for an intervention to claim effectiveness, the evaluations should meet all of the conditions of efficacy studies discussed in the previous section as well as the following criteria: generalizability of the findings, long term impacts, positive average effect across studies and no reliable iatrogenic effect observed on important outcomes.

Evaluation design

Generalizability. In order to translate the findings of efficacy evaluations into a wider range of population and settings, the standards concerning effectiveness stipulate that the generalizability of intervention effects should be tested across the following dimensions: a replication; and population subgroup analysis.

Replication. Most of the standards typically request two or more RCTs or QEDs conducted in different locations. Some of the standards consider that before an intervention is judged as effective and ready for scaling up data collection and analysis should be carried out by an independent evaluator who does not have any involvement with the developer of the intervention. For instance, the Clearinghouse for Military Family Readiness (USA) (2012[72]) requests at least one replication involving an external implementation team at a different site. The National Dropout Prevention Center (USA) (2019[81]) gives the highest score to programmes that were evaluated using an experimental or strong quasi-experimental design conducted by an external evaluation team. Further details on the approach adopted by Blueprints are described in Box 3.17.

Other standards do not necessarily request or establish any "independency" condition for a replication. Most of them specify the number of evaluations of the intervention and attention to the transferability of the findings to different context. For instance, CEBC (USA) (The California Evidence-Based Clearinghouse for Child Welfare, 2019[75]) requests at least two RCTs in different settings. Another example is Education counts (NZ), which considers the degree of applicability to New Zealand contexts; and specificity or generalisability of findings (Alton-Lee, 2004[10]).

Population subgroup analysis. Other standards explore generalizability through an analysis of the population subgroups (e.g. race, gender, social class). For instance, Society for Prevention Research Standards of Evidence (2015[53]) requests a statistical analysis of subgroup effects for each important subgroup to which intervention effects are generalized. SUPERU (NZ) (2017[52]) asks for evidence of the impact of the intervention on different subgroups in the target population.

Sample

The standards concerning effectiveness assume that the studies have the same robust approach about the sampling, as already was specified in the efficacy section.

Measurements

Most standards stipulate that effectiveness evaluations must meet the same requirements for measurement as in efficacy standards. Some of them request additionally the independency of the measurements from the participants and from the person delivering the intervention. For instance, EIF (UK) (2018[32]) request that at least one evaluation use a form of measurement that is independent of the study participants and independent of those who deliver the programme.

Box 3.17. Standards concerning Effectiveness: Blueprints

Blueprints is a project from USA funded by the Annie E. Casey Foundation, and hosted by the Center for the Study and Prevention of Violence (CSPV) to make available a registry of evidence-based positive youth development programmes on the health and well-being of children and teens.

The programmes are ranked under the following standards of effectiveness:

Promising programs:

- Intervention specificity: Clarity on the objective outcome, risk and/or factors expected from the study, target population, and components role to produce a 'change'.
- Evaluation quality: The programme has been evaluated by (a) at least one randomized controlled trial (RCT) or (b) two quasi-experimental (QED) evaluations.
- Intervention impact: There is a significant positive change in planned outcomes that can be attributed to the programme and there is no evidence of harmful effects.
- Dissemination readiness: organizational capability, manuals, training, technical assistance and other support required for implementation with fidelity in communities and public service systems.

Programmes meet the minimum standard of effectiveness and are recommended for local community and systems adoption.

Model programmes - meet these additional standards:

- Evaluation Quality: A minimum of (a) two high quality randomized control trials or (b) one high quality randomized control trial plus one high quality quasi-experimental evaluation.
- Positive intervention impact is sustained for a minimum of 12 months after the programme intervention ends.

Model Plus programmes – provide one additional standard:

- Independent Replication: In at least one high quality study, demonstrating desired outcomes, authorship, data collection, and proper analysis.

Model and Model Plus programmes meet a higher standard, provide greater confidence in the programme's capacity to change behaviour and developmental outcomes, and are recommended for large-scale implementation.

Source: Adapted from Blueprints for health youth development _(2018[82])

Approach of analysis

Most of the standards stipulate that at effectiveness evaluations must meet all the methodological requirements previously discussed in the efficacy standards including appropriate statistical analysis (e.g. Intent-to-treat) and baseline equivalence adjustments.

Finding and impacts

Most of the standards concerning effectiveness agree on requiring positive average effects and no evidence of negative effects or risk of harm. Other standards also consider the sustainability of the effect at the long term.

- **Positive average effect and no reliable iatrogenic effect observed.** Most of the standards agree on requesting positive average effect across studies and reporting no reliable iatrogenic effects observed on important outcomes. For instance, Society for Prevention Research Standards of Evidence (2015[53]) specifies that effectiveness can be claimed only for intervention populations, times, settings, and outcome constructs for which the average effect across all effectiveness studies is positive and for which no reliable iatrogenic effect on an important outcome has been observed. Another example is Dartington Service Design Lab (UK), which requests evidence of a positive effect and an absence of iatrogenic effects from the majority of the studies (Graham Allen, 2011[73]).
- Other standards adjust the programme's rating according to the average effects of the studies. These standards recognize a category for each of the possible results found in multiple studies. For example, CEBC (USA) (The California Evidence-Based Clearinghouse for Child Welfare, 2019[75]) presents three categories for the overall weight of evidence from several studies:
  - *Well supported category:* At least two RCTs have found the benefit of the practice;
  - *Evidence Fails to Demonstrate Effect:* Two or more RCTs have found the practice has not resulted in improved outcomes;
  - *Concerning Practice:* The overall weight of evidence suggests the intervention has negative effect.
- **Long term effects**. Most of these approaches agree on requesting sustained effects for at least 12 months. A few of them also accept effects for at six least months. For example, the Nest What Works for Kids (AU) (2012[1]) consider different periods of time: Supported level, the effect should be maintained at a 6-month follow-up; and for Well supported an effect must be maintained for at least one study at one-year follow-up. Another example is HomVEE (USA) (2018[83]), which evaluates the evidence across diverse outcome domains, such as duration of Impacts (information on the length of follow-up) and Sustained Impacts (impacts were measured at least one year after programme enrolment).

A postscript on systems-based approaches to evaluation

One important caveat to the standards reviewed in the efficacy and effectiveness chapters is that they primarily originate from traditional approaches to impact evaluation. These can be contrasted with systems-based approaches to evaluation (see Table 3.1). System based approaches to evaluation start from challenges faced when dealing with the open-ended nature of problems and issues including innovation and the goal complexity of the connected processes (Askim, Hjelmar and Pedersen, 2018[84]; Tõnurist, 2019[85]).

OECD has been also moving towards a systems approach to public sector innovation and has developed a model to look at innovation activities from an individual, organisational and systemic lens, which can then also feed into approaches to evaluation (Tõnurist, 2019[85]). The tensions between the traditional

approaches impact evaluation and the system based approaches has been further discussed by Tõnurist (2019[85]) and it is acknowledged that integrating these insights would be a useful next step for approaches to standards of evidence.

Table 3.1. Two Basic Approaches to Evaluate Public Sector Innovation

	Impact-focused approach	Systems approach
Purpose	Solid knowledge about effects and causality	Generate new practices; nurture innovation and learning
Participants	Controlled selection, few in number, randomisation, strategic sampling, control group	Self-selection, many participants, no control group
Contents	Narrow scope and accommodation	Wide scope; emergent, liberal accommodation
Evaluation	Comprehensive, external party, effect-oriented	Inspiring depictions, context-sensitive narratives, process-oriented

Source: Adapted from Tõnurist (2019[85]) and Askim, Hjelmar and Pedersen, (2018[84])

Key questions

Evaluation design

- Has the programme been evaluated in more than one location?
- Has at least one evaluation been carried out independently from the developer of the intervention?

Measurements

- Are the measurements tools used independently from the participants and the people involved in delivering the intervention? (e.g. administrative data, independent observer)

Findings and impacts

- Has the programme's impacts been analysed by population sub-groups?
- Are the effects of the intervention sustained over the medium or long term?

Cost (effectiveness) of Interventions

Measuring effective interventions requires not only evidence of their impacts, but evidence of their cost and value for money. Cost data provide information relevant to the financial planning and sustainable scale-up (Levin and Chisholm, 2016[86]); while a variety of methodological tools look to assess the benefits and costs associated with an intervention.

Why is it important?

Positive impacts at a very high price may not be in the interests of governments and citizens. Using economic evidence is important to demonstrate value for money for public programmes in a context of continued fiscal constraints. Increased understanding of interventions that achieve impact at a too high price would enable decision-makers to make more efficient decisions.

Summary of the mapping of existing approaches

A variety of different methodologies is taken by the existing standards. Some of them focus on reporting the existence of cost or economic evaluations of an intervention. Other standards request in their criteria the presence of cost information and related analysis for a policy or programme; whilst a final set of standards provide detailed guidance on carrying out and interpreting economic evaluations.

The Box 3.18 provides a first general description of the different types of economic evaluations used to understand their complexity and usefulness when a particular organization or government is taking an investment decision.

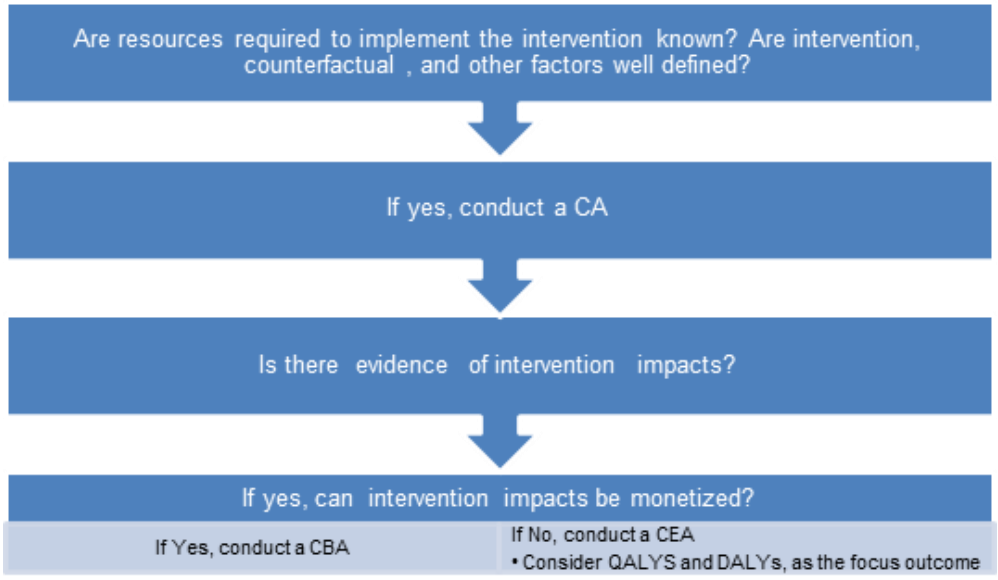
Box 3.18. Type of economic evaluations

- *Cost analysis (CA):* This evaluation can provide a complete account of the economic costs of all the resources used to carry out an intervention.
- *Cost-effectiveness analysis (CEA):* CEA requires that an indicator of effectiveness be compared with a cost from a single project. CEA can only offer guidance on which of several alternative policies (or projects) should be selected, given that one has to be selected. (OECD, 2018[87]).
- *Cost-utility analysis (CUA):* when alternative interventions produce different levels of effect in terms of both quantity and quality of life (or different effects), the effects may be expressed in utilities. The best known utility measure is the quality-adjusted life year, or QALY.
- *Cost-benefit analysis (CBA):* provides a model of rationality (or human preference). It forces the decision-maker to look at beneficiaries and losers–in both, spatial and temporal dimensions– in monetary values (OECD, 2018[87]).
- *Return on investment (ROI):* captures the percentage of return for every dollar invested. It is calculated by dividing intervention benefits net of intervention costs by intervention costs.
- *Multi-criteria analysis (MCA):* looks at manifold and diverse dimensions of policies and investment projects through the same analytical framework. MCA offers a broader interpretation of CEA since it openly features the existence of multiple objectives (OECD, 2018[87]).

Source: (The Cochrane Collaboration, 2011[88]; National Academies of Sciences, Engineering, and Medicine, 2016[89]).

In the same line, the National Academies of Sciences, Engineering, and Medicine provides a decision tree to determinate if an intervention is ready for an economic evaluation (See Figure 3.5). According to this tool, the first step corresponds to reviewing the available information about an intervention to determine if it is enough to answer the research question (e.g. are the counterfactuals well defined? are resources required to implement the intervention known?). At this stage, a researcher or policy maker should be able to conduct a CA. If there is evidence of intervention impacts, the researcher should consider conducting a CEA or CBA, which also relies on whether the interventions' impacts can be monetarized. If they can be, the researcher should conduct a CBA; otherwise, a CEA would be the best option. Other economic evaluations can be in consideration such as QALY, or DALY. This will depend on the perspective of the evaluation.

Figure 3.5. Decision tree for the different types of economic evaluation



Source: Adapted from the National Academies of Sciences, Engineering, and Medicine (2016[89]).

Cost information

A first step towards being able to carry out economic evaluation is to collect information on the costs of an intervention; most of the approaches (31 of 50) report information in term of the economic resources in materials and training in an intervention. For instance, Clearinghouse for Military Family Readiness (USA) (2012[72]) presents in the evidence summary information related to the cost of training per participants, and further available information related to the programme implementation.

Other approaches report in the evidence reviews if an intervention or programme has developed an economic evaluation. For example, What Works for Kids (Australia) (2012[1]) asks specifically if a cost benefit study has been undertaken and published. This information is also identified in the evidence portal of What Work Centre for Children's Social Care (UK) (2018[69]). A further example is Social Program that works (USA), which provides summaries of the available programme's benefits and costs in their evidence reviews.

Cost rating or assessment

Some of the approaches establish in their criteria a cost rating or assessment condition, separately from the quality design, to assess if an intervention provides cost information, or the settings to present this type of information. For instance, EMCDDA (EU) (2011[9]) suggests planning financial requirements in terms of cost estimations for the programme, and a detailed and comprehensive breakdown of costs. Other approaches provide tools to determine the resources required in future interventions or additional activities, such as EEF (See Box 3.19) and the Toolkit produced by the What Works Centre for Crime Reduction (UK) (2017[90]), which distinguish direct and indirect costs, and allow users to make a comparison of costs prior to the intervention being implemented and after the intervention, or in a different context.

Box 3.19. Cost rating bands for EEF Toolkit

The Education Endowment Foundation (UK) implements a cost rating reflecting the approximate additional cost per pupil of implementing such interventions in schools. This might include the cost of new resources required, additional training courses, and activities. The cost rating bands are organized as following:

- £ - Very low: up to about £2,000 per year per class of 25 pupils, or less than £80 per pupil per year.
- ££ - Low: £2,001 £5,000 per year per class of 25 pupils, or up to about £200 per pupil per year.
- £££ - Moderate: £5,001 to £18,000 per year per class of 25 pupils, or up to about £720 per pupil per year.
- ££££ - High: £18,001 to £30,000 per year per class of 25 pupils, or up to £1,200 per pupil.
- £££££ - Very High: over £30,000 per year per class of 25 pupils, or over £1,200 per pupil.

Source: Education Endowment Foundation (2018[25]).

Whereas few of the approaches requesting a Cost Benefit Analysis or Cost Effectiveness Analysis in their criteria. For instance, EPIC (EU) (2017[57]) assesses if the programme has been found to be cost-effective/cost-beneficial (i.e. the practice can deliver positive impact at a reasonable cost).

Standards concerning economic evaluations

To determine how best to invest public or private resources in social policies, decision makers require the use of economic evaluations to answer relevant questions, such as What does it cost to implement this intervention in a particular context and what are its expected returns? To what extent can these returns be measured in monetary or nonmonetary terms? Who will receive the returns and when? Is this investment a justifiable use of scarce resources relative to other investments? (National Academies of Sciences, Engineering, and Medicine, 2016[89]).

In this section additional standards that are mainly focused on establishing guidelines to evaluate economic evaluations will be introduced, particularly, with regard to criteria relating to cost analysis (CA), cost effectiveness analysis (CEA), and cost benefit analysis (CBA).

In the following sub-sections the standards for framing the evaluation will be outlined: identifying the impacts, determining the cost, valuing benefits and cost, and presenting the results.

Standards for framing an economic evaluation

Most of the approaches stipulate the importance of clearly stating the objectives of an economic evaluation regarding the information and resources available for a given intervention in order to establish which evaluation method to use. For instance, The National Academies of Sciences, Engineering, and Medicine (2016[89]) suggest that in order to determinate whether an intervention is ready for economic evaluation, this will depend on the question(s) of interest, the intervention specificity, and a well specified counterfactual condition.

Other approaches specify that the eligibility criteria, delivery setting time, and location must be well described, which relates to some of the issues of theory of change and logic model addressed. For instance, New South Wales Government (2017[91]) refers to the need for a programme logic to identify the

issues that a programme is seeking to address; its intended activities and processes; their outputs; and the intended programme outcomes. Another example is the work on Cost benefit analysis and the environment by the OECD (2018[87]) which specifies that it is necessary to mention all the direct and indirect participants involved in the policy, geographical boundary, and extension to wider limits.

Standards for identifying the programme/policy impacts

Most of the approaches stipulate that the outcomes used in CEA and CBA should come from robust designs to determinate unbiased impacts from an intervention (e.g. RCTs or QEDs). This builds on the issues of efficacy and effectiveness addressed in Section 0 and 0 (Research design, measurements, sample, potential impacts, and external validity). For instance, The National Academies of Sciences, Engineering, and Medicine (2016[89]) stipulate that for CEA or CBA not only is information on the resources used to implement the intervention required, there is also a need for credible evidence of impact. Another example is the Vera institute (USA) (2014[92]), which refers to quantifying the investment's impacts using evaluations that establish the causal link between an investment and its impacts.

Other approaches also highlight the use of meta-analysis or systematic reviews when multiple impact studies exist regarding a programme or similar intervention. For example, the Washington State Institute for Public Policy Benefit (USA) develops a meta-analytic approach to identify, screen, and code research studies in its cost-benefit analysis. The WSIPP also adjusts effect size regarding the methodical quality of the study and the longitudinal linkage (2017[17]).

Standards for estimating programme costs

Developing accurate estimates of the cost of an intervention is one of the main concerns in economic evaluations; and represents an opportunity to improve subsequent programme planning and implementation (Crowley et al., 2018[93]). According to this, some of the approaches agree on planning cost data collection, ideally, in the early stages of the intervention through standardized methodologies such as: a macro top-down approach; or a bottom-up approach (See Box 3.20). Some of them also provide information regarding tools to facilitate the process of data collection: for instance, CostOut, produced by the Center for Benefit-Cost Studies of Education (Vera Institute, 2014[92]; Crowley et al., 2018[93]), which was designed to simplify the estimation of costs and cost-effectiveness of educational or other social programmes. Other approaches provide information on current practices, such as the analytical report presented by EMCDDA (2017[94]), which compiled initiatives for estimating drug treatment costs across eleven countries—including US, Australia, Portugal, Italy, and Czech Republic and the European Union.

Box 3.20. Methods for costing an intervention

- *Macro top-down approach (costs to society)*: uses total public spending (or individual site budget or expenditure) data to provide gross average estimates of intervention costs.
- *Bottom-up approach (or Micro costing or Ingredients Methods):* relies on identifying all resources required to implement an intervention and then valuing those resources in monetary units to estimate intervention costs. This method requires the evaluator to:
  - Describe the theory of change or logic model guiding the programme;
  - Identify specific resources used to implement the programme;
  - Value the cost of those resources;
  - Estimate the programme's total, average, and marginal costs.

Source: (Crowley et al., 2018[93]; National Academies of Sciences, Engineering, and Medicine, 2016[89]; Vera Institute, 2014[92]).

Whatever the methodology chosen to measure the cost of an intervention, the majority of approaches agree on covering as much information cost categories as possible (See Box 3.21) to ensure unbiased analysis (Vera Institute, 2014[92]; NSW Goverment, 2017[91]). For instance, the National Academy of Sciences (2016[89]) not only considers personnel, space, materials, and supplies (in the micro costing method); but also found useful to register direct cost, indirect cost (e.g. volunteer time), fixed cost (do not vary with the number of participants served), and variable costs, particularly when the evaluator is interested in an intervention's marginal and steady-state (average) costs. Additional to these costs, Crowley et al (2018[93]) also suggests that resources needed to support programme adoption, implementation, sustainability, and monitoring should be included in cost estimates.

Box 3.21. Type of costs

- *Direct Costs:* value of all the goods, services, and other resources that are consumed in the provision of an intervention. (e.g. Total programme expenditures for the most recently completed fiscal year)
- *Fixed costs*
  - Capital cost: for setting up the service
  - Labour costs: salaries and fringe benefit amounts for all programme staff
  - Overhead cost: for running the services
- *Variable costs:* costs that change in relation to variations in an activity.
- *Indirect Costs*: travel, training, trainer fees, supplies, materials, utilities, office rent, office services or volunteering.
- *Intangible costs:* Monetary value of pain, suffering, distress etc. associated with treatment (measured through Willingness To Pay).

Sources: (Lomas et al., 2018[95]) (NPC Research & Portland State University's Center for the Improvement of Child and Family Services, 2019[96]).

Other approaches, such as the WSIPP (USA) (2017[17]), use several strategies in meta-analysis to construct programme cost estimates. Some of their principles are the following:

- If the programme evaluations they have meta-analysed contain information on the number of "physical resource units" used by the programme, then they summarize those units, and produce an estimate of the average cost.
- The per-participant programme costs represent the cost of the average person who enters the programme, rather than the cost of a participant who completes the programme.
- In addition to a per participant cost estimate, they also note the year in which the dollars are denominated.
- Programmes that involve multiple years of per-participant spending can be present valued with NPV equation, where the discount factor depends on the years.

Standards for valuing the costs and benefits of programmes

After having identified the resources used in an intervention and their outcomes, the approaches typically refer to how cost and benefits should be valued. This will depend on the type of economic evaluation, its purpose, and time horizon.

For CA, some of the approaches consider the market price of a resource as a good approximation for its opportunity cost. For example, CADTH (Canada) recommends that the fees and prices, listed in schedules and formularies of Canadian ministries of health, be considered as unit-cost measures when calculating the perspective of the public payer (2017[97]). Other approaches suggest that shadow prices can be used, as another method for valuing the resource, when a market price does not exist (National Academies of Sciences, Engineering, and Medicine, 2016[89]; Crowley et al., 2018[93]). Shadow prices are used to capture the appropriate economic value in terms of willingness to pay: what consumers are willing to forget to obtain a given benefit or avoid a given cost (Karoly, 2012[98]).

For CBA, most of the approaches agree on the three summary statistics or decision rules in CBA model (See Box 3.22) Net Present Value (NPV); the Benefit-Cost Ratio (BCR); and the Internal Rate of Return (IRR). Particularly, NPV requests that factors such as the discount rate, inflation and time horizon be taken into account (National Academies of Sciences, Engineering, and Medicine, 2016[89]; OECD, 2018[87]; OMB, 1992[99]; CADTH, 2017[97]; Crowley et al., 2018[93]; NSW Goverment, 2017[91]).

Box 3.22. Summary statistics on CBA

The OECD has worked extensively on the use of Cost Benefit Analysis (CBA) as a core tool of public policy. Their last publications on the topic reflect a timely update on recent developments in the theory and practice of CBA, and provide detailed information on the following definitions:

- *Net Present Value (NPV):* refers to the present value of net benefits so that: NPV = PV(B) – PV(C) with present values calculated at the social discount rate. Where PV(B) denotes the (gross) present value of benefits, and PV(C) the gross present value of costs.
  - Discount rate: Under the assumption that the social or shadow price of a unit of consumption in the future is lower than the price of a unit of consumption today. The discount rate simply measures the rate of change of the shadow price.
- *Benefit-Cost Ratio (BCR):* The general rules become: i) accept a project if: PV(B)/PV(C) > 1; ii) in the face of rationing: rank by the ratio PV(B)/PV(C); or, iii) in choosing between mutually exclusive projects: select the project with the highest benefit cost ratio.
- *Internal Rate of Return (IRR):* The net present value rule requires the use of some predetermined social discount rate rule to discount future benefits and costs. An alternative rule is to calculate the discount rate which would give the project a NPV of zero and then to compare this "solution rate" with the pre-determined social discount rate.

Source: OECD (2018[100]; 2006[101]).

Other approaches such as the Office for Management and Budget (USA) (2018[102]) provides further Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programmes. They stipulate and update information on treatment of inflation, Real Discount Rates (a forecast of real interest rates from which the inflation premium has been removed), and Nominal Discount Rates (a forecast of nominal or

market interest rates for calendar year). Additionally, the OECD (2018[100]) provides information about health valuations for valuing risks to life (VSL), and the value of a (statistical) life-year (VSLY) (See Box 3.23).

Box 3.23. Health Valuation

Many public policies are mainly designed to protect life and health. However, these policies come with substantial costs that are not immediately detected. Understanding how both government and individuals must make trade-offs between income and risk, has led evaluators to measure the value that individuals place on changes in risk levels for threats to life and health (Bosworth, Professor and Kibria, 2017[103]), through some of the following measurements:

Valuing risks to life

The procedure for measuring the value of a statistical life captures how much an individual is willing to pay (WTP) to secure a risk reduction arising from a policy or project, or the willingness-to-accept (WTA) a compensation for tolerating higher than "normal" risks; which is equal to the marginal rate of substitution between income (or wealth) and mortality risk.

Value of a (statistical) life-year

The VSLY approach replaces the assumption (implicit in the way that VSL is typically applied) that age does not matter with an alternative assumption that age not only matters but it matters in a particular way.

Source: (OECD, 2018[100]; Bosworth, Professor and Kibria, 2017[103]).

For CEA, some of approaches stipulate the need for a comprehensive measure of the intervention's economic costs from a societal perspective, and the examination of one or more no-monetized outcomes(s). However, the use of different units to measure outcomes limits the aggregation of them. This problem has been mitigated by the development of measures, such as quality-adjusted life years (QALYs) (also known as cost-utility analysis) or disability-adjusted life years (DALYs) (National Academies of Sciences, Engineering, and Medicine, 2016[89]) (Goverment of Netherlands, 2016[104]). Other approaches, such as NICE (UK) (2013[105]), provide parameters to consider an intervention cost effective (less than £20,000 per QALY gained or between £20,000 and £30,000 per QALY, if certain conditions are satisfied).

One of the problems that arise from valuing cost and benefits is **double counting.** This issue refers to outcomes (benefits or cost) that are inputs for other outcomes or can be linked within each other. A number of approaches mention the need or precautions to avoid double counting (OECD, 2018[100]; Goverment of Netherlands, 2016[104]; National Academies of Sciences, Engineering, and Medicine, 2016[89]). For instance, Crowley et al (2018[93]) suggests that one approach to manage double counting is to employ a series of "trumping rules" that isolate developmental pathways to ensure no double counting occurs. Another example is the Treasury from New Zealand, which provides practical examples concerning double counting (See Box 3.24).

Another problem that can arise when measuring cost and benefits are **externalities**. This issue refers to goods that, once produced, can be consumed simultaneously by any number of people and from which people can't be excluded. These can have negative or positive effects (2015[106]). OMB (USA) presents several examples where externalities are addressed, such as in the principle of Willingness-To-Pay, where

market prices provide an invaluable starting point for measuring costs, but prices sometimes do not adequately reflect the true value of a good to society; hence, the use of *shadow prices*[3] can avoid market distortions, such as externalities or taxes (1992[99]). Externalities also lead to measuring indirect cost and intangible costs as another way to avoid a measurement bias in the evaluation. The Treasury from New Zealand offers another example regarding externalities (See Box 3.24).

Box 3.24. Examples for avoiding double counting and externalities: The Treasury of New Zealand

The treasury created the *Guide to Social Cost Benefit Analysis* with the aim of encouraging that all public decisions be accompanied by at least a robust and comprehensive CBA, through main steps into the generic policy development process. For instance, they present some of the principles and examples to identify the costs and benefits (Step 3) of a programme or policy, such as avoidance of double counting and externalities.

Avoidance of double counting

- A depreciation charge is intended to reflect the 'consumption' of capital, or the reduction in the value of the capital investment over a specified period, but it may double count the cost of an investment if the construction costs were already included in the CBA. Accounting practice treats construction cost as capital expenditure and recognises depreciation as an operating cost.

Externalities

- A 'carbon' tax would be an example, though whether it could be set at the right level would depend on whether it was possible to determine the resource cost that carbon emissions give rise to (in terms of adverse climate consequences). The effect of the tax is for people to take the cost of the externality into account in their actions, i.e. to 'internalise' the externality. Externalities are quantified when doing CBAs in order to ensure that they are internalised in the spending decision or regulatory decision.

Source: Adapted from The treasury of New Zealand (2015[106]), Guide to Social Cost Benefit Analysis - July 2015.

Standards for handling estimate uncertainty

Because cost and benefits estimations are made prior to the implementation of the programme, the outcomes from an economic evaluation could take many different possible values; and create a level of uncertainty. Standards are needed for estimating, resolving, and reporting that uncertainty. A risk, on the other hand, refers to whether the information allows for the estimation of the full range of possibilities of an event in terms of their probabilities. (OECD, 2018[100]).

The majority of approaches focus on uncertainty, and request testing the economic projections using a variety of approaches to sensitivity analysis (See Box 3.25). For instance, The Pew Charitable Trusts (USA) (2013[107]) stipulates the need to conduct and report sensitivity analysis, and provide a range of possible outcomes to ensure methodological rigor and transparency from an economic evaluation. Another example is the *Regulatory Impact Analysis* (RIA) Guidelines by the Government of Ireland (2009[108]), which

suggest that any assumptions made in RIAs (and the MCAs and CBAs performed in this context) should be calculated for a variety or range of future values through a sensitivity analysis.

Most of the approaches suggest using the Monte Carlos Analysis to handle uncertainty. Other approaches propose different practices such as Partial sensitivity analysis and Break-even analysis (See Box 3.25). Additionally to this, some approaches stipulate the need for these methods not only to test the robustness of the findings, but also to present estimates within a confidence interval and their standards errors (Crowley et al., 2018[93]) (National Academies of Sciences, Engineering, and Medicine, 2016[89]).

Box 3.25. Mechanism to test uncertainty

Sensitivity analysis shows how responsive—or sensitive—a study's results are to changes in assumptions. These are some of the types of sensitivity analysis (Vera Institute, 2014[92]):

- *Partial sensitivity analysis*: In this analysis, you select one variable and change its value while holding the values of other variables constant, to determine how much the CBA results change in response.
- *Break-even analysis*: If you are unable to estimate a policy's most likely effects or cannot find comparable studies to help determine its best-case and worst-case scenarios, you can use a break-even analysis. This helps identify how large a policy's impact must be for its benefits to equal its costs, that is, to break even. By definition, breaking even results in a net present value of $0.
- *Monte Carlo analysis*: uses estimates of the probability distributions of costs and benefits, and other parameters used in CBA, to undertake a probabilistic analysis of the likely NPV to emerge from a particular project. The steps to Monte Carlo analysis are as follows (OECD, 2018[100]):
  1. Estimate the probability distributions for the parameters of interest. Where parameters are likely to be correlated, the joint probability distributions are estimated;
  2. Take a random draw of the parameters of interest of sample size;
  3. Estimate the NPV n times using the parameters drawn;
  4. Calculate the mean NPV across the n estimates and store the value;
  5. Repeat m times until one can plot the probability distribution of mean NPV conditional on the uncertain parameters with sample size n, with m repetitions;
  6. Evaluate the likelihood of a positive or negative NPV.

Source: (Vera Institute, 2014[92]; OECD, 2018[100])

Standards for reporting the results of economic evaluations

The procedure of reporting economic evaluation findings depends not only on the type of evaluation conducted, but also on the needs to promote transparency and comparability across studies. The standards are requested to provide best practices for reporting a clear record of how the evaluation was conducted and to support the verification of the findings by an independent researcher.

The majority of approaches provide guidelines on how and what to report in an economic evaluation. For instance, the National Academies of Sciences (2016[89]) offer a Checklist of Best Practices for Reporting

Economic Evidence according to the methodology implemented (CA, CEA or CBA). Another example is the Government of Netherlands (2016[104]), which provides guidance related to reporting input values, costs, and uncertainty analysis, within others.

Some of the approaches suggest more specific reporting requirements such as maintaining a common table of inputs and assumptions. For instance, Crowley et al (2018[93]) recommends, in their Standards for Reporting Findings from Economic Evaluations, implementing a two-tiered reporting system that includes a consumer-focused summary accompanied by a technical description (e.g. included as an appendix) that details the modelling and assumptions made to estimate costs and benefits. Another example is the Vera institute (USA) (2014[92]), which provides guidance for CBA on how to tabulate results, document the analysis, and interpret the findings (See Figure 3.6).

Figure 3.6. Sample table of CBA results

The Enhanced Services for the Hard-to-Employ Demonstration
Table 4.4
Net Benefits and Costs (in 2009 Dollars),
Full Sample
Center for Employment Opportunities

	Taxpayer ($)	Victim ($)	Participant ($)	Total ($)
Benefits				
Criminal justice	2,912	432	0	3,344
Employment				
Earnings	0	0	590	590
Tax payments and credits[a]	190	0	–190	0
Value of output from CEO transitional jobs[b]	4,576	0	0	4,576
Costs				
CEO program costs[c]	–3,603	0	0	–3,603
Net benefits (per person)	4,075	432	400	4,907
Benefit-cost ratio	2.13	NA	NA	2.36

SOURCES: Marginal costs were estimated using information provided by New York City and state criminal justice agencies, New York State executive budget agencies, and Taifa (1995). See Appendix Table F.1 for more detail. For additional employment data sources, see Table 3.1; for additional criminal justice data sources, see Table 3.2.

NOTES:
[a] Income taxes and federal and state Earned Income Tax Credit were based on tax rules for filing year 2009.
[b] The value of output from CEO transitional jobs was calculated using information from the Department of Citywide Administrative Services.
[c] CEO program costs were calculated using CEO's financial expenditure reports for Fiscal Year 2005 and CEO's management information system (MIS), adjusted for inflation.

Source: Vera Institute (2014[92])

Only a small number of approaches are concerned with the delivery (time) and accessibility of the evaluation findings. For instance, the Government of Ireland (2009[108]) discusses factors such as: Where should *Regulatory Impact Analysis (RIAs)* be published? And the Evidence-Based Policymaking Collaborative (USA) (2016[109]), which highlights the importance that CBA results are delivered in accessible, concise, and compelling ways, and completed in time to inform decision-makers' choices. They consider that adopting rigorous, replicable CBA methodologies and making data readily available to conduct analyses can help improve timeliness.

Key questions

Framing an economic evaluation

- Does the economic evaluation specify an empirical question, with an underlying model or theory?
- Does the evaluation adopt a societal perspective? If not, what are the limitations of the evidence used?
- Does the evaluation take into account options for agile design and delivery methods that could have implications for the results of the evaluation?

Identifying the programme/policy impacts

- Does the economic evaluation describe the methods used to identify the programme/policy impacts (Quality design, measurements, sample, approach of analysis, findings, and external validity)?
- Does the economic evaluation well-describe the impacts, including magnitude, significant and non-significant findings, significance levels, standard errors or standard deviations, methods of estimation?
- Does the economic evaluation describe any limitations to generalizability, internal and external validity of impacts, as well as confidence intervals?
- If the economic evaluation uses a systematic review to identify the programme impacts, does it stipulate the methodological criteria to select the studies?

Estimating programme costs

- Does the economic evaluation cover different cost categories (e.g., labour, equipment, training, materials and supplies, office space, indirect cost) according to the stages of an intervention or programme (e.g. adoption, implementation, or monitoring)
- Does the economic evaluation use micro costing procedures or ingredients methods to improve the quality of intervention cost estimates?
- Does the costing concern both direct and indirect costs?

Implementation and Scale up of Interventions

Knowledge of 'what works' – of which policies and programmes are effective, is necessary but not sufficient for obtaining outcomes for citizens. Increasingly, there is recognition that 'implementation matters'- that the quality and level of implementation of an intervention of a policy is associated with outcomes for citizens (Durlak, 1998[110]; Durlak and DuPre, 2008[111]).

Why is it important?

It is important to understand the features of policies and programmes, of the organisation or entity implementing them, along with the myriad of other factors that are related to adoption, implementation and sustainability of a policy or programme. This enables practical guidance to enable successful implementation and scale-up efforts. Increased attention to implementation has also been drawn by the work of economists such as Pr. Duflo, and the JPAL networks working on development issues. It has been estimated that interventions implemented correctly can achieve effects two or three times greater than interventions where problems with implementation have been experienced (Durlak and DuPre, 2008[111]).

Summary of the mapping of existing approaches

Of the approaches included in the mapping, the majority include some coverage issues concerning implementation and scale up of interventions.

Standards concerning the delivery and implementation of an intervention

Most of the approaches that cover implementation and scale up are focused on simply providing factual details about the delivery and implementation requirements of an intervention. These approaches include the Australian What Works for Kids, the Canadian Best Practices Portal, Spain's 'Prevención basada en la evidencia' and the Evidence Based Teen Pregnancy Programmes in the USA. Spain's approach provides information related to the delivery of an intervention, its materials and setting. The Canadian Best Practices Portal is another approach that provides key information about what is required to implement an intervention (Box 3.26).

Box 3.26. Canadian Best Practices Portal – Adaptability of an intervention

The Canadian Best Practices Portal is a compilation of multiple sources of information on interventions, best practices and resources for chronic disease prevention and health promotion recommended by experts. The Portal links to resources and solutions to plan programmes for promoting health and preventing diseases for populations and communities. The portal includes three types of interventions: Best Practices, Promising Practices and Aboriginal 'Ways Tried and True'.

From the overview of the interventions evaluated, the portal provides information related to the adaptability of the programme, in terms of:

- Implementation history;
- Expertise required for implementation within the context of the intervention;
- Available support for implementation;
- And, resources and/or products associated with the interventions.

Source: National Collaborating Centre for Methods and Tools (2010[112])

Other approaches provide more granularity about the implementation requirements of an intervention. The Evidence Based Teen Pregnancy Programs in the USA has a standalone section on implementation, which comprises of eight fields including implementation requirements and guidance and allowable adaptations. What Works for Kids also has a standalone section on implementation which includes the following fields:

- Training
- Can training be accessed in Australia?
- Who delivers the programme?
- Minimum practitioner qualifications
- Are there any licensing or accreditation requirements?
- Is there a manual that describes how to implement the programme?
- What are the required materials for the trainer?
- Are specific assessments required prior to implementation?
- Are particular tools required for implementation?

- Overall implementation / resourcing issues
- Is the programme scalable?
- Comments on the scalability of the intervention
- Setup costs
- Ongoing costs

Experiences of the implementation of an intervention

Some of the approaches that cover issues around implementation and scale up are focused on providing and categorising experiences of implementing an intervention. These experiences are typically the findings from process evaluations and qualitative studies. These approaches include The Community guide, the EMCDDA Best Practice Portal, the EU-Compass for Action on Mental Health and Well-Being and HomeVee.

The Community Guide is a resource that helps practitioners and policy makers to improve health and safety in their communities. As part of a ten-step process it includes details about the applicability and barriers to implementation for the recommended interventions. The EU-Compass for Action on Mental Health and Well-being also includes information on experiences of implementation and is described in Box 3.27. The EMCDDA Best Practice Portal has recently published a new database of programmes for implementation. This includes details of programmes that have been implemented in more than one European Country, along with details of experiences of implementation (EMCDDA, 2020[113]).

HomeVee is another approach that provides a summary of 'Implementation Experiences' based on the studies included in a review, focusing on:

- Characteristics of Model Participants,
- Location and Setting,
- Staffing and Supervision,
- Model Components,
- Model Adaptations or Enhancements,
- Dosage (Home visits), and
- Lessons Learned.

Box 3.27. EU-Compass for Action on Mental Health and Well-being

The EU-Compass for Action on Mental Health and Wellbeing is a web-based mechanism used to collect exchange and analyse information on policy and stakeholder activities in mental health. The EU-Compass collects examples of good practices from EU countries and stakeholders on an ongoing basis. These examples are evaluated using quality criteria agreed with EU countries and the Commission. Each year, a brochure with examples of good practices is published regarding one of the EU-Compass areas. The brochure contains a brief summary of each practice which addressed priority areas, the lessons learned, and level of implementation. Particularly, EU-Compass provides information related to the lessons learned from an implemented programme in three main sub-sections:

- What worked well & facilitators to implementation;
- What did not work & barriers to implementation;
- Recommendations for future adopters of this practice.

Source: EU-Compass for Action on Mental Health and Well-being (EU) (European Commission - Directorate-General for Health and Food Safety, 2017[49]).

Standards concerning the dissemination readiness and system readiness

A small number of approaches go further in providing detailed criteria against which dissemination readiness and/or system readiness could be assessed. These are features of the intervention or of the organisation or community adopting the intervention that have been shown to be related to adoption, implementation, or sustainability of the intervention (Society for Prevention Research Standards of Evidence, 2015[53]). The purpose of such approaches is to support the implementation and scale-up efforts of evidence-based interventions.

These approaches include the EU-Compass for Action on Mental Health and Well-being, the Green List Prevention, NESTA, Housing Associations' Charitable Trust, Blueprints, and SUPERU. The Green List Prevention includes six criteria to rate the 'implementation quality' of an intervention including whether 'support / technical assistance during implementation is available' and whether 'instruments for quality control during the implementation are available'. Blueprints includes five criteria on 'dissemination readiness' including that 'there are explicit processes for ensuring the intervention gets to the right persons'. SUPERU also developed comparable criteria, described in Box 3.28.

Box 3.28. SUPERU's intervention consistency and documentation criteria

The strength of evidence scale of Superu (NZ) consisted of five levels. The levels corresponded not just for ascending rankings for strength of evidence, but also to expectations about the type of evidence that can and should be generated about an intervention as it matures and grows; specifically, evidence about which elements of the intervention are necessary to implement with fidelity, and which can be adapted. Level 4 is defined as appropriate for mature, largescale interventions with a strong evidence base. Interventions that reach this level must contain the following information:

- There is evidence that the intervention is consistently delivered as planned and reaches its target groups.
- *Intervention consistency and documentation*
- There is clarity and documentation about what the intervention comprises.
- There is regular review of procedures, manuals and staff training processes.
- Information is available on the resources (money and people) required to deliver the intervention.
- Technical support is available to help implement the intervention in new settings.

Source: (SUPERU, 2016[114])

A limited number of approaches go further in either explicitly scoring the implementation readiness of an intervention. The Evidence Based Teen Pregnancy Programs conducts a detailed assessment of an intervention's 'Implementation Readiness' conducted based on materials and documents about the intervention and its implementation. Based on this assessment, an implementation readiness score is awarded by three component scores: (1) curriculum and materials, (2) training and staff support, and (3) fidelity monitoring tools and resources. The component scores are added together to give a total score, which ranges from 0 to 8, with higher scores indicating the interventions most ready to implement.

Key questions

Requirements

- Are materials available that specify the activities to be carried out and optimal methods of implementing the intervention?
- Is there guidance about who can implement the intervention and what education and training they must have to be able to implement the intervention successfully?
- Are training and technical assistance available for implementing the core components of the intervention?

Intervention Readiness

- Are tools available to ensure that the intervention is being implemented as intended?
- Is there a system in place for documenting adaptations to core components that occur during implementation?
- Is there a system to support regular monitoring of implementation?
- Is there a system in place to support planning and monitoring of service user recruitment?

System readiness

- Are there tools to enable the organization or community adopting the intervention to assess factors that are likely to impede or facilitate adoption and successful implementation of the intervention?
- Are there tools to enable the organization or community adopting the intervention measure their own capacity to implement in a high-quality fashion?

Real world experiences of implementation

- Are there existing resources available, which describe the findings of studies, examine the barriers, and facilitators of implementing the intervention?
- Are there any recommendations for adopters of the intervention that could be incorporated into the implementation of the intervention?
- Are the findings and conclusions of the assessment shared with and validated by a range of key stakeholders?

References

Agency for Healthcare Research and Quality (2012), *What Is the Evidence Rating?*, https://innovations.ahrq.gov/help/evidence-rating (accessed on 19 February 2019). [56]

Alton-Lee, A. (2004), *Guidelines for Generating a Best Evidence Synthesis Iteration*, Ministry of Education New Zealand, http://www.minedu.govt.nz (accessed on 14 February 2019). [10]

Askim, J., U. Hjelmar and L. Pedersen (2018), "Turning Innovation into Evidence-based Policies: Lessons Learned from Free Commune Experiments", *Scandinavian Political Studies*, Vol. 41/4, pp. 288-308, http://dx.doi.org/10.1111/1467-9477.12130. [84]

Axford, N. et al. (2005), "Evaluating Children's Services: Recent Conceptual and Methodological Developments", *British Journal of Social Work*, Vol. 35/1, pp. 73-88, http://dx.doi.org/10.1093/bjsw/bch163. [39]

Be You (2020), *The Be You Programs Directory*, https://beyou.edu.au/resources/tools-and-guides/about-programs-directory (accessed on 25 March 2020). [78]

Better Evaluation (2012), *Describe the theory of change*, https://www.betterevaluation.org/en/node/5280 (accessed on 20 October 2019). [41]

Blueprints for Health Youth Development (2015), *Evidence-Based Programs - Standards of Evidence*, https://www.blueprintsprograms.org/resources/Blueprints_Standards_full.pdf (accessed on 15 February 2019). [48]

Blueprints for health youth development (2018), *Blueprints Database Standards*, https://www.blueprintsprograms.org/resources/Blueprints_Standards_full.pdf (accessed on 15 February 2019). [82]

Bosworth, R., A. Professor and A. Kibria (2017), *THE VALUE OF A STATISTICAL LIFE: ECONOMICS AND POLITICS Primary Investigators*, https://strata.org/pdf/2017/vsl-full-report.pdf (accessed on 7 June 2019). [103]

CADTH (2017), *Guidelines for the Economic Evaluation of Health Technologies: Canada 4th Edition*, https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf (accessed on 17 April 2019). [97]

Child Trends (2018), , https://www.childtrends.org/what-works/eligibility-criteria. [71]

Clearinghouse for Labor Evaluation and Research (2017), *About CLEAR*, https://clear.dol.gov/about (accessed on 8 March 2019). [6]

Clearinghouse for Labor Evaluation and Research (2014), *Guidelines for reviewing implementation studies*, https://clear.dol.gov/sites/default/files/CLEAR_Operational%20Implementation%20Study%20Guidelines.pdf (accessed on 19 February 2019). [63]

Clearinghouse for Labor Evaluation and Research (2014), *Guidelines for reviewing quantitative descriptive studies*, https://clear.dol.gov/sites/default/files/CLEAROperationalDescriptiveStudyGuidelines.pdf (accessed on 19 February 2019). [60]

Clearinghouse for Military Family Readiness (2012), *Continuum of Evidence*, https://militaryfamilies.psu.edu/wp-content/uploads/2017/08/continuum.pdf (accessed on 30 January 2019). [72]

Coalition for Evidence-Based Policy (2010), *Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence*, http://coalition4evidence.org/wp-content/uploads/2010/02/Checklist-For-Reviewing-a-RCT-Jan10.pdf (accessed on 19 February 2019). [2]

College of Policing: What Work Network (2017), *Crime Reduction Toolkit*, https://whatworks.college.police.uk/toolkit/Pages/Toolkit.aspx (accessed on 30 April 2019). [51]

Crime Solutions (2013), *Practices Scoring Instrument*, https://www.crimesolutions.gov/pdfs/PracticeScoringInstrument.pdf (accessed on 18 February 2019). [5]

Crime Solutions (2013), *Program Scoring Instrument Version 2.0*, https://www.crimesolutions.gov/pdfs/program-rating-instrument-v2.0.pdf (accessed on 18 February 2019). [54]

Crowley, D. et al. (2018), "Standards of Evidence for Conducting and Reporting Economic Evaluations in Prevention Science", *Prevention Science*, Vol. 19/3, pp. 366-390, http://dx.doi.org/10.1007/s11121-017-0858-1. [93]

Drost, E. (2011), *Validity and Reliability in Social Science Research*, https://www3.nd.edu/~ggoertz/sgameth/Drost2011.pdf (accessed on 13 February 2019). [62]

Durlak, J. (1998), "Why program implementation is important", *Journal of Prevention & Intervention in the community*, Vol. 17/2, pp. 5-18. [110]

Durlak, J. and E. DuPre (2008), "Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation", *American journal of community psychology*, Vol. 41/3-4, pp. 327-350. [111]

Early Childhood Australia (2020), *KidsMatter has become Be You*, http://www.earlychildhoodaustralia.org.au/our-work/beyou/ (accessed on 22 April 2020). [79]

Early Intervention Foundation (2019), *10 steps for evaluation success*, Early Intervention Foundation, http://dx.doi.org/12345. [55]

Early Intervention Foundation (2018), *EIF Guidebook*, https://guidebook.eif.org.uk/eif-evidence-standards (accessed on 14 February 2019). [32]

Education Endowment Foundation (2018), *Technical appendix and process manual*, https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf (accessed on 1 February 2019). [25]

EFSA Guidance for those carrying out systematic reviews European Food Safety Authority (2010), "Application of systematic review methodology to food and feed safety assessments to support decision making", *EFSA Journal*, Vol. 8/6, p. 1637, http://dx.doi.org/10.2903/j.efsa.2010.1637. [37]

EMCDDA (2020), *Xchange prevention registry*, http://www.emcdda.europa.eu/best-practice/xchange (accessed on 21 April 2020). [113]

EMCDDA (2017), "Drug treatment expenditure: a methodological overview", http://www.emcdda.europa.eu/publications/insights/drug-treatment-expenditure-measurement_en. [94]

Epstein, D. and J. Klerman (2012), "When is a Program Ready for Rigorous Impact Evaluation? The Role of a Falsifiable Logic Model", *Evaluation Review*, Vol. 36/5, pp. 375-401, http://dx.doi.org/10.1177/0193841X12474275. [40]

European Commission (2018), *Guidance Document on Monitoring and Evaluation*. [45]

European Commission (2011), "Towards a new system of monitoring and evaluation in EU cohesion policy", https://ec.europa.eu/regional_policy/sources/docgener/evaluation/doc/performance/outcome_indicators_en.pdf. [43]

European Commission - Directorate-General for Health and Food Safety (2017), *Criteria to select best practices in health promotion and chronic disease prevention and management in Europe*, https://ec.europa.eu/health/sites/health/files/mental_health/docs/compass_bestpracticescriteria_en.pdf (accessed on 25 January 2019). [49]

European Monitoring Centre for Drugs and Drug Addiction (2020), *Best practice portal*, http://www.emcdda.europa.eu/best-practice_en (accessed on 27 April 2020). [19]

European Monitoring Centre for Drugs and Drug Addiction (2011), "European drug prevention quality standards", http://dx.doi.org/10.2810/48879. [9]

European Platform for Investing in Children (2017), *Review Criteria and Process*, https://ec.europa.eu/social/main.jsp?catId=1246&intPageId=4286&langId=en (accessed on 14 February 2019). [57]

Every Student Succeeds Act - ESSA (2019), *Evidence for ESSA: Standards and Procedures*, https://content.evidenceforessa.org/sites/default/files/On%20clean%20Word%20doc.pdf (accessed on 18 February 2019). [34]

Evidence-Based Policymaking Collaborative (2016), *Evidence Toolkit: Cost Benefit Analysis*. [109]

Ferri, M. and P. Griffiths (2015), "Good Practice and Quality Standards", in *Textbook of Addiction Treatment: International Perspectives*, Springer Milan, http://dx.doi.org/10.1007/978-88-470-5322-9_64. [18]

Gaffey, V. (2013), "A fresh look at the intervention logic of Structural Funds", European Commision, http://dx.doi.org/10.1177/1356389013485196. [44]

Ghate, D. (2018), "Developing theories of change for social programmes: co-producing evidence-supported quality improvement", *Palgrave Communications*, Vol. 4/1, p. 90, http://dx.doi.org/10.1057/s41599-018-0139-z. [42]

Gough, D., J. Thomas and S. Oliver (2019), *Clarifying differences between reviews within evidence ecosystems*, BioMed Central Ltd., http://dx.doi.org/10.1186/s13643-019-1089-2. [13]

Gough, D. and H. White (2018), *Evidence standards and evidence claims in web based research portals*, Centre for Homelessness Impact, https://uploads-ssl.webflow.com/59f07e67422cdf0001904c14/5bfffe39daf9c956d0815519_CFHI_EVIDENCE_STANDARDS_REPORT_V14_WEB.pdf (accessed on 8 March 2019). [3]

Goverment of Netherlands (2016), *Guideline for economic evaluations in healthcare*. [104]

Government of Ireland (2009), *How to conduct a Regulatory Impact Analysis*. [108]

Graham Allen (2011), *Early Intervention: The Next Steps An Independent Report to Her Majesty's Government*, http://www.childtrauma.org (accessed on 14 February 2019). [73]

Groeger-Roth, F. and B. Hasenpusch (2011), *Green List Prevention: Inclusion-and Rating-Criteria for the CTC Programme-Databank Crime Prevention Council of Lower Saxony*, Crime Prevention Council of Lower Saxony, https://www.gruene-liste-praevention.de/communities-that-care/Media/GreenListPrevention_Rating-Criteria.pdf (accessed on 14 February 2019). [50]

Guyatt, G. et al. (2008), "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations", *BMJ*, Vol. 336/7650, pp. 924-926, http://dx.doi.org/10.1136/bmj.39489.470347.ad. [38]

Health Evidence (2018), *Quality Assessment Tool*, https://www.healthevidence.org/documents/our-appraisal-tools/quality-assessment-tool-dictionary-en.pdf (accessed on 8 March 2019). [4]

Hollis, S. and F. Campbell (1999), "What is meant by intention to treat analysis? Survey of published randomised controlled trials", *British Medical Journal*, Vol. 42, p. 4, http://dx.doi.org/10.1136/bmj.319.7211.670. [68]

Home Visiting Evidence of Effectiveness (2018), *Assessing Evidence of Effectiveness*, https://homvee.acf.hhs.gov/Review-Process/4/Assessing-Evidence-of-Effectiveness/19/7 (accessed on 19 February 2019). [83]

Home Visiting Evidence of Effectiveness (2018), *Producing Study Ratings*, https://homvee.acf.hhs.gov/Review-Process/4/Producing-Study-Ratings/19/5 (accessed on 19 February 2019). [77]

Housing Associations' Charitable Trust (2016), *Standard for Producing Evidence - Effectiveness of Interventions – Part 1: Specification*, https://www.hact.org.uk/sites/default/files/StEv2-1-2016%20Effectiveness-Specification.pdf (accessed on 25 January 2019). [59]

Johns Hopkins University School of Education's Center for Data-Driven Reform in Education - CDDRE (n.d.), *Best Evidence Encylopedia*, http://www.bestevidence.org/aboutbee.htm (accessed on 18 February 2019). [80]

Johnson, S., N. Tilley and K. Bowers (2015), "Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis reviews", *Journal of Experimental Criminology*, Vol. 11/3, pp. 459-473, http://dx.doi.org/10.1007/s11292-015-9238-7. [27]

Karoly, L. (2012), "Toward Standardization of Benefit-Cost Analysis of Early Childhood Interventions", *Journal of Benefit-Cost Analysis*, Vol. 3/1, http://dx.doi.org/10.1515/2152-2812.1085. [98]

Karoly, L. (2010), "Toward Standardization of Benefit-Cost Analyses of Early Childhood Interventions", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.1753326. [115]

Levin, C. and D. Chisholm (2016), "Cost-Effectiveness and Affordability of Interventions, Policies, and Platforms for the Prevention and Treatment of Mental, Neurological, and Substance Use Disorders", *Mental, Neurological, and Substance Use Disorders: Disease Control Priorities*, Vol. 4, http://dx.doi.org/10.1596/978-1-4648-0426-7_ch12. [86]

Lomas, J. et al. (2018), "Which Costs Matter? Costs Included in Economic Evaluation and their Impact on Decision Uncertainty for Stable Coronary Artery Disease", *PharmacoEconomics - Open*, Vol. 2/4, pp. 403-413, http://dx.doi.org/10.1007/s41669-018-0068-1. [95]

Mathematica Policy Research (2016), *Identifying Programs That Impact Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors*, https://tppevidencereview.aspe.hhs.gov/pdfs/TPPER_Review%20Protocol_v5.pdf (accessed on 26 January 2019). [33]

National Academies of Sciences, Engineering, and Medicine (2016), *Advancing the Power of Economic Evidence to Inform Investments in*, The National Academies Press, http://dx.doi.org/10.17226/23481. [89]

National Collaborating Centre for Methods and Tools (2010), *Effective interventions: The Canadian Best Practices Portal*, McMaster University, Hamilton, http://www.nccmt.ca/resources/search/69 (accessed on 15 February 2019). [112]

National Dropout Prevention Center - NDPC (2019), *Rating system*, http://dropoutprevention.org/mpdb/web/rating-system (accessed on 19 February 2019). [81]

National Implementation Research Network (2018), *The Hexagon: An Exploration Tool. Hexagon Discussion & Analysis Tool Instructions*, https://implementation.fpg.unc.edu/sites/implementation.fpg.unc.edu/files/resources/NIRN_HexagonTool_11.2.18.pdf (accessed on 19 February 2019). [8]

National Institute for Health and Care Excellence (2013), *How NICE measures value for money in relation to public health interventions*, https://www.nice.org.uk/Media/Default/guidance/LGB10-Briefing-20150126.pdf (accessed on 1 May 2019). [105]

Nest What works for kids (2012), *Rapid Evidence Assessment*, http://whatworksforkids.org.au/rapid-evidence-assessment (accessed on 19 February 2019). [1]

NESTA (2013), *Standards of evidence: an approach that balances the need for evidence with innovation*, https://media.nesta.org.uk/documents/standards_of_evidence.pdf (accessed on 26 February 2019). [47]

New Zealand Treasury (2015), *Guide to Social Cost Benefit Analysis - July 2015*, New Zealand Treasury, Wellington, http://www.treasury.govt.nz/publications/guidance/planning/costbenefitanalysis/guide/ (accessed on 2 May 2018). [106]

Newhouse J.P., H. (1993), *Free for All, lessons from the Rand Health Insurance Experiment,*, https://doi.org/10.7249/CB199. [66]

Norwegian Institute of Public Health (2021), *Elementer i livsstilstiltak for vektreduksjon blant voksne personer med overvekt eller fedme*, https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2021/elementer-i-livsstilstiltak-for-vektreduksjon-blant-voksne-personer-med-overvekt-eller-fedme-rapport-2021-v2.pdf. [23]

Norwegian Institute of Public Health (2020), *A systematic and living evidence map on COVID-19*, https://www.fhi.no/contentassets/e64790be5d3b4c4abe1f1be25fc862ce/covid-19-evidence-map-protocol-20200403.pdf. [22]

NPC Research & Portland State University's Center for the Improvement of Child and Family Services (2019), *Conduct a Cost Analysis of Your Home Visiting Program*, http://www.homevisitcosts.com/organizing-your-data.php (accessed on 22 May 2019). [96]

NSW Goverment (2017), *Guide to Cost-Benefit Analysis*, https://www.treasury.nsw.gov.au/sites/default/files/2017-03/TPP17-03%20NSW%20Government%20Guide%20to%20Cost-Benefit%20Analysis%20-%20pdf_0.pdf. [91]

OECD (2018), *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264085169-en. [100]

OECD (2018), "Preface", in *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264085169-1-en. [87]

OECD (2017), "Making policy evaluation work: The case of regional development policy", *OECD Science, Technology and Industry Policy Papers*, https://dx.doi.org/10.1787/c9bb055f-en. [67]

OECD (2006), *Cost-Benefit Analysis and the Environment: Recent Developments*. [101]

Office of Manangment and Budget (2018), *2018 Discount Rates for OMB Circular No. A-94*. [102]

Oliver, S. et al. (2018), "Approaches to evidence synthesis in international development: a research agenda", *Journal of Development Effectiveness*, Vol. 10/3, pp. 305-326, http://dx.doi.org/10.1080/19439342.2018.1478875. [11]

OMB (1992), *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A94/a094.pdf (accessed on 17 September 2018). [99]

Oxman, A., J. Lavis and A. Fretheim (2007), "Use of evidence in WHO recommendations", *Lancet*, Vol. 369/9576, pp. 1883-1889, http://dx.doi.org/10.1016/S0140-6736(07)60675-8. [20]

Project Oracle Children and Youth Evidence Hub (2018), *Validation Guidebook: An overview of Project Oracle's validation process*, https://project-oracle.com/uploads/files/Validation_Guidebook.pdf (accessed on 14 February 2019). [46]

Puddy, R. and N. Wilkins (2011), *Understanding Evidence Part 1: Best Available Research Evidence. A Guide to the Continuum of Evidence of Effectiveness*, Centers for Disease Control and Prevention, Atlanta, https://www.cdc.gov/violenceprevention/pdf/understanding_evidence-a.pdf (accessed on 18 February 2019). [64]

Reduction, W. (2017), *Crime Reduction Toolkit*, The College of policing, https://whatworks.college.police.uk/toolkit/Pages/Toolkit.aspx (accessed on 30 April 2019). [90]

Review Srengthening Families Evidence (2019), *Review Process*, https://familyreview.acf.hhs.gov/ReviewProcess.aspx?id=3 (accessed on 19 February 2019). [76]

Saran, A. and H. White (2018), "Evidence and gap maps: a comparison of different approaches", http://dx.doi.org/10.4073/cmdp.2018.2. [15]

Scholtes, V., C. Terwee and R. Poolman (2010), "What makes a measurement instrument valid and reliable?", *Injury*, Vol. 42, pp. 236-240, http://dx.doi.org/10.1016/j.injury.2010.11.042. [61]

Shemilt, I. et al. (2010), "Evidence synthesis, economics and public policy", *Research Synthesis Methods*, Vol. 1/2, pp. 126-135, http://dx.doi.org/10.1002/jrsm.14. [12]

Social Programs That Work (2019), *Evidence Based Programs*, https://evidencebasedprograms.org/ (accessed on 19 February 2019). [70]

Society for Prevention Research Standards of Evidence (2015), "Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation", *Society for Prevention Research Standards of Evidence*, Vol. 16/7, pp. 893-926. [53]

Strengthening Families Evidence Review - SFER (2018), *Review Process*, https://familyreview.acf.hhs.gov/ReviewProcess.aspx?id=3 (accessed on 19 February 2019). [35]

SUPERU (2017), *An evidence rating scale for New Zealand. Understanding the effectiveness of interventions in the social sector*. [52]

SUPERU (2016), *Standards of evidence for understanding what works: International experiences and prospects for Aotearoa New Zealand*, SUPERU, Wellington, http://www.superu.govt.nz/sites/default/files/Standards%20of%20evidence.pdf (accessed on 18 April 2018). [114]

The California Evidence-Based Clearinghouse for Child Welfare (2019), *Scientific Rating Scale*, http://www.cebc4cw.org/ratings/scientific-rating-scale/ (accessed on 25 January 2019). [75]

The Campbell Collaboration (2019), *Campbell systematic reviews: Policies and Guidelines*, http://dx.doi.org/10.4073/cpg.2016.1. [30]

The Cochrane Collaboration (2021), *Cochrane Denmark*, https://www.cochrane.dk/nordic-cochrane-centre-copenhagen. [21]

The Cochrane Collaboration (2011), *Cochrane Handbook for Systematic Reviews of Interventions*, https://handbook-5-1.cochrane.org/ (accessed on 18 April 2019). [88]

The Community Guide (2018), *The Community Guide Methodology*, https://www.thecommunityguide.org/about/our-methodology (accessed on 19 February 2019). [36]

The EQUATOR Network (2020), *Enhancing the quality and transparency of health research*, https://www.equator-network.org/ (accessed on 21 April 2020). [28]

The Pew Charitable Trusts (2013), *States' Use of Cost-Benefit Analysis*. [107]

The UK Civil Service (2014), *What is a Rapid Evidence Assessment?*, https://webarchive.nationalarchives.gov.uk/20140402163359/http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/what-is (accessed on 22 April 2020). [16]

Tõnurist, P. (2019), *Evaluating Public Sector Innovation Support or hindrance to innovation?*, OECD, Paris. [85]

Vera Institute (2014), *Cost-Benefit Analysis and Justice Policy Toolkit*. [92]

Washington State Institute for Public Policy Benefit (2017), *Benefit-Cost Technical Documentation*. [17]

What Works Centre for Children's Social Care (2018), *Evidence standards*, https://wwc-evidence.herokuapp.com/pages/our-ratings-explained (accessed on 19 February 2019). [69]

What Works Centre for Local Economic Growth (2016), *Guide to scoring evidence using the Maryland Scientific Methods Scale*, https://whatworksgrowth.org/public/files/Methodology/16-06-28_Scoring_Guide.pdf (accessed on 24 January 2019). [65]

What Works Centre for Local Economic Growth (2015), *Evidence Review Apprenticeships*, https://whatworksgrowth.org/public/files/Policy_Reviews/15-09-04_Apprenticeships_Review.pdf (accessed on 13 May 2019). [29]

What Works Centre for Wellbeing (2017), *A guide to our evidence review methods*, https://whatworkswellbeing.org/product/a-guide-to-our-evidence-review-methods/ (accessed on 8 March 2019). [31]

What Works Clearinghouse (2020), *Standards Handbook (Version 4.1)*, https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf (accessed on 5 February 2019). [74]

What Works For Health (2010), *Evidence Rating: Guidelines*, http://whatworksforhealth.wisc.edu/evidence.php (accessed on 6 June 2019). [24]

What Works For Health (2010), *Methods*, http://whatworksforhealth.wisc.edu/evidence.php (accessed on 19 February 2019). [58]

White, H. (2019), "The twenty-first century experimenting society: the four waves of the evidence revolution", *Palgrave Communications*, Vol. 5/1, p. 47, http://dx.doi.org/10.1057/s41599-019-0253-6. [14]

Whiting, P. et al. (2016), "ROBIS: A new tool to assess risk of bias in systematic reviews was developed", *Journal of Clinical Epidemiology*, Vol. 69, pp. 225-234, http://dx.doi.org/10.1016/j.jclinepi.2015.06.005. [26]

Zaza, S. et al. (2000), *Data Collection Instrument and Procedure for Systematic Reviews in the Guide to Community Preventive Services*, http://www.thecommunityguide.org (accessed on 25 January 2019). [7]

Notes

[1] An extended family or community of related families who live together in the same area. (Taken from Oxford Dictionary)

[2] This will be broadly discuss in the effectiveness section.

[3] A shadow price is an estimate of an economic value when market-based values are unavailable (e.g., no market for buying and selling emotional regulation) (Karoly, 2010[115]). The quality and consensus on shadow prices can vary by substantive area. Sometimes an estimate is only appropriate for projections in certain circumstances and should not be generalized (Crowley et al., 2018[93]).

4 Conclusion

This chapter reviews the objectives, achievements, and limitations of this stocktaking of principles and standards for the use of evidence in OECD countries. The chapter recognises the results of this report as a first global effort on the road to create a supportive institutional framework for evidence-informed policy-making. This work complements OECD's previous work on building capacity for evidence-informed policy-making, as well as on a data driven public sector. In addition, the chapter discusses opportunities for future work. It points out the relevance of understanding current practice and exploring the potential for a shared approach to standards. This is an important step to ensure the reliability of public-sector decision-making processes and to ultimately nurture trust in decision-making in government.

What has this report achieved?

This report looks to provide a stocktaking of principles and standards for the use of evidence. Although other mappings exist at a national level, this report is perhaps unique in attempting to map these approaches across OECD countries. As a result, the number of approaches enables lessons to be reached about the strengths and gaps of existing practice. In discussing a range of detailed case study examples, there is a wealth of material that can be used to enable countries to reflect on their own practice and facilitate knowledge sharing.

In the future, this mapping could facilitate a shared approach to evidence across the OECD and could represent an initial exploratory step in identifying and developing guidance at the international level in the area of evidence and evaluation. Some stakeholders have already called for a standardised approach to standards to avoid the confusion and duplication which can arise from the proliferation of evidence standards (Vine, 2018[1]).

At this stage, the current mapping remains descriptive. It opens the floor for future discussions on the issues that should be consider for the review and production of evidence. This report represents an important component of OECD's work on evidence-informed policy and it also provides an important broadening of the discussion about evidence-based policy making, complementing the OECD work on a data driven public sector, which emphasises this area as an important policy cluster for the public sector (OECD, 2019[2]). The report should also be considered alongside work on Building capacity for evidence-informed policy-making (OECD, 2020[3]). In addition, a broader report focused on the role of policy evaluation in terms of institutionalisation, quality and use is also bringing comparative insights on a range of guidelines and approaches to structure the analysis and evaluation of public policies and programmes (OECD, 2020[4]).

Opportunities for future work

This report provides an initial resource for agencies and national experts thinking about the role of standards and principles in their evidence architecture. Reflecting both the nature of existing practice and the thematic interests of the project, the current review has primarily focused on public interventions and especially in the social policy area even if the ambition is to generate findings of general significance. This work could be complemented by future work exploring the synergies with emerging policy approaches in a data-driven public sector, where technological opportunities are also shaping a new environment for policy making. While the report makes reference to work on artificial intelligence for example, this is an area of rapid technological and policy development which could deserve future investigations.

The focus of the analysis of the standards of evidence reviewed in this report concerns primarily public interventions, as discrete decisions related to the concrete supply of public services, which is a bit more specific than the evaluation of policies as a whole, but is complementary. As a result, the report makes a specific but focused contribution to the full range of evidentiary considerations involved in policy-making. For example, it is recognised that evidence can play a role in the policy issues as developing a financial aid system in the area of development of in developing clinical guidelines for how physicians respond to an issue in the health area. However, such issues are not directly addressed in the current report and the communities of experts involved in development policy or in health care may either have developed similar stocktaking approaches. Engaging with such communities on the findings of the report may help to consolidate and improve common ground to strengthen evidence informed policy making. This is particularly true from a whole of government perspective, as there is a need to promote the evidence underlying public policy from across the range of policy fields.

Mapping existing practice is an important first step in understanding current practice and exploring the potential for a shared approach to standards. This report adopts a stocktaking neutral approach to offer a

general framework to describe and understand existing approaches to principles and standards. While the case can be made that some of the methods for assessing evidence described in this report are more rigorous than others, consensus has yet to emerge among the experts. Therefore, this report does not rate the quality of the approaches nor does it aim at any form of comprehensive ranking. What the report offers is to encapsulate in some portable format a presentation of the key steps that are generally involved in the design, implementation and evaluation of public interventions. While the approach is not normative as such and that all of the steps do not need to be considered jointly, it is clear that ignoring any of the steps presented here would not be conducive in any way to sound public governance. Considering some of them, at least in some significant way and ensuring space in the public sector for such considerations, should still be seen as an essential step in consolidating trust by citizens in the soundness of public-sector decision-making processes. Whilst the report is comprehensive, future work could also make this technical material accessible to a wider audience. One option would be to turn the material into a toolkit for public servants and experts to be used and promoted throughout the public sector, in some portable or web-based format.

Countries might also consider updating their standards to reflect recent and ongoing developments in policy evaluation and experimentation. As discussed in the Effectiveness chapter, existing approaches to standards of evidence rarely address systems based approaches to evaluation, which deal with open-ended problems and issues (Askim, Hjelmar and Pedersen, 2018[5]; Tõnurist, 2019[6]). These tensions between traditional linear approaches to impact evaluation and systems based approaches are part of the OECD approaches under its Observatory for Public Sector Innovation: integrating these insights could represent a next step for updating standards of evidence in the future.

Finally the field of data itself is rapidly changing as a result of the COVID 19 pandemic, with researchers, economists turning to high frequency and big data to track the economy in real time,[1] as traditional economic indicators might always be produced with some lag. This involves data from non-traditional data sources, including scraping the web, satellites, sensors, data from platforms. The OECD has started to analyse some of the current practices and applications (Johnstone N., 2019[7]). With the widespread deployment of the “Internet of Things”, such data sources are likely to growth exponentially in the future. The challenge will be to apply some quality filters and to ensure that some of these new data sources can be mobilised to formulate reliable analysis that can lead to meaningful conclusions for policy making.

References

Askim, J., U. Hjelmar and L. Pedersen (2018), “Turning Innovation into Evidence-based Policies: Lessons Learned from Free Commune Experiments”, *Scandinavian Political Studies*, Vol. 41/4, pp. 288-308, http://dx.doi.org/10.1111/1467-9477.12130. [5]

Johnstone N., A. (2019), *Using Digital Technologies to Improve the Design and Enforcement of Public Policies*, https://doi.org/10.1787/99b9ba70-en. [7]

OECD (2020), *Building Capacity for Evidence Informed Policy Making*. [3]

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

OECD (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/059814a7-en. [2]

Tõnurist, P. (2019), *Evaluating Public Sector Innovation Support or hindrance to innovation?*, OECD, Paris. [6]

Vine, J. (2018), *Standardising standards The case for shared standards in the evidence sector Alliance for Useful Evidence*. [1]

Note

[1] As highlighted in the OECD Work on New Approaches to Economic Challenges and the Innovation Lab. See. http://www.oecd.org/naec/projects/naecinnovationlabevents/.

Annex A. Overview of the principles for the use of evidence across jurisdictions

Table A.1. Principles for the Good Governance of Evidence

Country	Organisation	Framework	Appropriateness	Ensuring Integrity (Honest Brokerage)	Accountability	Contestability	Public representation in decision making	Transparency	Quality	Rigour	Effective Resource Management
EU	EuroScientist	The Brussels declaration on ethics & principles for science & society policy-making			YES	YES		YES			
EU	Commission of the European Communities	Communication from the commission on the collection and use of expertise by the commission: principles and guidelines Improving the knowledge base for better policies						YES	YES		
EU	European Food Safety Authority	Promoting Methods for Evidence use in Scientific assessments,				YES		YES	YES		

EU	European Food Safety Authority	Application of systematic review methodology to food and feed safety assessments to support decision-making,				YES		YES	YES	YES	
EU	European Food Safety Authority	EU-ANSA agencies' engagement in the EU research knowledge cycle									
Japan	Science Council of Japan	Proper relations between science and the government in policy making and implementation	YES		YES	YES	YES	YES	YES		
Malawi	Ministry of Health	Guidelines for Evidence Use in Decision-Making in the Health Sector in Malawi	YES					YES	YES		
New Zealand	Royal Society of New Zealand	Public Engagement Guidelines		YES			YES	YES			
New Zealand	Peter Gluckman	The Art of Science Advice to Government		YES			YES				
OECD	OECD	Scientific Advice for Policy Making				YES	YES	YES			
OECD	OECD	Recommendation of the OECD Council on Health Data Governance		YES		YES		YES			
UK	Justin Parkhusrt	The good governance of	YES		YES	YES	YES	YES	YES	YES	

		evidence									
UK	Louise Shaxson	Uncovering the practices of evidence-informed policy-making	YES		YES	YES	YES	YES	YES	YES	
UK	The Royal Society	Principles for good evidence synthesis for policy						YES		YES	
UK	National Insitutute for Health and Care Excellence-NICE	Principles for putting evidence-based guidance into practice			YES				YES		
UK	Government Office for Science	Guidelines on the Use of Scientific and Engineering Advice in Policy Making	YES					YES			
UK	Bond	An introduction to the principles for assessing the quality of evidence				YES		YES	YES	YES	
UK	Alliance for Useful Evidence	Evidence vs Democracy				YES				YES	
UK	Palgrave Communications (University of Oxford)	Three lessons from evidence-based medicine and policy: increase transparency, balance inputs and understand power	YES			YES		YES			
UK	Society for Prevention Research	Ethical Challenges in Promoting the Implementation of	YES	YES		YES					

		Preventive Interventions									
UK	Committee on Standards in Public Life	The 7 principles of public life		YES	YES			YES			
UK	Government Office for Science	Principles of scientific advice to government		YES			YES	YES			
UK	Nuffield Council on Bioethics	Emerging Biotechnologies: technology, choice and the public good				YES	YES	YES			
USA	Office of Management and Budget (OMB)	Building the capacity to produce and use evidence	YES	YES			YES	YES		YES	
USA	Institute of Medicine	Bridging the Evidence Gap in Obesity Prevention: A Framework to Inform Decision Making	YES					YES	YES		

Annex B. Mapping of Existing Standards of Evidence across a Range of Jurisdictions

Table B.1. Mapping of Standards of Evidence

#	Country	Organisation	Framework	URL	URL for evidence standard
1	Australia	Be you	The Be You Programs Directory	https://beyou.edu.au/get-started	https://beyou.edu.au/resources/tools-and-guides/about-programs-directory
2	Australia	ARACY Australian Research Alliance for Children and Youth	What Works for Kids (WW4K)	http://whatworksforkids.org.au/	http://whatworksforkids.org.au/rapid-evidence-assessment
3	Canada	Public Health Agency of Canada	Canadian Best Practices Portal	http://cbpp-pcpe.phac-aspc.gc.ca/resources/evidence-informed-decision-making/	https://www.nccmt.ca/registry/resource/pdf/69.pdf
4	Canada	McMaster University	Health Evidence	https://www.healthevidence.org/search.aspx	https://www.healthevidence.org/documents/our-appraisal-tools/quality-assessment-tool-dictionary-en.pdf
5	EU	European Commission	EU-Compass for Action on Mental Health and Well-being	https://ec.europa.eu/health/non_communicable_diseases/mental_health/eu_compass_en	https://ec.europa.eu/health/sites/health/files/mental_health/docs/compass_bestpracticescriteria_en.pdf
6	EU	European Platform for Investing in Children	Evidence Based Practices	https://ec.europa.eu/social/main.jsp?catId=1246&langId=en	https://ec.europa.eu/social/main.jsp?catId=1246&intPageId=4286&langId=en
7	EU	EMCDDA European Monitoring Centre for Drugs and Drug Addiction	European drug prevention quality standards	http://www.emcdda.europa.eu/emcdda-home-page_en	http://www.emcdda.europa.eu/system/files/publications/646/TD3111250ENC_318193.pdf

#	Country	Organisation	Framework	URL	URL for evidence standard
8	EU	EMCDDA European Monitoring Centre for Drugs and Drug Addiction	Best practice portal	http://www.emcdda.europa.eu/best-practice_en	http://www.emcdda.europa.eu/best-practice/evidence/about
9	Germany	Crime Prevention Council of Lower Saxony	Green List Prevention	http://lpr.niedersachsen.de/nano.cms/english	https://www.gruene-liste-praevention.de/communities-that-care/Media/GreenListPrevention_Rating-Criteria.pdf
10	New Zealand	SUPERU	A Quality Scale for New Zealand	https://www.superu.govt.nz/resources/evidence-rating-scale	https://www.superu.govt.nz/sites/default/files/Publications/Evidence%20Rating%20Scale.pdf
11	New Zealand	Education counts	Best Evidence Synthesis Iteration	https://www.educationcounts.govt.nz/home	https://www.educationcounts.govt.nz/__data/assets/pdf_file/0016/6640/BES-Development-Guidelines-27-07-04.pdf
12	Spain	Prevención basada en la evidencia	Criterios de selección de programas	http://www.prevencionbasadaenlaevidencia.net/index.php	http://www.prevencionbasadaenlaevidencia.net/index.php?page=Criterios
13	UK	Darlington Service Design Lab	Standards of evidence	https://dartington.org.uk/	https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/284086/early-intervention-next-steps2.pdf
14	UK	Project Oracle	Standards of evidence	https://project-oracle.com/what-weve-done/all-projects/	https://project-oracle.com/uploads/files/Validation_Guidebook.pdf
15	UK	Early Intervention Foundation	The Guidebook	https://guidebook.eif.org.uk/	https://guidebook.eif.org.uk/eif-evidence-standards
16	UK	Nesta	Standards of evidence	https://www.nesta.org.uk/	http://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf
17	UK	Bond	Evidence Principles	https://www.bond.org.uk/resources/evidence-principles	https://www.bond.org.uk/ngo-support/evidence-principles-download
18	UK	Centre for Analysis of Youth Transitions (CAYT)	Standards of evidence (CAYT)	http://cayt.mentor-adepis.org/	http://cayt.mentor-adepis.org/wp-content/uploads/2017/06/CAYT-Scoring-Application-Form-2017-FINAL.pdf
19	UK	What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	http://wwg.percipio.london/	https://whatworksgrowth.org/public/files/Methodology/16-06-28_Scoring_Guide.pdf

#	Country	Organisation	Framework	URL	URL for evidence standard
20	UK	Big Lottery Fund's Realising Ambition Programme	The confidence review	http://www.theconfidenceframework.org.uk/	http://www.theconfidenceframework.org.uk/different-sectors/
21	UK	Education Endowment Foundation	Teaching and Learning Toolkit	https://educationendowmentfoundation.org.uk/	https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf
22	UK	HACT-Ideas and Innovation in Housing	Standards for producing evidence	https://www.hact.org.uk/	https://www.hact.org.uk/sites/default/files/StEv2-1-2016%20Effectiveness-Specification.pdf
23	UK	Conservation evidence	What Works in Conservation	https://www.conservationevidence.com/	https://www.conservationevidence.com/content/page/79
24	UK	What Works Centre for Children's Social Care	Evidence Standards	https://whatworks-csc.org.uk/	https://wwc-evidence.herokuapp.com/pages/our-ratings-explained
25	UK	What Works Centre for Wellbeing	GRADE (Grading of Recommendations Assessment, Development and Evaluation)	https://whatworkswellbeing.org/	https://whatworkswellbeing.org/product/a-guide-to-our-evidence-review-methods/
26	UK	What Works Centre for Crime Reduction	EMMIE Framework	https://whatworks.college.police.uk/toolkit/Pages/About_the_CRT.aspx	https://whatworks.college.police.uk/toolkit/Pages/Quality-Scale.aspx
27	International	Campbell Collaboration	Campbell Collaboration Systematic Reviews: Policies and Guidelines	https://campbellcollaboration.org/	https://campbellcollaboration.org/library/campbell-collaboration-systematic-reviews-policies-and-guidelines.html
28	USA	Community Preventive Services Task Force (CPSTF)	The Community Guide	https://www.thecommunityguide.org/	https://www.thecommunityguide.org/about/our-methodology
29	USA	U.S. Department of Health & Human Services	Home Visiting Evidence of Effectiveness	https://homvee.acf.hhs.gov/	https://homvee.acf.hhs.gov/Review-Process/4/Overview/19
30	USA	What Works Clearinghouse	Find What Works from Systematic Reviews	https://ies.ed.gov/ncee/wwc/	https://ies.ed.gov/ncee/wwc/Handbooks
31	USA	Center for the Study and Prevention of Violence	Blueprints	http://www.blueprintsprograms.com	https://www.blueprintsprograms.org/resources/Blueprints_Standards_full.pdf
32	USA	California Department of Social Services	California Evidence-Based Clearinghouse for Child Welfare	http://www.cebc4cw.org/home	http://www.cebc4cw.org/ratings/scientific-rating-scale
33	USA	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of	Best Evidence Encyclopedia	http://www.bestevidence.org	http://www.bestevidence.org/aboutbee.htm

#	Country	Organisation	Framework	URL	URL for evidence standard
		Education			
34	USA	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education	Evidence for ESSA (Every Student Succeeds Act)	https://www.evidenceforessa.org/	https://content.evidenceforessa.org/sites/default/files/On%20clean%20Word%20doc.pdf
35	USA	Society for Prevention Research	Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science.	http://www.preventionresearch.org/	http://www.preventionresearch.org/wp-content/uploads/2011/12/Standards-of-Evidence_2015.pdf
36	USA	U.S. Department of Health and Human Services	Evidence Based Teen Pregnancy Programs	https://tppevidencereview.aspe.hhs.gov/Default.aspx	https://tppevidencereview.aspe.hhs.gov/pdfs/TPPER_Review%20Protocol_v5.pdf
37	USA	National Institute of Justice	Crimesolutions	https://www.crimesolutions.gov	https://www.crimesolutions.gov/about_evidencecontinuum.aspx
38	USA	Arnold Ventures	Social Programmes That Work	https://evidencebasedprograms.org/	http://toptierevidence.org
39	USA	Child Trends	What Works for Child and Youth Development	http://www.childtrends.org/what-works	http://www.childtrends.org/what-works/eligibility-criteria
40	USA	Washington State Institute for Public Policy (WSIPP)	Washington State Institute for Public Policy Benefit -Cost Results	http://www.wsipp.wa.gov/BenefitCost	http://www.wsipp.wa.gov/TechnicalDocumentation/WsippBenefitCostTechnicalDocumentation.pdf
41	USA	U.S Department of Justice	Office of Juvenile Justice and Delinquency Prevention Model Programs Guide	http://www.ojjdp.gov/programs/ProgSearch.asp	https://www.ojjdp.gov/mpg/Home/About
42	USA	University of Wisconsin Population Health Institute's	What Works for Health	http://whatworksforhealth.wisc.edu/search-options.php	http://whatworksforhealth.wisc.edu/evidence.php
43	USA	U.S. Department of Health and Human Services	The Agency for Healthcare Research and Quality (AHRQ)	https://www.ahrq.gov/cpi/about/profile/index.html	https://innovations.ahrq.gov/help/evidence-rating
44	USA	U.S. Department of Labor	Clearinghouse for Labor Evaluation and Research (CLEAR)	https://clear.dol.gov/	https://clear.dol.gov/about
45	USA	Clearinghouse for Military Family Readiness	Continuum of evidence	https://lion.militaryfamilies.psu.edu/programs/find-programs	https://militaryfamilies.psu.edu/wp-content/uploads/2017/08/continuum.pdf

#	Country	Organisation	Framework	URL	URL for evidence standard
46	USA	Suicide Prevention Resource Center	Evidence-Based Practices Project	http://www.sprc.org/	http://www.sprc.org/sites/default/files/ebpp_proj_descrip%20revised.pdf
47	USA	National Implementation Research Network	The Hexagon: An Exploration Tool	https://nirn.fpg.unc.edu/	https://implementation.fpg.unc.edu/sites/implementation.fpg.unc.edu/files/imce/documents/NIRN%20Hexagon%20Discussion%20Analysis%20Tool%20v2.2.pdf
48	USA	National Dropout Prevention Center	Model Programs Database	http://dropoutprevention.org/	http://dropoutprevention.org/mpdb/web/rating-system
49	USA	National Cancer Institute	Research-Tested Intervention Programs (RTIPs)	https://www.cancer.gov/	https://rtips.cancer.gov/rtips/reviewProcess.do
50	USA	Strengthening Families Evidence Review	Standards of evidence	https://familyreview.acf.hhs.gov/Default.aspx	https://familyreview.acf.hhs.gov/ReviewProcess.aspx?id=3

Table B.2. Type of Evidence Assessed and Key Standards of Evidence by Approach

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
1	Australia	Be you	The Be You Programs Directory	X		X	X	YES	YES	YES	NO	YES	NO	YES	NO	YES	NO
2	Australia	ARACY Australian Research Alliance for Children and Youth	What Works for Kids (WW4K)	X		X		NO	YES	YES	YES	YES	YES	YES	NO	NO	NO
3	Canada	Public Health Agency of Canada	Canadian Best Practices Portal	X				NO	YES	YES	YES	NO	NO	YES	NO	YES	NO
4	Canada	McMaster University	Health Evidence	X	X	X	X	NO	NO	YES	YES	YES	YES	NO	NO	NO	NO
5	EU	European Commission	EU-Compass for Action on Mental Health and Well-being	X				YES	NO	YES	YES	YES	YES	YES	YES	YES	NO
6	EU	European Platform for Investing in Children	Evidence Based Practices	X		X		NO	YES	YES	YES	YES	YES	YES	NO	NO	NO
7	EU	EMCDDA European Monitoring Centre for Drugs and Drug Addiction	European drug prevention quality standards	X		X	X	YES	YES	YES	YES	YES	YES	YES	YES	NO	NO
8	EU	EMCDDA European	Best practice portal	X	X	X	X	NO	NO	YES	YES	NO	NO	NO	NO	NO	YES

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
		Monitoring Centre for Drugs and Drug Addiction															
9	Germany	Crime Prevention Council of Lower Saxony	Green List Prevention	X				YES	YES	YES	YES	YES	NO	YES	NO	NO	NO
10	New Zealand	SUPERU	A Quality Scale for New Zealand	X	X			YES	YES	YES	YES	YES	YES	YES	NO	NO	YES
11	New Zealand	Education counts	Best Evidence Synthesis Iteration	X	X	X	X	YES	YES	YES	YES	NO	YES	YES	NO	NO	YES
12	Spain	Prevención basada en la evidencia	Criterios de selección de programas	X		X		YES	NO	YES	YES			YES	NO	NO	YES
13	UK	Darlington Service Design Lab	Standards of evidence	X		X		YES	YES	YES	YES	YES	NO	YES	YES	YES	NO
14	UK	Project Oracle	Standards of evidence	X		X	X	YES	YES	YES	YES	NO	YES	NO	YES	NO	NO
15	UK	Early Intervention Foundation	The Guidebook	X				YES	YES	YES	YES	YES	NO	YES	NO	NO	NO
16	UK	Nesta	Standards of evidence	X		X	X	YES	YES	YES	YES	YES	NO	YES	YES	NO	NO
17	UK	Bond	Evidence Principles	X		X	X	YES	NO	YES	NO	NO	NO	NO	NO	YES	YES
18	UK	Centre for Analysis of Youth Transitions (CAYT)	Standards of evidence (CAYT)	X		X	X	NO	YES	YES	YES	YES	NO	NO	NO	NO	NO

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
19	UK	What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	X		X		NO	YES	YES	NO	NO	NO	NO	NO	NO	NO
20	UK	Big Lottery Fund's Realising Ambition Programme	The confidence review	X		X	X	YES	YES	NO	NO	YES	YES	YES	YES	YES	NO
21	UK	Education Endowment Foundation	Teaching and Learning Toolkit	X	X	X		NO	YES	YES	NO	YES	YES	NO	NO	NO	NO
22	UK	HACT-Ideas and Innovation in Housing	Standards for producing evidence	X		X		YES	YES	YES	YES	NO	YES	YES	YES	YES	YES
23	UK	Conservation Evidence	What Works in Conservation	X		X		NO	NO	YES	NO	NO	NO	NO	NO	NO	NO
24	UK	What Works Centre for Children's Social Care	Evidence Standards	X	X			NO	NO	YES	YES	YES	YES	YES	YES	NO	YES
25	UK	What Works Centre for Wellbeing	GRADE (Grading of Recommendations Assessment, Development and Evaluation)	X		X	X	YES	YES	YES	YES	YES	YES	NO	NO	NO	YES
26	UK	What Works Centre for Crime Reduction	EMMIE Framework		X			YES	NO	YES	YES	YES	YES	YES	NO	NO	YES

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
27	International	Campbell Collaboration	Campbell Collaboration Systematic Reviews: Policies and Guidelines	X		X		YES	NO	YES	YES	YES	YES	YES	NO	NO	YES
28	USA	Community Preventive Services Task Force (CPSTF)	The Community Guide	X		X		YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
29	USA	U.S. Department of Health & Human Services	Home Visiting Evidence of Effectiveness	X		X		NO	NO	YES	YES	YES	NO	YES	NO	NO	YES
30	USA	What Works Clearinghouse	Find What Works from Systematic Reviews	X		X		NO	YES	YES	NO	NO	NO	NO	NO	NO	NO
31	USA	Center for the Study and Prevention of Violence	Blueprints	X				YES	NO	YES	YES	NO	NO	YES	YES	NO	YES
32	USA	California Department of Social Services	California Evidence-Based Clearinghouse for Child Welfare	X		X		NO	NO	YES	YES	NO	NO	YES	YES	YES	YES
33	USA	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education	Best Evidence Encyclopedia	X				NO	NO	YES	NO	NO	NO	YES	NO	NO	NO

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
34	USA	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education	Evidence for ESSA (Every Student Succeeds Act)	X	X			NO	YES	YES	NO	YES	NO	YES	NO	NO	NO
35	USA	Society for Prevention Research	Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science.	X				YES	NO	YES	YES	YES	YES	YES	YES	YES	YES
36	USA	U.S. Department of Health and Human Services	Evidence Based Teen Pregnancy Programs	X				NO	NO	YES	NO	NO	NO	YES	YES	NO	NO
37	USA	National Institute of Justice	Crimesolutions	X	X			YES	YES	YES	YES	NO	NO	YES	NO	NO	NO
38	USA	Arnold Ventures	Social Programmes That Work	X		X		NO	YES	YES	YES	NO	NO	NO	NO	NO	NO
39	USA	Child Trends	What Works for Child and Youth Development	X				NO	NO	YES	YES	NO	NO	NO	NO	NO	NO
40	USA	Washington State Institute for Public	Washington State Institute for Public Policy Benefit -	X				NO	NO	YES	YES	YES	YES	NO	NO	NO	NO

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
		Policy (WSIPP)	Cost Results														
41	USA	U.S Department of Justice	Office of Juvenile Justice and Delinquency Prevention Model Programs Guide	X		X		YES	NO	YES	YES	YES	NO	YES	NO	YES	NO
42	USA	University of Wisconsin Population Health Institute's	What Works for Health	X	X			NO	YES	YES	NO	NO	NO	NO	NO	NO	NO
43	USA	U.S. Department of Health and Human Services	The Agency for Healthcare Research and Quality (AHRQ)	X		X	X	NO	YES	YES	NO	YES	NO	YES	NO	NO	YES
44	USA	U.S. Department of Labor	Clearinghouse for Labor Evaluation and Research (CLEAR)	X		X		NO	YES	YES	NO	YES	NO	YES	NO	NO	YES
45	USA	Clearinghouse for Military Family Readiness	Continuum of evidence	X		X		NO	YES	YES	YES	YES	NO	YES	NO	NO	NO
46	USA	Suicide Prevention Resource Center	Evidence-Based Practices Project	X		X		YES	NO	YES	NO	YES	NO	YES	YES	NO	NO
47	USA	National Implementation Research Network	The Hexagon: An Exploration Tool	X		X		YES	YES	YES	YES	YES	NO	YES	YES	YES	NO

#	Country	Organization	Framework	Type of evidence assessed				Theory of change/ Logic Model	Design and development	Efficacy	Effectiveness	Cost		Implementation			
				Impact evaluation	Systematic review	Quantitative methods	Qualitative methods					Cost information	Cost-Benefit evaluation	Requirements	Intervention Readiness	System readiness	Experiences
48	USA	National Dropout Prevention Center	Model Programs Database	X		X		NO	YES	YES	YES	YES	NO	YES	NO	NO	NO
49	USA	National Cancer Institute	Research-Tested Intervention Programs (RTIPs)	X		X		YES	YES	YES	NO	YES	YES	YES	YES	NO	NO
50	USA	Strengthening Families Evidence Review	Standards of evidence	X		X		NO	NO	YES	NO	NO	NO	NO	NO	NO	YES

Table B.3. Rating and ranking of quality of evidence by approach

#	Organisation	Framework	Use of an assessment approach or scale
1	Be you	The Be You Programs Directory	No. Programs must: align with one or more of the five professional learning domains (Mentally Healthy Communities, Family Partnerships, Learning Resilience, Early Support, Responding Together) align with the Australian Curriculum or National Quality Framework be supported by a training/delivery/implementation manual or guide be offered as more than a one-off session (i.e., offer multiple, sequential sessions which, either as a set series of sessions or on an as-needs basis) be targeted at one of the following audiences as the intended beneficiary, for example: children, young people, parents, carers or families; early childhood educators, Out of Hours School Care have at least one research or evaluation study which demonstrates: a positive impact on mental health outcomes for children or young people a minimum of 20 participants in the study who received the program at least pre and post testing conducted on the group that received the program.
2	ARACY Australian Research Alliance for Children and Youth	Nest What Works for Kids (WW4K)	Yes. Well supported •No evidence of risk or harm. •If there have been multiple studies, the overall evidence supports the benefit of the program. •Clear baseline and post-measurement of outcomes for both conditions. •At least two RCTs have found the program to be significantly more effective than comparison group. Effect was maintained for at least one study at one-year follow-up. Supported •No evidence of risk or harm. •If there have been multiple studies, the overall evidence supports the benefit of the program. •Clear baseline and post-measurement of outcomes for both conditions. •At least one RCT has found the program to be significantly more effective than comparison group. Effect was maintained at 6-month follow-up. Promising •No evidence of risk or harm. •If there have been multiple studies, the overall evidence supports the benefit of the program. •Clear baseline and post-measurement of outcomes for both conditions. •At least one study using some form of contemporary comparison group demonstrated some improvement outcomes for the intervention but not the

#	Organisation	Framework	Use of an assessment approach or scale
			comparison group. Emerging •No evidence of risk or harm. •There is insufficient evidence demonstrating the program's effect on outcomes because: ◦the designs are not sufficiently rigorous (i.e. they do not meet the criteria of the above programs), or ◦the results of rigorous studies are not yet available.
3	Public Health Agency of Canada	Canadian Best Practices Portal	No. Promising Practices: A Promising Practice is defined as an intervention, program, service, or strategy that shows potential (or "promise") for developing into a best practice. Promising practices are often in the earlier stages of implementation, and as such, do not show the high level of impact, adaptability, and quality of evidence as best practices. However, their potential is based on a strong theoretical underpinning to the intervention. Aboriginal Ways Tried and True: Aboriginal 'Ways Tried and True' (WTT) refers to successful practices implemented in First Nations, Inuit, and Métis contexts to address local challenges. Success is measured not only by effectiveness, but also by how the intervention was designed and carried out. Interventions are intended to inspire and support public health practitioners, program developers, evaluators, and others by sharing information on programs and processes that have worked in Aboriginal contexts. Best Practices: A Best Practice is defined as an intervention, program, or initiative that has, through multiple implementations, demonstrated: high impact (positive changes related to the desired goals), high adaptability (successful adaptation and transferability to different settings), and high quality of evidence (excellent quality of research/evaluation methodology, confirming the intervention's high impact and adaptability evidence).
4	McMaster University	Health Evidence	Yes. Strong: Reviews with a score of 8 or higher in the Yes column Moderate: Reviews with a score between 5-7 in the Yes column Weak: Reviews with a score of 4 or less in the Yes column
5	European Commission	EU-Compass for Action on Mental Health and Well-being	Yes, although not numbered. Detailed criteria around three issues: Exclusion Criteria assess the following aspects: Relevance Intervention Characteristics Evidence and Theory base Ethical aspects Core criteria assess the following aspects:

#	Organisation	Framework	Use of an assessment approach or scale
			Effectiveness Efficiency Equity <u>Qualifier criteria assess the following aspects:</u> Transferability Sustainability Participation Intersectoral collaboration
6	European Platform for Investing in Children	Evidence Based Practices	Yes. Criteria to determine the evidence level are organised according to three categories: **•<u>Evidence of effectiveness:</u>** *Comparison group* + Evaluation utilises at the minimum pre/post design with appropriate statistical adjustments employed in order to control for selection + +: Study design uses a convincing comparison group to identify practice impacts, including randomised-control trial (experimental design) or some quasi-experimental designs *Statistical significance* + Significant ($p<0.1$), positive results are shown on at least one relevant outcome + + Significant ($p<0.05$), positive results are shown on at least one relevant outcome *Effect size* + No requirement + + Effect size of at least 10% of a standard deviation. *Sample size* + Sample size of at least 20 in each group + + Sample size of at least 50 in each group *Outcomes* + Outcomes are directly or indirectly related to outcomes identified in topic definitions + No significant negative outcomes reported (excluding those negative outcomes that might be due to chance) + + No significant negative outcomes reported (excluding those negative outcomes that might be due to chance) + + Outcomes are directly related to outcomes identified in topic definitions + + Outcome assessments have been validated, where applicable + + Outcome assessments conducted at baseline and follow-up, where applicable *Attrition* + No requirement + + Attrition is less than 25% or has been accounted for using an acceptable procedure, where applicable *Location:* At least one evaluation that meets the above criteria must have been conducted within EU member state(s) **•<u>Transferability</u>** *Replication*

#	Organisation	Framework	Use of an assessment approach or scale
			+ Practice has been evaluated in at least one additional population beyond the original study population** (broadly defined) in such a way that at least meets the basic criteria for internal validity as specified in the evidence of effectiveness criteria (e.g. significant positive results for at least one outcome are found, uses a comparison group, etc.) + +Same requirements as for + but in addition the practice has been found to be cost-effective/cost-beneficial (i.e. the practice can deliver positive impact at a reasonable cost) *Practice materials:* Practice materials (curriculum, etc.) are available, or documentation is sufficient, such that program can be replicated **•Enduring impact** *Follow-up conducted* An evaluation of the practice which meets the basic criteria for inclusion has conducted a follow-up of at least 2 years, and continues to find positive ($p<0.1$) and direct impact on at least one outcome **Evidence-based practices** on this site are assigned one of **three evidence levels**: •Emergent Practice: An "emergent practice" has achieved at least a + in "evidence of effectiveness." •Promising Practice: A "promising practice" has achieved at least a + in "evidence of effectiveness" and a + in at least one of the other two categories, "transferability" and "enduring impact." •Best Practice: A "best practice" has achieved at least a + in each of the three evidence categories, including "evidence of effectiveness", "transferability" and "enduring impact."
7	EMCDDA European Monitoring Centre for Drugs and Drug Addiction	European drug prevention quality standards	No, it is an eight-stage project cycle with cross-cutting considerations. Organised in an **eight-stage project cycle**, the Standards cover the following areas: Stage 1: Needs assessment Stage 2: Resource assessment Stage 3: Programme formulation Stage 4: Intervention design Stage 5: Management and mobilisation of resources Stage 6: Delivery and monitoring Stage 7: Final evaluations Stage 8: Dissemination and improvement **Cross-cutting considerations** are relevant for each project stage and are therefore placed in the centre of the project cycle**.** These Standards relate to: (A) sustainability and funding, (B) communication and stakeholder involvement, (C) staff development, (D) the ethics of drug prevention.
8	EMCDDA European Monitoring Centre for Drugs	Best practice portal	**Yes,**

#	Organisation	Framework	Use of an assessment approach or scale
	and Drug Addiction		**Evidence ratings** The available information on the effects of specific interventions are examined and then ranked them as described below. Beneficial: Interventions for which precise measures of the effects in favour of the intervention were found in the systematic reviews of randomised controlled trials (RCTs), and that were recommended in guidelines with reliable methods for assessing evidence (such as GRADE*). An intervention ranked as 'beneficial' is suitable for most contexts. Likely to be beneficial: Interventions that were shown to have limited measures of effect, that are likely to be effective but for which evidence is limited, and/or those that are recommended with some caution in guidelines with reliable methods for assessing evidence (such as GRADE). An intervention ranked as 'likely to be beneficial' is suitable for most contexts, with some discretion. Trade-off between benefits and harms: Interventions that obtained measures of effects in favour of harm reduction and/or are recommended in guidelines with reliable methods for assessing evidence (such as GRADE), but that showed some limitations or unintended effects that need to be assessed before providing them. Unknown effectiveness: Interventions for which there are not enough studies or where available studies are of low quality (with few patients or with uncertain methodological rigour), making it difficult to assess if they are effective or not. Interventions for which more research should be undertaken are also grouped in this category. Evidence of ineffectiveness: Interventions that gave negative results if compared with a standard intervention, for example. **Quality of evidence:** High quality evidence— one or more up-to-date systematic reviews that include high-quality primary studies with consistent results. The evidence supports the use of the intervention within the context in which it was evaluated. Moderate quality evidence— one or more up-to-date reviews that include a number of primary studies of at least moderate quality with generally consistent results. The evidence suggests these interventions are likely to be useful in the context in which they have been evaluated but further evaluations are recommended. Low quality evidence— where there are some high or moderate quality primary studies but no reviews available OR there are reviews giving inconsistent results. The evidence is currently limited, but what there is shows promise. This suggests these interventions may be worth considering, particularly in the context of extending services to address new or unmet needs, but should be evaluated.
9	Crime Prevention Council of Lower Saxony	Green List Prevention	Yes. Ratings of both programmes and evaluations. **Programme ratings** Level 1: Theoretically well grounded. Detailed criteria on the Conceptual Quality, Implementation Quality and Evaluation Quality Level 2: Probable Effectiveness Level1 and at least one evaluation study 1 to 3 stars with (predominantly) positive results. Level 3: Proven Effectiveness Level 1 and at least one evaluation study 4 or 5 stars with (predominantly) positive results and at least sufficient conclusiveness. **Ratings of evaluations** ***** Five Stars Randomized Controlled Trial (RCT) with follow-up (not less than 6 month, also below) **** Four Stars

#	Organisation	Framework	Use of an assessment approach or scale
			Quasi-Experimental Design (QED) with follow-up *** Three Stars RCT without follow-up, QED without follow-up. ** Two Stars "Clinical" RCT or QED with or without follow-up (not in routine context). Pre-post assessment with control-group(s) in routine context * One Star Benchmark / Norm-reference-study, Theory of Change – study No stars Participant-satisfaction assessment, Pre-post assessment without control-group, Goal-attainment study, Quality-assurance-study.
10	SUPERU	A Rating Scale for New Zealand	Yes, there are two scales. **The strength of evidence scale:** Level 0 - a pilot of a new initiative. Level 1 - Intervention is in its early stages of implementation, or planned but not yet implemented. This intervention's evidence base will be built over time. Level 2 - Typically, this intervention has been in operation for around one to three years. It has met all level 1 criteria and has been evaluated at least once. The evaluation indicates some effect, but it may not yet be possible to directly attribute outcomes to it. This intervention's evidence base will continue to be built over time. Level 3 - Typically, this intervention has been in operation for around three to 10 years. It has an established design which is consistently implemented, and quality assurance procedures are in place. It has met all the level 2 criteria, plus it has at least one evaluation that provides evidence about impact. It also has some information available that will help with implementation in new contexts. Level 4 - Typically, this intervention has been in operation for around eight years or longer and is large scale or high risk, justifying extra evaluation effort. It has met all the level 3 criteria, plus it has been replicated at least once. It has been evaluated at least twice and the evaluations provide strong evidence about effectiveness and impact, insights into how the intervention causes change, what works well or less well for different participants, and cost-benefit. There is support for implementation in new contexts. **The effectiveness scale:** Beneficial Mixed effects No effect Harmful Not applicable
11	Education counts	Best Evidence Synthesis Iteration	Yes, although not numbered. To evaluate the evidence, they consider: • which outcomes have been considered and how the evidence has linked influences to outcomes; • whether the research design was appropriate to address the research question; • whether there are untenable assumptions framing the study; • the nature of the data

#	Organisation	Framework	Use of an assessment approach or scale
			• whether the research is robust according to the method followed • who has been included in the sample or focus of the research • the role of context • credibility; • validity; • verifiability; • confidence in the findings (internal validity); • how issues of causation have been addressed; • how issues of agency and change processes have been addressed; • insights about what did and did not work; • how the findings illuminate influences on lower and higher achieving learners, and how issues of diversity are implicated in the educational processes or influences considered; • the range of findings about what did and did not influence learner outcomes, and the potential for making inter-links with these across other studies are considered; • degree of applicability to New Zealand contexts; and • specificity or generalisability of findings, with cautious attention to the significance of context.
12	Prevención basada en la evidencia	Criterios de selección de programas	Yes. (Original in Spanish)) ****Strong: Well-evaluated programs whose effect have been demonstrated through different studies. ***Moderate: Programs that having proven to be effective require more research to show that their effects maintain at long term. **Low: Programs whose effectiveness is not sufficiently demonstrated and it is necessary to investigate more about it to know the usefulness of the program. *Very Low/Not evidence: There is no evidence, or it is indirect, insufficient or contradictory according to the studies carried out with the program.
13	Darlington Service Design Lab	Standards of evidence	Yes, although the scale is not numbered. Within each of the four dimensions there are sub-categories which rank the intervention's evidence as "good enough" or "best". 1) Evaluation Quality; • Have been subjected to an evaluation that compares outcomes for children receiving the intervention with children with the same needs who do not receive the intervention; • Ideally, have been independently evaluated using a well–executed randomised controlled trial. 2) Impact; •A positive effect size, a standard measure of impact that provides comparable data regardless of the outcomes assessed; • No harmful effects or negative side–effects of the intervention. 3) Intervention Specificity; • Who is being served; • What impact on which aspects of children's health and development will be achieved; • The reason – the logic behind – why the intervention will achieve the outcome. 4) System Readiness • Having a clear indication of unit cost and staffing requirements;

#	Organisation	Framework	Use of an assessment approach or scale
			• Explicit processes to measure the fidelity of implementation and to address common implementation problems.
14	Project Oracle	Standards of evidence	Yes. Standard 1: We know what we want to achieve- Theory of Change and Evaluation Plan Standard 2: We have seen there is a change - Indication of impact Standard 3: We believe the change is caused by us - Evidence of impact Standard 4: We know why and how the change happened, this works elsewhere - Model ready Standard 5: We know why and how the change happened, this works everywhere - System ready
15	Early Intervention Foundation	The Guidebook	Yes. • Level 4 recognises programmes with evidence of a long-term positive impact through multiple high-quality evaluations. • Level 3 recognises programmes with evidence of a short-term positive impact from at least one high-quality evaluation. • Level 2 recognises programmes with preliminary evidence of improving a child outcome, but where an assumption of causal impact cannot be drawn. • NL2 (not level 2) distinguishes programmes whose most robust evaluation evidence does not meet the Level 2 threshold for a child outcome, so do not yet have direct evidence about the scale of impact of the programme at a 'preliminary' level. • NE (found not to be effective in at least one rigorously conducted study) is reserved for programmes where there is evidence from a high-quality evaluation of the programme that it did not provide significant benefits for children. This rating should not be interpreted to mean that the programme will never work, but it does suggest that the programme will need to adapt and improve its model, learning from the evaluation.
16	Nesta	Standards of evidence	Yes. A 1 to 5 scale. 1) You can give an account of impact. 2) You are gathering data that shows some change amongst those using or receiving your intervention. 3) You can demonstrate that your intervention is causing the impact, by showing less impact amongst those who don't receive the product/service. 4) You can explain why and how your intervention is having the impact that you have observed and evidenced so far. An independent evaluation validates the impact. In addition, the intervention can deliver impact at a reasonable cost, suggesting that it could be replicated and purchased in multiple locations. 5) You can show that your intervention could be operated by someone else, somewhere else, whilst continuing to have positive and direct impact on the outcome, and whilst remaining a financially viable proposition.
17	Bond	Evidence Principles	Yes. **Principles** 1.Voice and inclusion: the perspectives of people living in poverty, including the most marginalised, are included in the evidence, and a clear picture is provided of who is affected and how: 2.Appropriateness: the evidence is generated through methods that are justifiable given the nature of the purpose of the enquiry 3. Triangulation: the evidence has been generated using a mix of methods, data sources, and perspectives. 4. Contribution: the evidence explores how change happens, the contribution of the intervention and factors outside the intervention in explaining change: 5. and Transparency: the evidence discloses the details of the data sources and methods used, the results achieved, and any limitations in the data

#	Organisation	Framework	Use of an assessment approach or scale
			or conclusions. Each of the five principles has four questions and each question can be answered on a scale of 1-4. Scores for each of the questions are then added up and an overall score for the principles out of 16 is provided. Depending on the score, the principle is then assigned to a scale: 1) weak, 2) minimum standard, 3) good standard 4) gold standard.
18	Centre for Analysis of Youth Transitions (CAYT)	Standards of evidence	Yes. **Assessing impact grades (Score 0-4)**, they consider: a) Reach: the extent to which the programme attracts its intended audience and b) Significance: the effect that the programme is having on young people to influence health and wellbeing. **Level of evidence grades (Score 0-7)** 0 - Basic 1 - Descriptive, anecdotal, expert opinion 2 - Study where a statistical relationship (correlation) between the outcome and receiving services is established 3 - Study which accounts for when the services were delivered by surveying before and after 4 - Study where there is both a before and after evaluation strategy and a clear comparison between groups who do and do not receive the youth services 5 - As above but in addition includes statistical modelling to produce better comparison groups and of outcomes to allow for other differences across groups 6 - Study where intervention is provided on the basis of individuals being randomly assigned to either the treatment or the control group. 7 - Various studies that evaluate an intervention which has been provided through random allocation at the individual level. **Overall Programme Performance (Score 0-4)**
19	What Works Centre for Local Economic Growth	The Maryland Scientific Methods Scale (SMS)	Yes. Level 1: Either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups or periods. Level 2: Use of adequate control variables and either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. In (a), control variables or matching techniques used to account for cross-sectional differences between treated and control groups. In (b), control variables are used to account for before-and-after changes in macro-level factors. Level 3: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference). Justification given to choice of comparator group that is argued to be similar to the treatment group. Evidence presented on comparability of treatment and control groups. Techniques such as regression and (propensity score matching may be used to adjust for difference between treated and untreated groups, but there are likely to be important unobserved differences

#	Organisation	Framework	Use of an assessment approach or scale
			remaining. Level 4: Quasi-randomness in treatment is exploited, so that it can be credibly held that treatment and control groups differ only in their exposure to the random allocation of treatment. This often entails the use of an instrument or discontinuity in treatment, the suitability of which should be adequately demonstrated and defended. Level 5: Reserved for research designs that involve explicit randomisation into treatment and control groups, with Randomised Control Trials (RCTs) providing the definitive example. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance. There should be limited or, ideally, no occurrence of 'contamination' of the control group with the treatment.
20	Big Lottery Fund's Realising Ambition Programme	The confidence review	No. "The Confidence Framework addresses the five dimensions that Realising Ambition assessed as being essential for effective replication – service design, service delivery, ability to monitor impact, ability to determine benefit and the prospects for sustainability"'
21	Education Endowment Foundation	Teaching and Learning Toolkit	Yes. **Security of evidence criteria:** In term of : Quantity and type of study; Outcomes,; Causal inference; Consistency requirements; Effect Size requirements (from Four padlocks) **Ranking** *1. Very limited: One padlock:* Single studies with quantitative evidence of impact with effect size data reported or calculable (such as from randomised controlled trials, well-matched experimental designs, regression discontinuity designs, natural experiments with appropriate analysis); and/or observational studies with correlational estimates of effect related to the intervention or approach; but no publically available meta-analyses. 2. *Limited: Two padlocks:* At least one publically available meta-analysis 3. *Moderate: Three padlocks:* Two or more publically available meta-analyses which meet the following criteria: they have explicit inclusion and search criteria, risk of bias discussed, and tests for heterogeneity reported. They include some exploration of methodological features such as research design effects or sample size. 4. *Extensive: Four padlocks:* Three or more meta-analyses which meet the following criteria: they have explicit inclusion and search criteria, risk of bias discussed, and tests for heterogeneity reported. They include some exploration of the influence of methodological features such as research design effects or sample size on effect size. The majority of included studies should be from school or other usual settings. *5. Very Extensive: Five padlocks:*

#	Organisation	Framework	Use of an assessment approach or scale
			Three or more meta-analyses which meet the following criteria: They have explicit inclusion and search criteria, risk of bias discussed, and tests for heterogeneity reported. They include some exploration of the influence of methodological features such as research design effects or sample size on effect size. The majority of included studies should be from school or other usual settings.
22	HACT-Ideas and Innovation in Housing	Standards of evidence	No, evidence should be assessed in a seven-step process. 1) Describe; 2) Design; 3) Proceed; 4) Plan; 5) Protocol; 6) Study; 7) Findings They establish the Purpose, limitations and intended usage evidence at different levels *(Standard for Producing Evidence – Effectiveness of Interventions –Part 1: Specification, page 18)* Level 1:Exploration and Development Level 2: Effectiveness Level 3: Scaling-up
23	Conservation evidence	What Works in Conservation	Yes, although not numbered 1. Experts are asked to read the summarized evidence in the synopsis and then score to indicate their assessment of the following: Effectiveness: 0 = no effect, 100% = always effective. Certainty of the evidence: 0 = no evidence, 100% = high quality evidence; complete certainty. This is certainty of effectiveness of intervention, not of harms. Harms: 0 = none, 100% = major negative side-effects to the group of species/habitat of concern. 2. The median score from all the experts' assessments is calculated for the effectiveness, certainty and harms for each intervention. 3. Effectiveness *categorization* is based on these median values (i.e. on a combination of the size of the benefit and harm and the strength of the evidence), as and listed as follow: a. Beneficial b. Likely to be beneficial c. Trade-offs between benefits & harms d. Unknown effectiveness e. Unlikely to be beneficial f. Likely to be ineffective or harmful
24	What Works Centre for Children's Social Care	Evidence Standards	Yes. **Overall effectiveness**: looking at the consistency of effect across different research studies *Negative effect:* The balance of evidence suggests that the intervention has a negative effect (meta-analysis OR most of the studies)

#	Organisation	Framework	Use of an assessment approach or scale
			Mixed or no effect: The balance of evidence (including the pooled effect size from meta-analysis where available) suggests that the intervention has no effect overall, or studies show a mixture of effects. *Tends to positive effect:* The balance of evidence suggests that the intervention has a positive effect. There are one or more studies showing a negative effect, but either there was a meta-analysis OR most of the studies that showed a positive effect. *Consistently positive effect:* Most published studies have positive effects and none have negative effects for this outcome. Some individual studies may show no effect. However, either the pooled effect (in a meta-analysis) or most studies AND the studies involving most of the participants have a positive effect. **Strength of evidence:** looking at how confident we can be about a finding, based on how the research was designed and carried out. They overall framework is provided by the EMMIE system. They have adapted the EMMIE-Q to provide a four-point rating for strength of evidence. **0** -*Very low strength evidence*: No acceptable quality studies **1** - *Low strength evidence:* One or two acceptable quality studies **2** - *Moderate strength evidence: Three* or more acceptable quality studies. High quality review therefore possible. Between 0-3 EMMIE-Q requirements are met. **3** - *High strength evidence:* Three or more acceptable quality studies. High quality review therefore possible. Between 4-6 EMMIE-Q requirements are met including all themes marked* (see below). An acceptable quality study must have the following characteristics (definition used by the Early Intervention Foundation-EIF): 1.The sample is sufficiently large to test for the desired impact (e.g. a minimum of 20 participants in the treatment group AND comparison group). 2. The study must use valid measures. These measures should reliable, standardised and validated independently of the study. 3. Comparability of groups is addressed in selection and/ or analysis. 4. An 'intent-to-treat' design is used. 5. The study should report on overall and differential attrition. EMMIE-Q requirements 1. A transparent and well-designed search strategy* 2. High statistical conclusion validity (at least four of the following are necessary for a study to be considered sufficient)* (a) Calculation of appropriate effect sizes (b) The analysis of heterogeneity (c) Use of a random effects model where appropriate (d) Attention to the issue of dependency (e) Appropriate weighting of individual effect sizes in the calculation of mean effect sizes 3. Sufficient assessment of the risk of bias 4. Attention to the validity of the constructs, with only comparable outcomes combined and/or exploration of the implications of combining outcome constructs* 5. Assessment of the influence of study design (e.g. separate overall effect sizes for experimental and quasi-experimental design) 6. Assessment of the influence of unanticipated outcomes or spin-offs on the size of the effect (e.g. quantification of displacement or diffusion of benefit) Requirements 1-4 (highlighted by *) are considered particularly important, and are required for any review to achieve a rating of 3, which is the highest rating in the scale.
25	What Works Centre for Wellbeing	GRADE (Grading of Recommendations Assessment, Development	Yes. High quality: Further research is very unlikely to change our confidence in the estimate of effect

#	Organisation	Framework	Use of an assessment approach or scale
		and Evaluation)	Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate Very low quality: Any estimate of effect is very uncertain
26	What Works Centre for Crime Reduction	EMMIE Framework	Yes. EMMIE rates systematic review evidence against five dimensions: Effect: focuses on whether the evidence suggests the intervention led to an increase, decrease or had no impact on crime.; Mechanisms: focuses on what it is about the intervention that could explain its effect; Moderators: focuses on the circumstances and contexts in which the intervention is likely (or unlikely) to work ; Implementation: focuses on the conditions that should be considered when implementing the intervention ; Economic cost: focuses on the costs associated with the intervention, both direct and indirect, and whether there is any evidence of cost-benefit.
27	Campbell Collaboration	Campbell Collaboration Systematic Reviews: Policies and Guidelines	No
28	Community Preventive Services Task Force (CPSTF)	The Community Guide	Yes, although not numbered. The CPSTF uses the terms below to describe its findings. Recommended The systematic review of available studies provides strong or sufficient evidence that the intervention is effective. The categories of "strong" and "sufficient" evidence reflect the degree of confidence the CPSTF has that an intervention has beneficial effects. They do not directly relate to the expected magnitude of benefits. The categorization is based on several factors, such as study design, number of studies, and consistency of the effect across studies. Recommended Against The systematic review of available studies provides strong or sufficient evidence that the intervention is harmful or not effective. Insufficient Evidence The available studies do not provide sufficient evidence to determine if the intervention is, or is not, effective. This does NOT mean that the intervention does not work. It means that additional research is needed to determine whether or not the intervention is effective.
29	U.S. Department of Health & Human Services	Home Visiting Evidence of Effectiveness	Yes. HomVEE assigns a rating of high, moderate, or low to each effectiveness study according to the quality of causal evidence it provides. • High rating**:** Only randomized controlled trials (RCTs) can receive a high rating. RCTs receive a high rating if they do not have substantial design problems and if they control for differences in baseline characteristics and baseline measures. • Moderate rating**:** RCTs receive a moderate rating if they do not have substantial design problems but did not control for differences in the pre-program characteristics. Quasi-experimental design studies (QEDs) receive a moderate rating if they demonstrate that there were no differences between the treatment and comparison groups in their baseline characteristics and baseline measures, and if they control for baseline measures. • Low rating: RCTs and QEDs that do not meet the criteria for high or moderate ratings receive a low rating **Assessing evidence of effectiveness** To meet HHS' criteria for an "evidence-based early childhood home visiting service delivery model," models must meet at least one of the following criteria:

#	Organisation	Framework	Use of an assessment approach or scale
			• At least one high- or moderate-quality impact study of the model finds favourable, statistically significant impacts in two or more of the eight outcome domains; or • At least two high- or moderate-quality impact studies of the model using non-overlapping analytic study samples find one or more favourable, statistically significant impacts in the same domain. In both cases, the impacts considered must either (1) be found for the full sample or (2) if found for subgroups but not for the full sample, be replicated in the same domain in two or more studies using non-overlapping analytic study samples. For results from single-case designs to be considered toward the HHS criteria, three additional requirements must be met: • At least five studies examining the intervention meet the WWC's pilot single-case design standards without reservations or standards with reservations (equivalent to a "high" or "moderate" rating in HomVEE, respectively). • The single-case designs are conducted by at least three research teams with no overlapping authorship at three institutions. • The combined number of cases is at least 20. The HomVEE team examined and reported other aspects of the evidence for each model based on all high- and moderate-quality studies available, including the following: • Quality of Outcome Measures. • Replication of Impacts • Subgroup Findings • Unfavourable or Ambiguous Impacts. • Evaluator Independence • Magnitude of Impacts (HomVEE reported effect sizes when possible)
30	What Works Clearinghouse	Find What Works from Systematic Reviews	Yes. a) The results are sorted by *evidence of effectiveness*: ***Effectiveness Rating Key:*** it is based on the quality of research, the statistical significance of findings, the magnitude of findings, and the consistency of findings across studies Positive: strong evidence that intervention had a positive effect on outcomes. Potentially Positive: evidence that intervention had a positive effect on outcomes with no overriding contrary evidence. Mixed: evidence that intervention's effect on outcomes is inconsistent. No Discernible: no evidence that intervention had an effect on outcomes. Negative: strong evidence that intervention had a negative effect on outcomes b) The program also lets you compare interventions (Max. 5 interventions). It will allow you to see basic information for each intervention, such as grades examined, program type, delivery method, and the effectiveness rating. c) For single-case design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the consistency of demonstrated effects.
31	Center for the Study and Prevention of Violence	Blueprints	Yes. **Blueprints considers four criteria:**

#	Organisation	Framework	Use of an assessment approach or scale
			• Evaluation quality—Can we be confident in a program's evaluation • Intervention impact—How much positive change in key developmental outcomes can be attributed to the intervention • Intervention specificity—Is the intervention focused, practical, and logical • Dissemination readiness—Does the program have the necessary support and information to be successfully implemented **Program Criteria:** **Promising Programs** meet the following standards: • Intervention specificity: The program description clearly identifies the outcome the program is designed to change, the specific risk and/or protective factors targeted to produce this change in outcome, the population for which it is intended, and how the components of the intervention work to produce this change. • Evaluation quality: The evaluation trials produce valid and reliable findings. This requires a minimum of (a) one high quality randomized control trial or (b) two high quality quasi-experimental evaluations. • Intervention impact: The preponderance of evidence from the high quality evaluations indicates significant positive change in intended outcomes that can be attributed to the program and there is no evidence of harmful effects. • Dissemination readiness: The program is currently available for dissemination and has the necessary organizational capability, manuals, training, technical assistance and other support required for implementation with fidelity in communities and public service systems. European programs have not undergone the Blueprints certification process to determine dissemination readiness. **Model Programs** meet these additional standards: • Evaluation Quality: A minimum of (a) two high quality randomized control trials or (b) one high quality randomized control trial plus one high quality quasi-experimental evaluation. • Positive intervention impact is sustained for a minimum of 12 months after the program intervention ends. **Model Plus Programs** meet one additional standard: • Independent Replication: In at least one high quality study demonstrating desired outcomes, authorship, data collection, and analysis has been conducted by a researcher who is neither a current or past member of the program developer's research team and who has no financial interest in the program.
32	California Department of Social Services (CDSS)	California Evidence-Based Clearinghouse for Child Welfare	Yes. **Scientific Rating Scale** 1. Well-Supported by Research Evidence Multiple Site Replication and Follow-up 2. Supported by Research Evidence Randomized Controlled Trial and Follow-up: 3. Promising Research Evidence At least one study utilizing some form of control (e.g., untreated group, placebo group, matched wait list study) has been established 4. Evidence Fails to Demonstrate Effect Two or more randomized controlled trials (RCTs) have found the practice has not resulted in improved outcomes, when compared to usual care. The studies have been reported in published, peer-reviewed literature. 5. Concerning Practice If multiple outcome studies have been conducted, the overall weight of evidence suggests the intervention has a negative effect upon clients served;

#	Organisation	Framework	Use of an assessment approach or scale
			and/or NR. Not able to be Rated on the CEBC Scientific Rating Scale **Measurement Tools Rating Scale:** based on the level of psychometrics (e.g., sensitivity and specificity, reliability and validity) found in published, peer-reviewed journals A - Psychometrics Well-Demonstrated: 2 or more published, peer-reviewed studies have established the measure's psychometrics. B - Psychometrics Demonstrated: 1 published, peer-reviewed study has established the measure's psychometrics. C - Does Not Reach Acceptable Levels of Psychometrics: A preponderance of published, peer-reviewed studies have shown that the measure does not reach acceptable levels of psychometrics. NR - Not Able to Be Rated: Published peer-reviewed studies demonstrating the measure's psychometrics are not available. *Sensitivity: a measure of how well a test identifies people with a specific disease or problem *Specificity: a measure of how well a test excludes people without a specific disease or problem *Reliability: the extent to which the same result will be achieved when repeating the same measure or study again *Validity: the degree to which a result is likely to be true and free of bias.
33	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education	Best Evidence Encyclopedia	Yes. Strong Evidence of Effectiveness: At least one large randomized or randomized quasi-experimental study and one additional large qualifying study, or multiple smaller studies, with a combined sample size of 500 and an overall weighted mean effect size of at least +0.20. Moderate Evidence of Effectiveness: Two large matched studies, or multiple smaller studies with a collective sample size of 500 students, with a weighted mean effect size of at least +0.20. Limited Evidence of Effectiveness: Strong Evidence of Modest Effects: Studies meet the criteria for "Moderate Evidence of Effectiveness" except that the weighted mean effect size is +0.10 to +0.19. Limited Evidence of Effectiveness: Weak Evidence with Notable Effect: A weighted mean effect size of at least +0.20 based on one or more qualifying studies insufficient in number or sample size to meet the criteria for "Moderate Evidence of Effectiveness".
34	Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education	Evidence for ESSA (Every Student Succeeds Act)	Yes. The organization recognizes four levels of evidence. The top three levels require findings of a statistically significant effect on improving student outcomes or other relevant outcomes. Strong evidence: At least one well-designed and well-implemented experimental (i.e., randomized) study. Moderate evidence: At least one well-designed and well-implemented quasi-experimental (i.e., matched) study. Promising evidence: At least one well-designed and well-implemented correlational study with statistical controls for selection bias. The fourth level is a program or practice that does not yet have evidence qualifying for the top 3 levels, and can be considered evidence-building and under evaluation.
35	Society for Prevention Research	Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science.	Yes, although not numbered. Standards for Efficacy Standards for Effectiveness

#	Organisation	Framework	Use of an assessment approach or scale
			Standards for Scaling Up of Evidence-Based Interventions
36	U.S. Department of Health and Human Services	Evidence Based Teen Pregnancy Programs	Yes. **Study quality rating:** In terms of study design, Attrition, Baseline equivalence, Reassignment, Confounding factors. 1. High 2. Moderate 3. Low All impact studies meeting the criteria for a high or moderate study quality rating are considered eligible for providing credible evidence of program impacts. **The program's evidence of effectiveness (by domain) is classified as** 1. Positive impacts: Evidence of uniformly favourable impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies. 2. Mixed impacts: Evidence of a mix of favourable, null, and/or adverse impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies. 3. Indeterminate impacts: Evidence of uniformly null impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies. 4. Negative impacts: Evidence of uniformly adverse impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies.
37	National Institute of Justice	Crime solutions	Yes. a) *Programs* undergo an eight-step review and evidence-rating process. b) *Practices* undergo a seven-step review and evidence-rating process c) Then they address program and practice evaluations in an *evidence continuum* with two axes: (1) Effectiveness and (2) Strength of Evidence. *Effectiveness* is determined by the outcomes of an evaluation in relation to the goals of the program or practice. *Strength of evidence* for programs is determined by the rigor and design of the outcome evaluation, and by the number of evaluations. Rated as Effective: Programs and practices have strong evidence to indicate they achieve criminal justice, juvenile justice, and victim services outcomes when implemented with fidelity. Rated as Promising: Programs and practices have some evidence to indicate they achieve criminal justice, juvenile justice, and victim services outcomes. Included within the promising category are new, or emerging, programs for which there is some evidence of effectiveness. Inconclusive Evidence: Programs and practices that made it past the initial review but, during the full review process, were determined to have inconclusive evidence for a rating to be assigned. Rated as No Effects: Programs have strong evidence indicating that they had no effects or had harmful effects when implemented with fidelity
38	Arnold Ventures	Social Programmes that Work	Yes. Suggestive tier: Programs that have been evaluated in one or more well-conducted RCTs (or studies that closely approximate random assignment) and found to produce sizable positive effects, but whose evidence is limited by only short-term follow-up, effects that fall short of statistical significance, or other

#	Organisation	Framework	Use of an assessment approach or scale
			factors. Such evidence suggests the program may be an especially strong candidate for further research, but does not yet provide confidence that the program would produce important effects if implemented in new settings. Near top tier: Programs shown to meet almost all elements of the Top Tier standard, and which only need one additional step to qualify. This category primarily includes programs that meet all elements of the Top Tier standard in a single study site, but need a replication RCT to confirm the initial findings and establish that they generalize to other sites. This is best viewed as tentative evidence that the program would produce important effects if implemented faithfully in settings and populations similar to those in the original study. Top tier: Programs shown in well-conducted RCTs, carried out in typical community settings, to produce sizable, sustained effects on important outcomes. Top Tier evidence includes a requirement for replication – specifically, the demonstration of such effects in two or more RCTs conducted in different implementation sites, or, alternatively, in one large multi-site RCT (Is this equivalent to effectiveness?). Such evidence provides confidence that the program would produce important effects if implemented faithfully in settings and populations similar to those in the original studies.
39	Childs Trends	What Works for Child and Youth Development	No, only Eligibility Criteria for Analysis
40	Washington State Institute for Public Policy (WSIPP)	Washington State Institute for Public Policy Benefit - Cost Results	No
41	U.S Department of Justice	Office of Juvenile Justice and Delinquency Prevention Model Programs Guide	Yes. Based on the reviewers' assessment of the evidence, programs included in the Model Programs Guide and *CrimeSolutions.gov the* **evidence ratings** are: Effective : Programs have strong evidence indicating they achieve their intended outcomes when implemented with fidelity. Promising: Programs have some evidence indicating they achieve their intended outcomes. Additional research is recommended. No Effects: Programs have strong evidence indicating that they did not achieve their intended outcomes when implemented with fidelity. * The rating is given for a single study OR more than one study. * A single study icon is used to identify programs that have been evaluated with only one study. A multiple studies icon is used to represent a greater extent of evidence supporting the evidence rating.
42	University of Wisconsin Population Health Institute's	What Works for Health	Yes. **Evidence Rating:** Scientifically Supported •1 or more systematic review(s), or at least 3 experimental studies, or 3 quasi-experimental studies with matched concurrent comparisons •Strong designs •Statistically significant favourable findings Some Evidence •1 or more systematic review(s), or at least 2 experimental studies, or 2 quasi-experimental studies with matched concurrent comparisons, or 3 studies with unmatched comparisons or pre-post measures •Statistically significant favourable findings

#	Organisation	Framework	Use of an assessment approach or scale
			•Less rigorous designs •Limited effect(s) Expert Opinion •Generally no more than 1 experimental or quasi-experimental study with a matched concurrent comparison, or 2 or fewer studies with unmatched comparisons or pre-post measures •Expert recommendation supported by theory, but study limited •Study quality varies, but is often low •Study findings vary, but are often inconclusive Insufficient Evidence •Generally no more than 1 experimental or quasi-experimental study with a matched concurrent comparison, or 2 or fewer studies with unmatched comparisons or pre-post measures •Study quality varies, but is often low •Study findings vary, but are often inconclusive Mixed Evidence •1 or more systematic review(s), or at least 2 experimental studies, or 2 quasi-experimental studies with matched concurrent comparisons, or 3 studies with unmatched comparisons or pre-post measures •Studies have statistically significant findings •Body of evidence inconclusive Evidence of Ineffectiveness •1 or more systematic review(s), or at least 2 experimental studies, or 2 quasi-experimental studies with matched concurrent comparisons, or 3 studies with unmatched comparisons or pre-post measures Studies have: •Strong designs •Significant unfavourable or ineffective findings, or •Evidence of harm
43	U.S. Department of Health and Human Services	The Agency for Healthcare Research and Quality (AHRQ)	Yes. Strong: The evidence is based on one or more evaluations using experimental designs based on random allocation of individuals or groups of individuals. The results of the evaluation(s) show consistent direct evidence of the effectiveness. Moderate: While there are no randomized, controlled experiments, the evidence includes at least one systematic evaluation of the impact of the innovation using a quasi-experimental design, which could include the non-random assignment of individuals to comparison groups, before-and-after comparisons in one group, and/or comparisons with a historical baseline or control. The results of the evaluation(s) show consistent direct or indirect evidence of the effectiveness. However, the strength of the evidence is limited by the size, quality, or generalizability of the evaluations, and thus alternative explanations cannot be ruled out.

#	Organisation	Framework	Use of an assessment approach or scale
			Suggestive: While there are no systematic experimental or quasi-experimental evaluations, the evidence includes non-experimental or qualitative support for an association between the innovation and targeted health care outcomes or processes, or structures in the case of health care policy innovations. This evidence may include non-comparative case studies, correlation analysis, or anecdotal reports. As with the category above, alternative explanations for the results achieved cannot be ruled out.
44	U.S. Department of Labor	Clearinghouse for Labor Evaluation and Research (CLEAR)	Yes. Although only for Causal Studies High Causal Evidence This means there is strong evidence that the effects estimated in this study are solely attributable to the intervention being examined. This does not necessarily mean that the study found positive impacts, only that the analysis meets high methodological standards and the causal impacts estimated, whether positive, negative, or null, are credible. Currently, only well-implemented randomized controlled trials can receive this rating. Moderate Causal Evidence This means there is evidence that the effects estimated in the study are attributable at least in part to the intervention being examined. However, there may be other factors that were not accounted for in the study that might also have contributed. Causal studies that meet CLEAR evidence guidelines for no experimental designs (including randomized controlled trials with high attrition) can receive this rating. Low Causal Evidence This means there is little evidence that the effects estimated in the study are attributable to the intervention being examined, and other factors are likely to have contributed to the results. This does not imply that the study's results are not useful for some purposes, but they should be interpreted with caution. Causal studies that do not meet criteria for a high or moderate evidence rating receive this rating. They present, separately, Guidelines for reviewing quantitative descriptive studies and one for implementation studies.
45	Clearinghouse for Military Family Readiness	Continuum of evidence	Yes. Criteria to evaluate evidence: Significant Effect, Sustained Effect, Successful, External Replication, Study Design, and Additional Criteria Regarding Study Execution **Continuum of Evidence:** 1. Effective 2. Promising 3. Unclear 4. Ineffective
46	Suicide Prevention Resource Center	Evidence-Based Practices Project	Yes. **Scoring Criteria** Reviewers rated the quality of program evaluations using 10 items (See Table 1). Items were scored on a scale of 1-5 or 0-5. (A more detailed description of these items can be found in the Appendix.) 1. Theory 2. Intervention fidelity 3. Design

#	Organisation	Framework	Use of an assessment approach or scale
			4. Attrition 5. Psychometric properties of measures 6. Analysis 7. Threats to validity 8. Safety 9. Integrity 10. Utility **Classification Criteria** Classifications of programs as insufficient current support, promising, or effective were based solely upon the average scores for two items: integrity and utility. After averaging the scores of the reviewers, the lower average score of the two determined the classification level. *Insufficient current support* < 3.5 *Promising:* 3.5 - 3.9 *Effective* 4.0 - 5.0
47	National Implementation Research Network	The Hexagon: An Exploration Tool	Yes, the rating criteria is for each of the following indicators in a scale from 1 to 5, 5 being the best. **Implementing site indicators** Fit with current initiatives 1. Alignment with community, regional, state priorities. 2. Fit with family and community values, culture and history 3. Impact on other interventions & initiatives 4. Alignment with organizational structure Need 1. Target population identified. 2. Disaggregated data indicating population needs 3. Parent & community perceptions of need 4. Addresses service or system gaps Capacity to implement 1. Staff meet minimum qualifications. 2. Able to sustain staffing, coaching, training, data systems, performance assessment, and administration: Financial capacity, Structural capacity AND Cultural responsivity capacity 3. Buy-in process operationalized: Practitioners AND families **Program indicators** Evidence 1. Strength of evidence—for whom in what conditions: Number of studies, Population similarities, Diverse cultural groups AND Efficacy or Effectiveness 2. Outcomes – Is it worth it? 3. Fidelity data 4. Cost – effectiveness data Usability

#	Organisation	Framework	Use of an assessment approach or scale
			1. Well-defined program 3. Mature sites to observe 5. Several replications 6. Adaptations for context Supports 1. Expert Assistance 2. Staffing 3. Training 4. Coaching & Supervision 5. Racial equity impact assessment 6. Data Systems Technology Supports (IT) 7. Administration & system
48	National Dropout Prevention Center	Model Programs Database	Yes. Strong Evidence of Effectiveness These programs have been in existence for three years or more. They were evaluated using an experimental or strong quasi-experimental design conducted by an external evaluation team and have strong empirical evidence demonstrating program effectiveness in reducing dropout and/or increasing graduation rates and/or having significant impact on dropout-related risk factors. Moderate Evidence of Effectiveness These programs have been in existence for three years or more. They were evaluated using a quasi-experimental design conducted by an external or internal evaluation team and have adequate empirical evidence demonstrating program effectiveness in reducing dropout and/or increasing graduation rates and/or having significant impact on dropout-related risk factors. Limited Evidence of Effectiveness These programs may be relatively new programs. They were evaluated using a limited evaluation design (single group pre- and post-test) conducted by an external or internal evaluation team. They have promising empirical evidence demonstrating program effectiveness in reducing dropout and/or increasing graduation rates and/or having significant impact on dropout-related risk factors that requires confirmation using more appropriate experimental techniques. Insufficient Evidence of Effectiveness These programs require additional information before a rating category is determined.
49	National Cancer Institute	Research-Tested Intervention Programs (RTIPs)	No. Intervention evaluation and program materials are evaluated in four areas for the RTIPs review. **Research Integrity** Research Integrity reflects the overall confidence reviewers can place in the findings of a program's evaluation based on its scientific rigor. The Research Integrity rating system comprises 16 criteria scored by independent experts. Scores on each criterion are given on a 5-point scale ranging from low quality to high quality. The overall integrity score is an average of the 16 criteria reflecting the merits of the science that went into the program evaluation. **Intervention Impact** Intervention Impact describes whether, and to what degree, a program is usable and appropriate for widespread application and dissemination. This rating is determined by the Review Coordinators. Population Reach and Effect Sizes are separately rated on a 5-point scale; these ratings are then combined using the RTIPs Intervention Impact rating table to determine the impact score.

#	Organisation	Framework	Use of an assessment approach or scale
			Dissemination Capability Dissemination Capability refers to the readiness of program materials for use by others as well as a program's capability to offer services and resources to facilitate dissemination. The rating is given on a 5-point scale ranging from low quality (1.0) to high quality (5.0). Dissemination capability is measured through the assessment of three areas: •Quality of implementation materials •Training and technical assistance protocols •Availability of quality assurance materials to determine whether implementation was carried out with high fidelity to the original model **RE-AIM** RE-AIM is a five-step framework designed to enhance the quality, speed, and public health impact of efforts to translate research into practice: •Reach your intended target population •Effectiveness or efficacy •Adoption by target staff, settings, or institutions •Implementation consistency, costs, and adaptations made during delivery •Maintenance of intervention effects in individuals and settings over time
50	Strengthening Families Evidence Review	Standards of evidence	Yes. High Rating *Randomized controlled trials received a high rating if:* •The sample was randomly assigned to at least two conditions (for example, treatment and comparison groups). •The sample met the What Works Clearinghouse (WWC) standards for low levels of overall and differential attrition. •The sample members were not reassigned after random assignment was conducted. •There were no confounding factors, when one part of the design lined up exactly with either the treatment or comparison groups. •The analysis included statistical adjustments for selected measures (baseline measures of the outcomes, race/ethnicity, and socioeconomic status) if the treatment and comparison groups were not equivalent on these measures at baseline. Moderate Rating *Randomized controlled trials received a moderate rating if:* •The sample members were not reassigned after random assignment was conducted. •The sample met the WWC standards for low levels of overall and differential attrition. •There were no confounding factors. •The study included groups that were not equivalent on selected baseline measures (baseline measures of the outcomes, race/ethnicity, or socioeconomic status, but the analysis does not include statistical adjustments. OR •The study had high rates of overall or differential attrition OR sample members were reassigned after random assignment was conducted. •There were no confounding factors. •There was baseline equivalence of the treatment and comparison groups on selected measures (baseline measures of the outcomes, race/ethnicity, and socioeconomic status). •The analysis included statistical adjustments for the selected measures. *Quasi-experimental designs received a moderate rating if:* •There were no confounding factors.

#	Organisation	Framework	Use of an assessment approach or scale
			•There was baseline equivalence of the treatment and comparison groups on selected measures (baseline outcomes, race/ethnicity, and socioeconomic status). •The analysis included statistical adjustments for the selected measures. *Pre/post or other designs received a moderate rating if:* •Not applicable; these studies cannot receive a moderate rating because there is no comparison group. Low Rating •A study received a low rating if it included participant outcomes but did not meet the criteria for a high or moderate rating. Unrated •We did not rate studies that do not include participant outcomes.

OECD Public Governance Reviews

Mobilising Evidence for Good Governance

TAKING STOCK OF PRINCIPLES AND STANDARDS FOR POLICY DESIGN, IMPLEMENTATION AND EVALUATION

Governments are seeking to improve evidence-based policy making as well as trust in decision-making processes. This report offers a first global mapping of principles for the good governance of evidence in policy making, as well as standards of evidence from a significant range of OECD countries and international research bodies. Reflecting both the nature of existing practices and the various facets that contribute to quality evidence, the report takes stock of the full range of considerations involved in providing evidence across the policy cycle when designing public sector interventions, especially in the social policy area. The report also represents a first step in identifying and developing guidance at the international level in the area of evidence and evaluation.

PRINT ISBN 978-92-64-75266-5
PDF ISBN 978-92-64-85138-2