Organisation for Economic Co-operation and Development

EDU/WKP(2020)24

Unclassified **English - Or. English**

16 November 2020

DIRECTORATE FOR EDUCATION AND SKILLS

Education Policy Evaluation – Surveying the OECD Landscape
Working Paper 236

Gillian Golden (OECD)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Diana Toledo Figueroa (diana.toledofigueroa@oecd.org)
Gillian Golden (gillian.golden@oecd.org)

JT03468545

OECD EDUCATION WORKING PAPERS SERIES

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This Working Paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Table of Contents

Acknowledgements

This paper was prepared by Gillian Golden as part of the Education Policy Outlook Team, during her secondment to the OECD Directorate for Education and Skills from the Irish Department of Education and Skills.

This paper is part of the deliverables of the Education Policy Outlook for the Programme of Work and Budget 2017-18. It aims to provide a synopsis of education policy evaluation practices across OECD countries, by drawing from the Education Policy Outlook knowledge base on education policies, as well as a large array of further international evidence. The document is aimed at governments and evaluating bodies interested in gaining a broader understanding of, and building capacity for, education policy evaluation.

The paper was prepared under the supervision of Diana Toledo Figueroa (project leader of the Education Policy Outlook), and benefitted from comments as well from Daniel Salinas, Gabriele Marconi, Paulo Santiago and Andreas Schleicher. Stephen Flynn, Jonathan Wright, Savannah Saunders and Rachel Linden contributed to the editing and final formatting of this document.

Abstract

This paper aims to survey the current landscape of education policy evaluation across OECD countries and economies by examining recent trends and contextual factors that can promote more robust education policy evaluation, as well as identifying key challenges. It takes a view of policy evaluation as an activity that takes place throughout the entire policy cycle, before, during, and after a reform is implemented. It proposes a supporting framework for education policy evaluation that integrates institutional factors which can help to build robust underpinnings for policy evaluation. It also presents some specific considerations to take into account for individual policy evaluation processes.

Analysis of more than 80 evaluations across OECD education systems provides an indication of the diversity of approaches taken in the policy evaluation process. Key findings refer to the "who", "when", "what", "how", "for what" and "what next" of policy evaluation processes through a comparative lens.

Executive summary

There is an imperative on education systems to design and implement policies according to the best available evidence on value and efficacy. To do this, education systems need to build a stronger culture of policy evaluation. The evidence analysed for this working paper signals that some more emphasis is being placed on policy evaluation across OECD countries; however, definitions, concepts and practices are varied, and there is not yet a systematic strategy for institutionalising or improving education policy evaluation practices in most OECD education systems.

This paper aims to survey the current landscape of education policy evaluation across OECD countries and economies by examining recent trends and contextual factors that can promote more robust education policy evaluation, as well as identifying key challenges. It takes a view of policy evaluation as an activity that takes place throughout the entire policy cycle, before, during and after a reform is implemented. It proposes a supporting framework for education policy evaluation that integrates institutional factors which can help to build robust underpinnings for policy evaluation. It also presents some specific considerations that need to be taken into account for individual policy evaluation processes.

Key trends and challenges

This paper identifies converging developments that form the foundation of the current trends on education policy evaluation across OECD countries. Modern governance structures favour more streamlined and efficient public management, and new funding models reward evidence-based decisions and better policy results. A rapidly expanding range of evaluation and assessment processes provide education policy actors with a wealth of data. Combined with new technologies and methodologies, these can serve to underpin policy evaluation. Governments are placing more emphasis on funding and legislating for evaluation, as well as synthesising evaluation results and conveying their messages to a broader audience.

However, governments continue to face many challenges in successfully institutionalising policy evaluation, which affect both the supply of and demand for policy evaluations. On the supply side, difficulties can arise in: effectively resourcing evaluation; choosing appropriate evaluation metrics and methods; and knowing when to evaluate, how to prove the effect of a reform, and how to place the decision of the effectiveness of a reform in terms of the broader and changing needs of the education system. On the demand side, socio-political considerations can undermine the evaluation process or even discourage evaluations from taking place.

The components of policy evaluation culture

Education reform may play out differently, even in similar contexts, due to variations in the local interpretation of a policy. This is one reason for education systems being characterised as complex systems. It is vital for actors involved in the policy evaluation process to understand systemic features, relationships and externalities when assessing reform impact, as this forms a crucial part of policy evaluation.

To overcome barriers to effective policy evaluation, governments will need to mobilise strong institutional support for evaluation within the system through regulation, funding and guidance. How the foundation is laid is crucial for avoiding evaluation becoming a routine box-ticking exercise or not being developed as a structural practice at all.

Developing a mind-set of evaluative thinking across education systems goes hand-in-hand with government actions to lay the financial and regulatory groundwork for policy evaluations. This means emphasising and valuing a deeper and sometimes more critical enquiry process, being prepared to question assumptions and the status quo, and viewing mistakes and system failure as necessary parts of the learning process.

The evidence reviewed for this paper indicates that there is no "one-size-fits-all" approach to evaluating policies. However, the lack of a standard approach does not necessarily imply a lack of quality or maturity in the field of education policy evaluation. Instead, it may highlight that multiple methods of policy evaluation are needed to properly take into account the complexity and variety inherent in the education ecosystem. Governments or other evaluating bodies need to combine analysis of the policy context with a strategy and portfolio that ensures the most suitable methodological selection for evaluation is made for each reform process. The ability to formulate evaluation questions and choose the best approach is in itself a skill for which capacity needs to be built.

Education policy evaluation practices across the OECD

The specific evaluation strategies that governments may use depend on their evaluation objectives, as well as the resources and availability of relevant data and evidence. To maximise utility, budgets and methodology for individual evaluations should be proportional to the strategic importance of the reform and future plans for the reform.

Analysis of more than 80 evaluations across OECD education systems provides an indication of the diversity of approaches taken in the policy evaluation process. Key findings include:

- Evaluators of education policy take many different forms. The "**who**" of evaluation ranges from internal ministry staff and evaluation units to specific education research institutions and other policy research institutes, which often have more autonomy in their actions and publications. International organisations also play a strong role in policy evaluation, both for programmes they directly fund and as part of various peer review processes. Evaluation by committee can bring a diverse range of expertise to the process of evaluation, with many notable examples across OECD systems evident from the analysis.[1]
- The decision on **when** to evaluate is closely tied to the intended purpose of the evaluation. Most evaluations reviewed for this paper took place during or after implementation; it appears that very little formal evaluation takes place before implementation. Strengthening the evaluation of policies *ex-ante* offers the opportunity to develop a robust theory of change that takes account of all available evidence.
- Measuring policy impact begins with defining clear targets at the outset of policy development and considering **what** key metrics are associated with progress or success. A tension can exist for evaluators between identifying the measures which provide the best evidentiary standards and the feasibility and cost of getting the information necessary to use these measures. Two separate families of measures

[1] This paper acknowledges the large array of actors that can have responsibility for undertaking policy evaluation processes. For simplification matters, when referring to the actors in charge of these processes this paper may simply refer to governments (as the key stakeholder of evaluation processes) or evaluating bodies.

were identified within the analysis. One inferred the impact of reforms on outputs and outcomes, and the second examined how the reform has changed perceptions, processes and practices, such as institutional practice, systemic knowledge or capacity.

- An analysis of **how** evaluations are carried out show a diversity of methods, both quantitative and qualitative, with most evaluations combining some aspects of both. Randomised controlled trials are a "true" experiment where the researcher retains control of conditions, and are often considered as the gold standard for evaluating education policies. However, despite a political climate where there is a growing focus on "what works", the use of randomised controlled trials in education policy is still relatively rare, reflecting both technical difficulties and, in some cases, ethical concerns. Another experimental evaluation approach arises when information or conditions are available which allow for a natural experiment.
- The question of "**for what**" should always be borne in mind when planning an evaluation, as evaluation is only effective and its cost can only be justified if it promotes learning, i.e., if the results are used to inform future policies or modify existing approaches. Recent national and international initiatives provide some evidence of increasing efforts to develop effective strategies to communicate the results of evaluations to policy makers and other stakeholders.
- Finally, in terms of the **"what next"**, education policy evaluations are not by any means commonplace across OECD countries, but increasing data availability and new methodologies may provide for an expansion of evaluation activity in the future. While most evaluations reviewed in this paper cannot causally attribute changes to the implementation of the policy, evaluation processes can still help to provide actionable advice to policymakers. Policymakers looking to improve evaluation capacity in education systems can focus on creating an evaluative culture and mind-set throughout the system, building a portfolio of tools and methods that can be applied across different contexts, and ensure that the right questions are asked by evaluators.

1. Framing the discussion

1.1. Introduction: The evaluation imperative in education policy

1.1.1. *Education policy evaluation as an opportunity*

Education policy conditions in OECD countries and economies are constantly evolving. Today's children are growing up in environments where the traditional model of the family is changing, and are being educated in a context of increasing globalisation and urbanisation (OECD, 2016[1]). This is taking place while an ongoing "race between education and technology" plays out (Goldin and Katz, 2008[2]). How well our education systems rise to these challenges will determine how well prepared the children of today are for their future lives in this rapidly changing environment.

Against this background, two key factors indicate an increasing role for policy evaluation in education:

1) Disparities in how well policies work on the ground.

Increasing evidence establishes links between higher levels of education and better individual and collective outcomes. Educational attainment impacts outcomes as diverse as employment, wages, skills acquisition, health, happiness and civic engagement (Ma, Pender and Welch, 2016[3]; OECD, 2017[4])). Evidence shows that there is a relationship between better educational outcomes and certain systemic features, such as high-quality early childhood education and care, equitable school systems, increasing autonomy for some decisions, and improving accountability mechanisms. Research has noted some international convergence in terms of the types of policy actions introduced across education systems (Jakobi and Teltemann, 2011[5]; Anderson-Levitt, 2003[6]). Recent analysis of international evidence by the OECD Education Policy Outlook project (Box 1.1) also finds some agreement on the general principles of policy action to take (OECD, 2018[7]).

However, evidence also shows that even where similar policies are introduced in different countries and rolled out across education systems, what eventually happens in the classroom is often very different from what policy makers intended, with substantial variations between or even within schools (Datnow, 2005[8]), as well as across different national contexts. It is clear that policy convergence at higher levels of the system does not always equate to convergence in how these policies reach and benefit students. The policy evaluation process often provides implementation perspectives which can shed light on how and why policies produce disparate or unintended effects (OECD, 2018[7]).

2) The need for better targeting of investment in education

The social and economic costs of failed education policy are high, so it is vital that governments target investment correctly and efficiently to yield improved outcomes (Kearney and Yelland, 2010[9]). This concern for funding efficiency has only become more prevalent in the context of 2008 post-crisis constraints on public education budgets in many OECD countries and economies (OECD, 2015[10]).

Most governments in are gradually integrating evaluation and performance measurement into the budget allocation process and public finances. Performance budgeting, which requires performance measures to be included either alongside funding allocations or to directly inform funding provision, has become commonplace in the OECD in the past

decade, including in education systems (OECD, 2015[11]; OECD, 2017[12]). In a 2014 survey conducted by the OECD, representatives from 23 out of 24 countries indicated that there was at least some focus on performance in their budgeting system (OECD, 2014[13]). Moreover, the remit of supreme audit institutions (external auditors established by constitutions or supreme law-making bodies) is expanding from a focus purely on financial audits to examining the performance of expenditure or value for money of certain initiatives in many OECD countries and economies (OECD, 2016[14]).

Within this context, the policy evaluation process can be a useful tool for making policy decisions and assessing the value of reforms. Appropriate, well-executed and well-resourced policy evaluation can provide an understanding of how interventions work and how well they work, can be used to improve existing policies, can provide an evidence base for future action, and can help to justify and account for the expenditure of public funds (HM Treasury, 2011[15]).

Box 1.1. The Education Policy Outlook

Within the OECD Directorate for Education and Skills, the Education Policy Outlook (EPO) is an analytical observatory of key educational reforms in OECD countries and economies since 2013, with a focus on reforms implemented since 2000. Six policy levers frame its analysis (OECD, 2015[16]):

- **Students:** How to raise outcomes for all in terms of equity and quality and preparing students for the future.
- **Institutions:** How to raise the quality of instruction through school improvement and evaluation and assessment.
- **Systems:** How to align the governance and funding of education systems to be effective.

Through its three strands of work, the project follows the premise that knowledge of education policy is as valuable as the capacity to use it. The Education Policy Reform Dialogues promote evidence-informed conversations among countries of good education policy practices in terms of: what key education priorities education systems share to help individuals reach their potential, what has worked and, why it has worked in different education contexts. Comparative analysis of the lifecycles of education policies being implemented or already in place, across national or subnational education systems, and combining qualitative and quantitative evidence, supports this peer-learning process. Finally, the Education country policy profiles complete this series, which are reports designed for policy makers, analysts and practitioners who are seeking information and analysis of education policy, taking into account the importance of national context.

Source: OECD (2019[17]), Education Policy Outlook 2019: Working Together to Help Students Achieve their Potential, OECD Publishing, Paris, https://doi.org/10.1787/2b8ad56e-en.

There is not yet an institutionalised framework and culture of evaluation of education policies across the OECD (OECD, 2015[16]). In recent years, many education systems have

been expanding their frameworks for evaluating institutions, teachers and students, and there is a growing well of data and information available from such evaluations. However, practices for systematically evaluating the impacts of education policies appear to not yet have reached a critical mass. An OECD review of evaluation and assessment frameworks (OECD, 2013[18]) concluded that: "there is only an emerging culture of systematically evaluating the impact and outcomes of different educational interventions and again these efforts may be hindered by a lack of reliable and comparable information on student outcomes". Previous analysis of the OECD Education Policy Outlook database of reforms showed that among the significant and novel policies which had been implemented in the period 2007-2014, less than one in ten had been evaluated.

In addition, there is little comparative knowledge on evaluation at the policy level. The variety in the large number of policy evaluations examined for this paper also shows no one predominant paradigm for judging the worth of reforms; approaches vary significantly across and even within countries. OECD countries and economies participating in the Education Policy Outlook project have repeatedly indicated a strong need for more research and guidance in the area of education policy evaluation.

As the range of evaluative activities continues to expand in OECD education systems, the increasing availability of performance evidence within education systems offers the possibility to develop new education improvement processes, where a strong and rigorous culture of evaluating policy reforms becomes institutionalised. This paper outlines some of the actions that can be taken to enhance and improve policy evaluation practices in order to strengthen the basis for making decisions.

1.2. Objectives and methodology

1.2.1. *Objectives*

This review brings together research and learning from OECD projects and elsewhere to identify pathways for education systems to develop a strong policy evaluation infrastructure. It includes specific country examples from a comparative database of over 80 policy evaluation processes from 25 OECD education systems. These examples were provided by education systems through different interactions with the Education Policy Outlook project (See Box 1.1). Specific objectives are:

- **Exploring trends:** To scan the horizon for key political, contextual and methodological trends in education policy evaluation and, more generally, public sector decision support mechanisms which offer insight into the future direction of education policy evaluation in OECD countries.
- **Drawing lessons:** To present key principles derived from the evidence to form the basis of building and institutionalising a framework for education policy evaluation.
- **Mapping practices:** To develop a comparative digest of relevant examples of policy evaluation strategies employed recently in OECD countries.

The rest of this paper proceeds as follows: the remainder of this section (Section 1) clarifies the language and scope of this study and develops the motivation for studying education policy evaluations. Section 2 summarises the state of the art of education policy evaluation and proposes a framework for policy evaluation in OECD systems. Section 3 analyses some building blocks which can promote a robust evaluation culture. Section 4 examines recent education policy evaluations in OECD countries, based on different facets of the policy

evaluation process (the "who", the "when", the "what", the"how" and the "for what" of policy evaluation). Finally, a short reflection from the review is presented.

1.2.2. *Methodology*

This review uses the evidence base of the Education Policy Outlook project. The first round of the project's analytical work, which took place between 2013 and 2017, consisted of the development of a comparative report and policy snapshots for all OECD countries. A total of 32 countries also had individual policy profiles published. The project undertook a new survey exercise (through the Education Policy Outlook National Survey for Comparative Policy Analysis) during 2016/2017 with a strong focus on implementation and evaluation. The Education Policy Outlook team is supported in their work by a network of national co-ordinators who act as liaison for individual countries and validate the work programme and outputs of the project.

The three distinct components to the research strategy for this paper are:

1) A comprehensive search and review of relevant academic literature on policy evaluation practices using a search strategy on ERIC, Google Scholar and the OECD's internal search function. This covered terms such as "education policy evaluation", "education reform evaluation", "education evidence" and "policy evaluation". Based on these initial search results, a large quantity of abstracts and papers were reviewed and judged most relevant for further detailed examination for each aspect of the review. The search was generally restricted to papers published since 1995 in order to concentrate the search on the state of the art of the field, while also allowing for the relative paucity of literature. However, relevant papers with an earlier publication date were included if they had been referenced prominently in research papers published since 1995, or where there was a particular relevance for the discussion throughout the paper. In total, over 400 peer-reviewed abstracts were examined and over 180 papers were reviewed in detail.

2) A survey of guidance documents and related publications was carried out through searching the publications of four major international bodies: the OECD, the European Union, the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the World Bank. The topics discussed included policy evaluation, education policy evaluation and wider related contextual topics, such as performance budgeting, innovation and complexity in education. The search strategy followed a similar structure to component 1 above, but was also augmented by pre-existing author knowledge of projects and initiatives ongoing within international organisations which were considered relevant to the analysis.

3) A comparative digest of over 80 recent policy evaluations was developed to support the review and provide richness to the analysis. In order to be included in the digest, the evaluation must have been in relation to a specific policy or reform implemented in the country since 2000, it must have been reported to the Education Policy Outlook team as part of the policy surveys carried out in 2013 and 2016 or included in an Education Policy Outlook country profile, and information on the results and the approach of the evaluation must be publicly available. Evaluation reports were identified, collected and then analysed with a key set of analytical questions in mind to distil the features of the evaluation strategy employed and make entries into the comparative digest. Annex A contains a summary of the analytical questions used in this review process.

This digest, referred to throughout this paper as the Education Policy Outlook evaluations digest, is not exhaustive. Education policy evaluation activities in OECD countries and economies are likely to extend beyond the scope of what is included. However, as the

evaluations were reported by countries themselves in relation to prominent recent policy reforms it provides a wide-ranging snapshot of key policy evaluations which have taken place recently in OECD countries. It therefore gives an indication of the status quo of institutions, methods and outcomes of policy evaluation in OECD countries and economies at the time of drafting this paper. Together with a review of the academic literature on education policy evaluation, it provides the evidentiary basis for this paper.

1.3. Developing a clear language

1.3.1. *What is policy evaluation?*

Defining "policy"

It is important to clearly define the "policy" part of "policy evaluation", and distinguish between "processes, programs and politics" (McConnell, 2010[19]). The term "education policy" can cover ideological or strategic approaches to education by those that govern or seek to govern the system, as well as their specific reform actions or initiatives. In keeping with the wider theme of the Education Policy Outlook focus on education reforms, this analysis focuses on evaluations of specific actions and initiatives. These actions and initiatives are the activation of the strategy or ideology of the government and are reform-oriented. Examples include introducing new operational rules or regulations, increasing expenditure or changing funding arrangements, developing new curricula, or enhancing assessment models.

For example, from the school perspective, education reforms are understood as "any planned changes in the way a school or school system functions, from teaching methodologies to administrative processes" (RAND Corporation, n.d.[20]). This analysis relates then to specific actions or initiatives which have been designed and implemented to improve the education system in some way. Policy evaluation is also distinct from system evaluation: policy evaluation refers to devising and assessing the level of success of measures that address the issues identified, while system evaluation refers to the diagnosis of systemic challenges or policy priorities.

Following the pattern of other recent OECD analysis in this area (Viennet and Pont, 2017[21]), the terms "policy" and "reform" may be used interchangeably in this paper, with the understanding that where "policy" is mentioned it refers to reforms of the status quo in an education system through specific actions or initiatives.

Defining "evaluation"

It is also relevant to differentiate between what this paper understands as evaluation and other forms of judgement or monitoring of education policies. With a broad range of terminology used to describe evaluation in literature, this term can take on varied meanings depending on the context in which it is used.

Vedung characterises evaluation as "distinguishing the worthwhile from the worthless, the precious from the useless" (Vedung, 2000[22]). The OECD has also defined evaluation in relation to its utility rather than specific processes by classing it as "analytical assessments addressing results of public policies, organisations or programmes that emphasise reliability and usefulness of findings" (OECD, 2005[23]). This paper therefore broadly defines evaluation as the assessment of policies in a structured manner in order to reliably determine their merit and value according to the specific criteria. This definition distinguishes evaluation from other processes which examine policies without attempting

to assess their impact or success within a specific context, such as monitoring the "fidelity of implementation" of policies (Century and Cassata, 2016[24]).

This definition is broader than the process commonly known as "programme impact evaluation". Programmes are narrow operational processes with defined goals, budgets and timeframes that seek to address a very specific objective. Programme evaluation processes are often linked to international development activities that fund specific initiatives and have established evaluation infrastructures to assess the impact of the funding.

In terms of scope, the education policy evaluation process is generally directed towards policy makers as the primary audience for the output, as opposed to other strands of educational research aimed at improving praxis in classrooms and schools. Kelleghan and Stufflebeam (2003[25]) also identify the following three education specific considerations when evaluating education policy:

- Education policy is often used as an instrument of broader social policy. Education is a key social service in most countries. It is generally the only social service which aims to actively engage all citizens for a sustained period of their lives. This makes the audience and the stakeholder group for educational evaluations very large compared to that in many other policy areas.
- Education policy evaluation as a field has grown as a result of the development of other evaluation activities, such as student assessments and curriculum and programme accreditation. The success of education policy evaluation may depend on the availability of outputs from other evaluation processes within an education system.
- Teachers have a unique and multifaceted role to play in policy evaluation, not only as stakeholders but as objects of evaluation and as evaluators themselves.

1.3.2. *The timing of policy evaluation*

The "policy cycle" is the oldest and arguably most well recognised explanatory model for policy processes. While different versions exist in literature, the policy cycle model generally presents policy making as a disparate set of stages which follow from each other, such as design, development, implementation and evaluation. In this model, evaluation occurs at the end of the policy process and then feeds into the beginning of another policy cycle (Howlett and Ramesh, 1995[26]). This framing of the policy process has been criticised as alien to the reality of the development and implementation of policy (Fischer and Miller, 2007[27]). At the same time, it is often promoted as a necessary simplification to allow governments to navigate the complexity of the process (Bridgman and Davis, 2003[28]).

Defining evaluation as an assessment of value implies that it should be entwined throughout the policy process, since evaluation often aims to distil lessons to inform future iterations of the policy or future policies. Evaluation can continue right through the policy cycle, examining where and how the policy is adding or has added value. Policy evaluation is already recognised in many views of the policy process as being an integrated part of the entire policy process. Fischer and Miller (2007[27]) for example, outline the role of evaluation as follows:

> *The plausible normative rationale that, finally, policy making should be appraised against intended objectives and impacts forms the starting point of policy evaluation* […] *Evaluation studies are not restricted to a certain point in the policy*

cycle; instead, the perspective is applied to the whole policy-making process and from different perspectives in terms of timing.

A more holistic theoretical perception of evaluation has become increasingly prevalent in recent years. For example, the European Commission reconceptualised evaluation to promote stronger links between different evaluative processes throughout the lifetime of a policy (Smismans, 2015[29]). This paper also adopts the view of evaluation as a process which is, or should be, interwoven throughout the policy lifecycle. This reflects actual practice as reported by countries to the Education Policy Outlook project. The potential value and expected impact of the policy can also be assessed before implementation. While evaluation before implementation was not commonly reported by countries, many policies were evaluated both during and after implementation (see Section 5).

2. The state of the art of education policy evaluation

As discussed in Section 1, while there is a recognised imperative for evaluation, the evaluation of education policies is not routinely carried out. This suggests that a renewed approach to the evaluation process and new modes of generating evidence may be required to overcome the barriers to strengthening evaluation. This section examines the history, recent developments and challenges to education policy evaluation. It then presents a framework which could be used to develop more effective policy evaluation processes. Later sections will examine in more detail the individual components of this framework to indicate how systemic principles and practices could develop to meet the evaluative needs of 21st century education systems.

2.1. Where we are and how we got here

2.1.1. *A (very) brief history of education policy evaluation*

The history of education policy evaluation is encompassed by the wider history of government policy evaluations. Evidence suggests that informal evaluations of policy initiatives have been taking place in some form for more than 150 years (Madaus, Stufflebeam and Scriven, 1983[30]). Beginning in the 1960s amid greater union mobilisation and increasing calls for public accountability, the specific field of education policy evaluation began to become more professionalised and expanded both conceptually and methodologically. There was a substantial increase from the mid-1970s onwards in the volume of professional literature, manuals and specific journals for education policy evaluation (Worthen and Sanders, 1987[31]).

Despite the early growth of the field, by the 1980s, low levels of systematic, effective evaluation of education reforms were well recognised as an ongoing challenge for education systems (Comfort, 1982[32]). The 1990s onwards saw an increasing focus internationally on policy evaluation as a tool to improve educational quality, competitiveness and equity (Stufflebeam, 2001[33]). This period also coincided with a move towards greater international co-operation, as evidenced by the first international projects for comparative education indicators, which have become a key underlying input to assist evaluations in many countries (OECD, 1994[34]). Impact evaluations for international development programmes (including education initiatives) have also grown both in number and sophistication (Cameron, Mishra and Brown, 2016[35]).

Despite the expansion of education policy evaluation over previous decades, it appears that education policy evaluation is still not institutionalised and that evaluative activities in education policy are still not practiced sufficiently (Kitamura, 2009[36]) (OECD, 2015[16]). However, as evaluations slowly permeate into systemic practice, debate is shifting towards the issue of quality and how best to ensure that reforms are evaluated to a high enough standard so that useful and actionable conclusions can be drawn (National Audit Office, 2013[37]; OECD, 2007[38])

2.1.2. *Recent trends promoting more effective evaluation*

As education systems evolve in response to rapidly changing economic and societal demands, more adaptive approaches to regulation and evaluation are needed to contribute

to effective governance. There is also some increasing recognition of the importance of integrating evaluation into all stages of the policy cycle, as discussed in Section 1.

Specific trends which could contribute positively to the coverage and quality of education reform evaluation are explored below.

A new period of smarter public management

The 2000s saw a new period of public administration that aimed for better value for money through the greater coherence and integration of public services. During this period an increased focus on standard setting emerged, as well as a greater role for specialist policy evaluation functions (OECD, 2015[39]). As discussed previously, performance budgeting is also becoming more prevalent across OECD countries, and many supreme audit institutions are broadening their focus to examine how programmes are working. Various recent country level frameworks that aim to increase the positive impact of public investments also favour the policy evaluation process. For example, New Zealand adopted a social investment strategy to reduce long-term costs by investing earlier in improving outcomes for vulnerable groups, which requires the robust evaluation of social policies (Boston and Gill, 2017[40]).

Other recent innovative funding models also aim to elevate the role of policy evaluation. Social impact investment models are one such rapidly growing area targeting initiatives that provide a social as well as financial benefit (Wilson, 2014[41]). According to the evidence collected, there is an urgent perceived need to develop more robust evaluative measures and standards to assess social impact and therefore the return on investment (OECD, 2015[42]). As social investment funds are often funnelled towards educational initiatives, education policy evaluation is likely to be strengthened in importance as a result.

A broader range of evaluation frameworks and evidence across education systems

In individual countries and across international organisations there are increasingly sophisticated mechanisms for analysing education systems and tracking progress and outcomes of students (OECD, 2013[18]). Moreover, modern international assessments also provide a range of contextual information, which is useful for policy makers to gain a comparative insight into which policies may be working well and which may need to be improved. As the range of evaluative measures across the system increases in size and sophistication, policy evaluation can become easier to undertake given the availability of underlying evidence to support the process.

The expansion of technological and methodological capacity

The range of data available to policy makers is rapidly expanding. For example, the increasing availability of policy relevant big data and open data[2] offers opportunities for greater insight into the impact of policies. This directly leads to an increase in the volume of data available for policy evaluation purposes (European Commission, 2010[43]).

In most cases the capacity to use such data is still in its infancy (Bakhshi and Mateos-Garcia, 2016[44]). However, computational power, analytical capacity and new technologies

[2] Big data is data which is "characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" (De Mauro, Greco and Grimaldi, 2016[153]). Open data refers to data which can be freely accessed, used, modified and shared for any purpose (Open Knowledge International, n.d.[154]).

continue to become available, which can make sense of the exponentially increasing volume of data and information. Large-scale computational structures are being built that would not have been possible previously, such as agent-based models, microsimulation of policy initiatives, geospatial models for education planning, and social network analysis (see Section 4). As data availability and capacity expands, methodologies for causal evaluation of policy impact have also become more widely used (Schlotter, Schwerdt and Woessmann, 2011[45]). This greater sophistication of data and methods of analysis bring new opportunities for policy evaluation, as well as new expectations in terms of how policy makers can be held accountable.

More emphasis on synthesis and brokerage

More attention is being paid as well to interpreting and using evaluation results to improve policy learning. Diverse practices have been developed in recent years which support the learning process, such as quality standards for systematic reviews and synthesis evaluation (Olsen and Reilly, 2011[46]). Meta-evaluation, a method for judging the quality and utility of evaluations, is a vital quality element to ensure that *"*evaluations provide sound findings and conclusions; that evaluation practices continue to improve; and that institutions administer efficient, effective evaluation systems" (Stufflebeam, 2001[47]).

There is a growing tendency to applying quality standards to evaluations and distilling their results. This is reflected by an increase in brokerage organisations and clearinghouses. These organisations compile the results of the evaluations carried out across education systems into formats that are more easily accessible for various audiences. In some cases, they assess the quality of the evaluation processes themselves (see Section 4). The use of other formalised methods for evaluating evidence for policy, such as the "best evidence synthesis" method (Slavin, 1995[48]), have also begun to feature in the education policy development process in some OECD countries.

2.1.3. *Challenges remain*

Despite the progress achieved and the positive trends outlined above which promote greater education policy evaluation, evidence shows that governments still face a number of key methodological, technical and practical challenges when attempting to evaluate the outcome of education policies. In addition to a possible challenge of ensuring sufficient buy-in of education actors towards review and evaluation, these challenges may prevent governments to undertake policy evaluations more systematically. The key challenges are explored below.

The temporal challenge

Reform takes time to bed down, and all the outcomes of reforms are not always measurable within the evaluation timeline (Levin, 2001[49]). The course of a reform may have multiple trajectories and stages, and impacts may therefore diverge across time periods and among different levels of the system (Pollitt and Bouckaert, 2017[50]). This complexity presents difficulties for evaluators in terms of measuring the right outcome in the right place at the right time. It also raises broader questions around the definition of evaluation and the understanding of the time period it should cover.

The question of sustainability of policies and practices over a longer period of time is often not addressed within evaluations. Often, it is unclear how the evidence of impact gathered at one point in time during a policy evaluation can be leveraged to predict likely long-term effects of the policy (Chelimsky, 2014[51]).

The resourcing challenge

Evaluation is also resource intensive and often requires specialised tools and skills sets, or the involvement of a third-party evaluation institution. Many of the evaluations analysed in Section 4 involved large-scale surveys and data cleaning as part of the process. Limited capacity and resources can cause evaluation to become the "poor relation" of the policy process: policy makers may prefer to prioritise finite budgets on implementation rather than evaluation, or may lack the capacity to identify how to evaluate policies effectively (Guskey, 2000[52]). However, without evaluative evidence it is difficult to judge whether resources have been spent effectively. Efficiency reviews carried out in many national systems, such as value for money (VFM) audits, often criticise the lack of evaluation evidence (National Audit Office, 2013[37]).

The causality challenge

The causal impact of education reforms can also be difficult to measure, and incorrect causal conclusions can have significant costs for education systems (Cook, 2002[53]). Policy evaluation reports collected indicate that indirect associations are often made without proving causal links between the policy and changes to outcomes (see Section 4). Many general methodological challenges also add to the difficulties of determining causal effects, such as identifying counterfactuals or determining causal parameters (Heckman, 2008[54]). As a result, it is challenging to translate and distil the results of the evaluation in an evidential manner to contribute towards future policy development.

The socio-political challenge

Barriers to creating an effective evaluation culture can come from conflicts related to creating demand for policy evaluation within the system (Rutter, 2012[55]). Robust evaluation can be hampered by a lack of political will; evaluations therefore may not always be rationally planned but may reflect the socio-political dynamics of the particular educational context (Bamber and Anderson, 2012[56]). Evaluation results may become available during an inconvenient time in the political cycle and carry political risks, for example a political party can be tied closely to a particular policy position and therefore linked to an underperforming reform.

In the same way, an incumbent government may feel less enthusiastic about acknowledging positive evaluation results for a policy introduced and undertaken by a different government administration. Such dynamics can act as a barrier to institutionalising reform evaluation in education systems. They can also foster a strong aversion to policy failure, which evidence shows is a challenge to innovative policy developments in the public sector (OECD, 2015[57]). Fear of failure can contribute to a lack of motivation for encouraging systemic practices that tend to highlight inadequacies, such as policy evaluation.

2.2. A framework for education policy evaluation

Policy makers need to take into account a range of considerations when trying to design evaluation processes to generate the best possible evidence. These considerations include: when in the policy cycle the evaluation should take place, why the evaluation is being carried out, what should be measured by the evaluation, who should do it, how should it be done, and for what purpose it will be used. The answers to these questions are driven by the objective of the policy, but also by the availability of the relevant resources to address the questions. In the absence of strong capacity to address these considerations

strategically, the evaluation results may suffer in quality or be unsuitable for the purpose intended.

At the same time, outside of the design of any one policy evaluation, mounting evidence shows that an understanding of the context into which reforms are implemented is a key success factor. Policy implementation needs "conducive contexts" in order to be successful (Viennet and Pont, 2017[21]). The same education reforms may be "enacted" differently in similar contexts due to differences in interpretation of the intention and mechanics of the policy (Braun, Maguire and Ball, 2010[58]). Thus, policy makers and evaluators need as much insight as possible into the systemic features, relationships and externalities which comprise the context of an education reform.

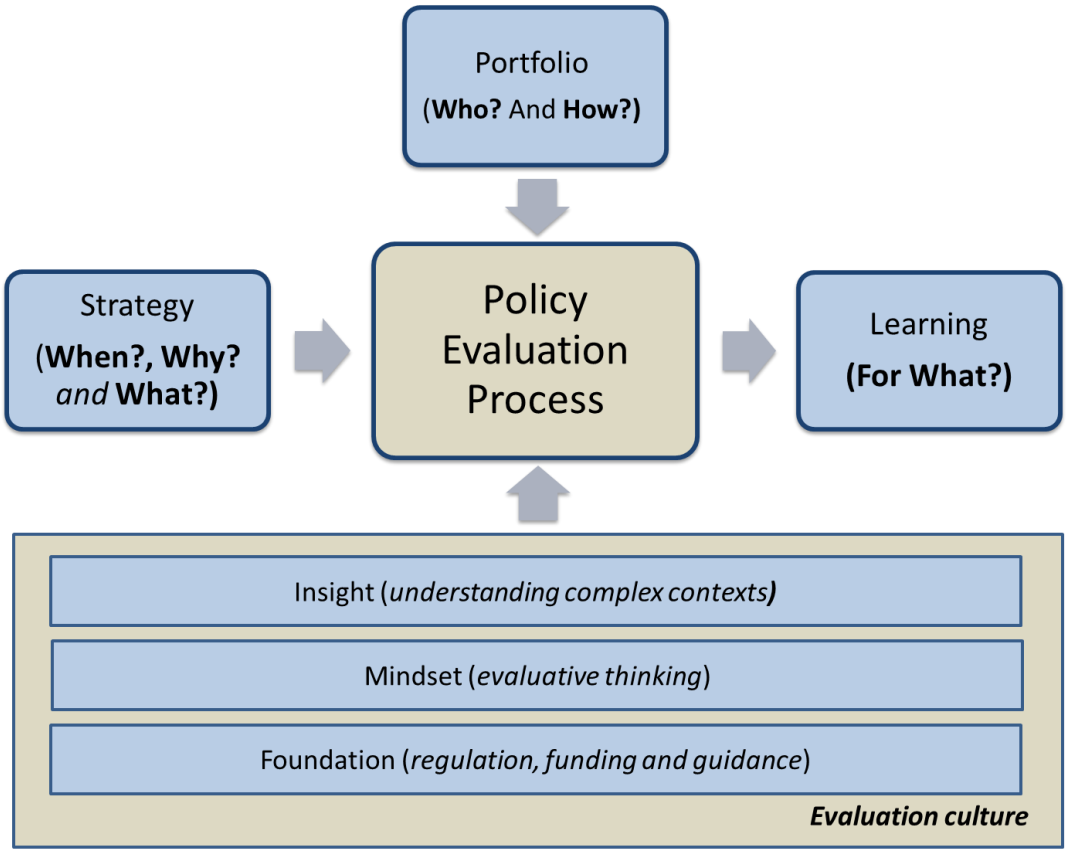
To counteract the methodological, resource and socio-political challenges outlined in the previous section, policy evaluation needs to become an integral and beneficial part of the policy-making process. This requires governments to mobilise strong institutional support for policy evaluation within the system in terms of regulation, funding and guidance. Building on this foundation, embedding a systemic evaluative mind-set can pay off in terms of increasing the quality and regularity of evaluations. Strategically, maintaining a portfolio of versatile approaches to evaluation can also ensure that a variety of evaluation tactics can be deployed as the context evolves.

In summary, signals in the policy environment, recent country practices, and relevant literature point to two related pathways through which education systems can improve policy evaluation processes and results:

- Developing the following capacities can improve the quality of individual evaluation processes: developing strong foundations for evaluation activity, improving the ability to assess policy contexts, and fostering a cultural mind-set which favours evaluation.
- For individual education policies, having a comprehensive evaluation strategy and access to a diverse portfolio of evaluators and methods can pay off. Governments should give deep consideration to the questions of who, what, why, how and for what, to ensure that the most suitable modality for evaluation is chosen.

Figure 2.1 presents a visual model of how these pathways could be incorporated and combined into a robust framework for policy evaluations within an education system.

Figure 2.1. A framework for the education policy evaluation process



This framework is used as the basis of the analysis presented in the remainder of this paper. Section 3 explores the elements of the evaluation culture (insight, mind-set and foundation), while Section 4 looks at the questions of "who", "how", "when", "why", "what", and Section 5 examines the element of "for what", using examples from the education policy outlook evaluations digest.

3. Building blocks for a strong evaluation culture

3.1. Insight: Working to understand complex contexts

Policy ecosystems are comprised of core policy priorities for improvement, the existing context of the system in which policies interact, key actors (through their engagement and capacities) and the key systemic arrangements needed to make policies feasible and effective (OECD, 2018[7]). Education policy ecosystems are well recognised as complex systems, and the application of complex systems theory to education has become more prevalent in recent years (Davis and Sumara, 2008[59]; Mason, 2008[60]).

Growing decentralisation and school autonomy means a greater number actors with decision-making abilities in OECD education systems, and increasing numbers of systemic relationships and interactions. This presents challenges for the overall governance of the system, and new thinking is required in terms of policy delivery across increasingly heterogeneous systems (Burns and Köster, 2016[61]).

Complex systems, as implied by the name, are generally difficult to describe and analyse in their totality. They are: "that which the mind cannot easily comprehend and whole constituent parts cannot be easily disentangled" (Alhadeff-Jones, 2008[62]). They are also difficult to govern: the OECD New Approaches to Economic Challenges Report concluded that complex systems cannot be successfully steered with simple linear mechanisms (OECD, 2015[10]). An overly simplified approach to governance is therefore less effective in complex policy ecosystems and might explain why policies can produce undesired or unexpected results, even when scaling up from a successful pilot.

It is often commented that education reforms do not easily take hold in classrooms, and that the context into which they are introduced matters. Not taking into account these initial conditions in complex education systems, nor understanding how they might lead to variability in policy implementation, can spell failure for the policy in question (Trombly, 2014[63]). A strong systemic understanding can help identify possible small adaptations that could or did nudge the system in desired directions. It can also highlight relationships or features of the system that are most associated with positive impacts.

However, to date very little of the insight gained on system complexity has been converted to actionable tools and instruments for policy makers to use. Few "practitioners" exist with the ability to apply complex systems thinking to education policy. Nevertheless, research is beginning to emerge that offers some practical direction on ways to identify and represent key contextual elements in a complex education system (Box 3.1).

Box 3.1. Five principles for evaluating policies in complex contexts.

Recent research has aimed to develop practical approaches to enhance insight into complex education policy contexts. Reforms operating in complex systems require an approach to evaluation which respects and takes into account this complexity, while accepting that achieving complete understanding of a complex system is unlikely.

Preskill and Gopal developed propositions for evaluating complexity, many of which are relevant to the process of assessing the policy context of a reform, including a focus on identifying key systemic relationships, dynamics and feedback loops (Preskill and

Gopal, 2013[64]). OECD research on governing complex education systems also identified important governance factors, including taking a whole-of-system view, understanding network structures, building trust and ensuring high capacity within the system to adapt to changing circumstances (Burns and Köster, 2016[61]).

Van Geert and Steenbeek take a more pragmatic view, arguing that complex educational systems in reality cannot be fully described as they encompass everything from the historical development of the system to the specifics of teacher-student knowledge transfer in the individual classroom (Van Geert and Steenbeek, 2014[65]). The community of policy makers, along with other communities within education systems, are therefore obliged to create simplex models of the system (i.e. a simplified mental model of the system which informs their everyday practice).

During policy implementation, the differences in the simplex models among these communities can clash, causing adverse impacts to the effectiveness of the policy.

Taking this research into account, the following five principles can be taken into account when evaluating in complex systems:

Principle 1: Aim to understand and describe the structure of the whole system, including structure, formal and informal networks, relationships and interdependencies, while accepting that this will not be fully possible.

Principle 2: Distinguish between system aspects which are known and describable (for example, the system history and known organisational structures) and those where the dynamics may not be fully describable (influencing processes, informal relationships, power balances, trust levels, capacity for change, risk acceptance).

Principle 3: For the dynamics which are not fully describable, aim to identify the variety of simplex models which are currently in use by different stakeholders, and how these might promote or inhibit reform success.

Principle 4: Identify likely feedback loops, information channels and points of influence within the system which have the potential to interact with the policy reform either positively or negatively.

Principle 5: Identify areas of stability and instability in the system that have particular importance for considering the possible effects of a new reform. Consider how policy innovations are likely to, or have already, interacted with these areas, and whether linear or non-linear models best represent the system dynamics.

Source: Developed by the author based on: Preskill, H. and S. Gopal (2013[64]), *Evaluating Complexity: Propositions for Improving Practice,* FSG, Boston, www.pointk.org/resources/files/Evaluating_Complexity.pdf; Burns, T., Köster, F. and M. Fuster (2016[61]), *Education Governance in Action: Lessons from Case Studies*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264262829-en.; Van Geert, P., and H. Steenbeek (2014[65]), *The good, the bad and the ugly? The dynamic interplay between educational practice, policy and research*, Complicity: An International Journal of Complexity and Education, The University of Alberta, Alberta, https://journals.library.ualberta.ca/complicity/index.php/complicity/article/view/22962/17093.

3.1.1. *Microdata infrastructures for contextual analysis*

Administrative microdata has a particularly strong potential to enhance policy evaluation, given the large databases available in OECD countries. Microdata from different administrative sources have been linked in many OECD countries and economies in order to develop decision support data infrastructures (Table 3.1). This type of infrastructure has the potential to serve as the backbone of the evidence base both for introducing new

education policies and evaluating them. For contextual analysis, high-quality data infrastructures can efficiently inform the assessment of the status quo. In some domains, they are used for simulating and "testing" the likely impact of policies (such as resource planning and allocation). These data sources can also be valuable for examining the effects of policies after implementation, taking into account specific sub-contexts within a system.

While there is a greater recognition of the significant potential of harnessing microdata to gain new insights into the efficacy of policies, for the most part the development and full exploitation of this type of decision support infrastructures is at a nascent stage. However, at a national level some notable examples of exploiting micro-level data to support education policy decision making and policy evaluation already exist in OECD countries and economies (Table 3.1).

Table 3.1. Decision support data infrastructures for education policy

National level data infrastructures to support education policy in OECD countries

Country	Model name	Description
Denmark	DREAM	The educational component of the DREAM model predicts future education levels, transition trends and labour market participation based on the characteristics and behaviour of the current Danish population (derived from register data), which provides a rich set of information to assist in education decision support.
Australia	AURIN	AURIN is a national collaboration delivering e-research infrastructure to empower better decisions for Australia's urban settlements and their future development. Data from over 60 different providers are combined for contextual analysis and evidence-based decision making.
New Zealand	Integrated Data Infrastructure (IDI)	The IDI is a linked longitudinal dataset that combines unit-record administrative information from a range of agencies and organisations. The IDI is maintained by Statistics New Zealand under strict privacy and confidentiality protocols. All records within the IDI are anonymised and any statistical outputs that use IDI data go through a confidentialisation process to ensure that the privacy of individuals is fully protected.
Germany	GeRDI (Generic Research Data Infrastructure)	The aim of GeRDI is to enable all scientists in Germany, especially those who hold only small amounts of data, to store, share and re-use research data across disciplines. This new project is rolling out over a three year period from 2016, the results will inform the creation of a National Research Data Infrastructure in Germany.
United Kingdom	Administrative Data Research Network (ADRN)	The ADRN was launched in 2013 with the aim of enabling researchers to access de-identified, linked administrative data from various sources for research purposes. Researchers are trained and accredited and go through an approvals process before being allowed to access the data in a secured environment. All research proposals must demonstrate a social benefit and researchers must commit to making their results publicly available in order to access the data.

At the international level, some projects and skills centres are in development to exploit the value of administrative data. For example the OECD Science Technology and Innovation microdata lab collects and links administrative and commercial micro-level datasets (OECD, 2014[66]). A European Union Competence Centre on Microeconomic Evaluation (CC_ME) also opened in 2016 with a focus on building skills for evaluation design and methodology using microdata (European Commission, 2017[67]).

Significant challenges remain before the potential of these techniques for education evaluation can be fully realised. The administrative databases currently in existence in most

OECD countries and economies were generally not designed with research in mind. As a result, serious technical barriers exist in many cases to linking administrative data from disparate databases. Strong legal impediments also exist in many jurisdictions to using administrative microdata for non-administrative purposes. There is political caution in providing access to microdata to researchers given privacy concerns and strong digital rights movements in many OECD countries. Further developmental work may be required in some jurisdictions to provide an appropriate legal basis for using microdata for research purposes.

3.2. Mind-set: Evaluative thinking across the education policy cycle

Evaluation can be considered as process based using the paradigm of the evaluation policy, i.e. the rules and principles that guide evaluative actions (Trochim, 2009[68]). However, it can also be described as a collective mind-set that arises through the development of the overall evaluative culture. Evaluative culture is:

> *…an organizational culture that deliberately seeks out information on its performance in order to use that information to learn how to better manage and deliver its programs and services, and thereby improve its performance. Such organization values empirical evidence on the results—outputs and outcomes—it is seeking to achieve. (Mayne, 2008[69])*

This implies evaluation as a mind-set; rather than a process or activity. Thus, evaluative thinking is a "way of doing business". Embedding such thinking across an organisation is the opposite of treating evaluation as a simple "box-ticking" exercise (Griñó et al., 2014[70]). Developing effective evaluation capacity therefore does not only imply simply carrying out more and broader evaluations. It involves developing the capacity to employ evaluative thinking across the policy process and effectively using evaluation as the link between policy design, development and implementation. The increasing centrality of this concept in the evaluation field, and the importance of developing capacity for evaluative thinking, are also highlighted by Buckley et al. (2015[71]), who define evaluative thinking as follows:

> *Evaluative thinking is critical thinking applied in the context of evaluation, motivated by an attitude of inquisitiveness and a belief in the value of evidence that involves identifying assumptions, posing thoughtful questions, pursuing deeper understanding through reflection and perspective taking, and informing decisions in preparation for action.*

While specific evaluation processes are often left to specialists, evaluative thinking is something that everyone involved in the reform process can engage in, and capacity can be built to embed the mind-set across the system (Griñó et al., 2014[70]). Evaluative thinking can involve analysing and questioning the assumptions underpinning policy actions that are often taken for granted. It can consist of the critical examination of relevant evidence to avoid the "evidence blinkers" modality of only developing policy where there is already supporting evidence. In short, it implies developing the skills to make the types of critical inquiry into policy processes that will yield maximum insight. The OECD has outlined a series of key ideas for evaluating education systems that can help to promote more innovative and effective learning environments (Box 3.2).

There is a clear link between the principles of evaluative thinking highlighted in Box 3.2, and those provided in the previous section for analysing the complex context into which a reform will be or has been implemented. Both sets of principles privilege a deeper and more critical enquiry process and a search for insight into how the reform process is likely to interact with different contextual scenarios. Within educational institutions, fostering evaluative thinking is a promising way to promote the use of data and evidence to improve provision for individual students (Wyatt, 2017[72]). It also can serve as a means to involve all stakeholders in the improvement of policies through collaborative inquiry processes (Earl and Timperley, 2015[73]).

Box 3.2. The OECD perspective: Evaluative thinking for greater educational success

Recent OECD work highlights the importance of evaluative thinking in developing educational innovations that are responsive and impactful. The Innovative Learning Environments Handbook (OECD, 2017[74]) draws together a number of basic ideas which can embed and promote evaluative thinking. The handbook conceives of evaluative thinking not as part of an unstructured innovation process, but as a discipline combining sequential steps with feedback loops in order to ensure that all aspects of the innovative process are comprehensively questioned and evaluated. According to the handbook, core opportunities where evaluative thinking can be engaged include:

- When **defining an educational innovation** by considering the intention, roots, philosophy and expected impact. These considerations can be revisited as more evidence becomes available later in the process.
- When engaging **stakeholders** by involving them in the process of critical thinking about the innovation and considering their viewpoints and cultural differences.
- When **designing evaluation processes** by developing a clear vision of what is required from the evaluation process and thinking deeply about the range of questions the evaluation needs to answer, from both an internal viewpoint and to ensure external accountability.
- When **using evidence** by developing theories of action and collaboratively planning the evidence-gathering methods in advance, considering whether the evidence gathered is fit-for-purpose, and giving more and deeper contemplation to evidence interpretation and insight rather than privileging evidence gathering.

Source: OECD (2017[74]), The OECD Handbook for Innovative Learning Environments, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264277274-en.

3.3. A strong foundation through regulation, funding and guidance

Every education system needs a functioning evaluation infrastructure that is fit for purpose and an evidence delivery system that can feed into the policy process. Making this a reality requires strong foundations to provide appropriate resources, regulation and guidance to the process of policy evaluation.

3.3.1. *Resourcing education policy evaluation*

Education evaluation and education research in general are often viewed as underfunded compared to other areas of social policy research, although concrete comparative indicators are not available on an international scale (Burkhardt and Schoenfeld, 2003[75]). Furthermore, a tension exists between growing calls for more evidence-based decision making and resourcing in education, and trends toward the decentralisation of governance. This decentralising trend can lead to fragmentation in the procuring and funding of evaluation (OECD, 2007[38]). It is therefore critical that systemic arrangements are in place to provide adequate access, resources and guidance for policy evaluation processes.

Many options exist for channelling funding towards policy evaluation. One traditional approach involves setting a proportion of the policy budget aside specifically for the evaluation process. This is often a feature of specific public policy initiatives funded by international bodies, such as the European Commission, which mandate policy evaluation in the funding conditions. However, there is often not a clear link between the cost of implementing and the cost of evaluating a reform. Appropriate funding levels can depend on the strength of "proof" required of the reform impact, as well as the evidence and data available or built in to the implementation process and the amount of measurement precision required (Lagarde, Kassirer and Lotenberg, 2012[76]).

Budgets for individual evaluations should also be related to the strategic importance of the reform. One option for rational funding is through the resourcing of specialist agencies or bodies mandated to evaluate education policy. Many governments have developed initiatives in this area in recent years, either through developing more general policy evaluation institutions that also focus on assessing education policies, or specific institutes for education policy evaluation (see Section 5).

Finally, investment in internal administrative and analytical capacities within national education ministries can pay off for developing evaluation capacity. Many education ministries and/or departments have internal monitoring and evaluation units, or units concerned with policy analysis. However, silos can exist between these units and systemic data and evidence generated elsewhere in the system, such as in education management and information systems (EMIS), teacher payrolls and achievement data. Such silos can seriously hamper the effective evaluation of the actions of national ministries (Hua and Herstein, 2003[77]). Investing in technical or organisational solutions which break down these barriers can greatly enhance the effectiveness of evaluation processes and help build a stronger evaluative culture among education policy makers.

3.3.2. *Regulation and guidance*

Connecting evaluators with the knowledge and guidance they need is essential to avoid poorly designed evaluations which do not meet the needs or expectations of policy makers. There are many sets of general guidelines and frameworks for policy evaluation available from academic sources or in the body of literature to guide the evaluation profession. Governments in some countries have also produced official guidance for evaluating reforms, although few examples of official education specific evaluation frameworks exist (see Box 3.3 for an example from New South Wales). Notable official general public policy evaluation guidelines include:

- The **Instructions for Official Studies and Reports Norway** (*Utredningsinstruksen*) have been central to the Norwegian system of assessment,

submission and review procedures for official studies, regulations, propositions and reports to the *Storting* (Parliament) since 2000. In 2016, the instructions were revised to make evaluations simpler, clearer and more complete, with overly complex rules removed. Minimum requirements are laid out in six questions that each evaluation report must answer. Proportionality is also an important component: resources invested in the evaluation should be proportional to the resources invested in the initiative itself. Other changes aim to promote the involvement of ministries and other stakeholders earlier in the assessment process (DFØ, 2016[78]).

- The **Magenta Book** issued by the **United Kingdom** Treasury was revised in 2011 to provide broader ranging guidance and advice on evaluation processes. It is aimed at analysts and policy makers across all levels of government with a view to improving policy design and implementation in order to promote effective evaluation, as well as improving evaluation processes themselves. It highlights the benefits of robust, proportionate evaluations for policy and also presents technical guidance and best practice standards for those practically involved in the evaluation process (HM Treasury, 2011[15]).

Although ethical guidelines targeted at reform evaluators exist, evaluations tend to be more preoccupied with technical and methodological issues (Bechar and Mero-Jaffe, 2014[79]). Still, there is a need to ensure that evaluations meet ethical standards and give due consideration to all participants in the evaluation process. A responsible and truthful approach by the evaluators, informed by ethical standards, can help to ensure that the rights and concerns of all involved in the process are respected, which can improve evaluation quality.

There must be no doubts around the independence of evaluators, the development and circulation of reports, and the privacy of information from individuals involved in evaluations (Burgess, 2005[80]). Evaluations can create conflict and resistance in those who are being evaluated. This evaluation anxiety may reflect fear of critical judgement, previous negative experiences with the evaluation process, and concerns about changes to working conditions or power structures that may occur as a result of the evaluative process (Taut and Brauns, 2003[81]). It is important for evaluators to be aware of these concerns and actively mitigate against them through building trust and facilitating dialogue, communicating possible benefits of the process for stakeholders, and stressing the focus of the evaluation on the policy rather than the people (Taut and Brauns, 2003[81]).

Box 3.3 A modern policy and framework for evaluation in education: New South Wales, Australia

An education specific policy evaluation framework was introduced in 2014 in New South Wales, Australia, which recognises evaluation as "an integral part of managing government programs at every stage of the policy cycle", and aims to strengthen evaluation practices to improve performance and accountability. Key principles for evaluation laid down as part of the framework are:

- Evaluation will be planned early during the design of programmes.
- Evaluation will be appropriately resourced as part of programme design, taking into account what is feasible and realistic to achieve within time and budget constraints.
- Evaluation will be rigorous, systematic and objective, with appropriate scale and design.
- Evaluation will be conducted with a suitable level of expertise and independence.
- Stakeholders will be identified and actively involved in the evaluation process.
- Evaluation will be timely and strategic to influence decision making.
- Evaluation will be transparent and open.

The framework provides guidelines for how to prioritise evaluation and choose the most appropriate evaluation strategy given the reform aims and stage of implementation. The five criteria considered most important to prioritise and focus evaluation activity include: 1) the financial scale of the programme; 2) the strategic alignment of the programme to government priorities; 3) external requirements (for example those funded from other sectors); 4) the existing evidence base (with priority given to those that have not been recently evaluated or with an inadequate evidence base to assess the policy); and 5) methodological considerations as to whether the programme can be effectively evaluated. Guidelines are also provided for judging the appropriateness and scale of the evaluation activity against the scale and importance of the programme itself, and for the appropriate governance arrangements for the execution of the evaluation plan.

Source: New South Wales Department of Education and Communities (2014[82]), Evaluation Framework, https://education.nsw.gov.au/policy-library/associated-documents/evaluationframework.pdf.

While guidance and standards are desirable for evaluation processes, this does not imply that all evaluations should be carried out using the same strategy. There is no one correct way to carry out policy evaluations in education systems, they need to draw upon a diversity of methods in order to promote true learning (Sanderson, 2002[83]). Specific evaluation strategies depend on particular policy characteristics, evaluation objectives, resources, and the availability of relevant data and evidence. Evaluations can be purely quantitative, purely qualitative, or take a mixed-methods approach that combines both quantitative and qualitative analysis. A broad range of methods are employed in education policy evaluations that cover experimental and quasi-experimental methods, causal comparative methods, surveys, case studies, interviews, and narrative approaches (Mertens, 2015[84]).

Different evaluation methods may be selected depending on the aims of the evaluation. However, the issue is not as simple as choosing among a suite of tools and techniques. Evaluation should first and foremost be driven by the principles and relevant issues rather than the available methods (OECD, 2007[38]). Designing an effective evaluation also

involves a theoretical analysis of the change process of the reform. This theoretical step is often overlooked, but it can ensure that the right evaluation questions are posed and improve understanding as to why a reform has impacted in a certain way, which informs future policy decisions (White, 2013[85])

It follows that governments need to combine analysis of the policy context with a strategy and portfolio that ensure the most suitable methodological selection for evaluation is made for each reform process. This ability to formulate evaluation questions and choose the best approach is in itself a skill which requires capacity building at the level of the policy maker. Taking these considerations into account shifts the focus from implementing specific evaluation methodologies towards building the capacity to take various approaches at different points, i.e., to harness methodological pluralism to generate the variety of evidence that might be required in different contexts (Shlonsky and Mildon, 2014[86]).

4. Recent policy evaluation practices in OECD countries

This section uses a digest of over 80 key policy evaluation processes which have taken place in recent years across the OECD to structure a comparative analysis of approaches to the education policy evaluation process. Examples presented showcase the diversity of views of the education policy evaluation processes, the methods used, the objectives of policy evaluations and the purposes for which they are used. Analysis is organised according to the key facets of policy evaluation processes which were identified in the framework proposed in Section 2.

4.1. Why?

Policy evaluations tend to be conducted for two main purposes: gather evidence that can be used formatively to improve the policy, and, to assess the impact of the policy in a summative manner (Scriven, 1991[87]). The line between formative and summative evaluation is often not very clearly drawn, and both formative and summative approaches are often taken to the same reform according to evidence from reported OECD policy evaluations.

As pilots and experiments have become more commonplace, there has been an increasing onus placed on evaluators to develop more formative approaches to evaluating policy implementation and take on the role of policy "change agent" (Martin and Sanderson, 1999[88]). Evaluation with a formative intention could include gathering information on the experiences of stakeholders affected by the reform, with a view to making changes or improvements to the reform. Ideally, evaluation for formative purposes acts as a virtuous evaluation-feedback-action mechanism that promotes the policy improvement process (Scheerens, Glas and Thomas, 2007[89]).

In contrast, while formative evaluation might be carried out "to improve" during the development and rollout process, the purpose of summative evaluation is "to proof" (Van Den Akker, Bannan and Kelly, 2013[90]), i.e., to critically examine whether the intervention has been effective and to gather evidence to support its continuation, modification or termination. In OECD countries, summative exercises tend to serve as foundation for decision making at different levels of the education system (OECD, 2013[18]). Impact evaluations can be carried out as an independent exercise, or may also consist of gathering and synthesising the evidence from one or more evaluations carried out throughout the rollout of a reform.

Other more comprehensive categorisations of evaluation purposes also exist. For example, Chen combines two types of evaluation functions (improvement and assessment) with two programme stages (process and outcomes), to arrive at four evaluation purposes: constructive process evaluation, conclusive process evaluation, constructive outcomes evaluation and conclusive outcome evaluation (Chen, 2015[91]). In reality, overlaps in these purposes may also occur, for example, a constructive outcomes assessment may also be intended to be used for process improvements.

4.2. When?

This section examines the question of when evaluation tends to take place within the education policy cycle. Policies can be evaluated before, during or after implementation. Recent reported policy evaluations in OECD countries and economies show that the timing

of the evaluations within the reform process varies greatly, as does the period of time between implementation and evaluation.

The decision on when to evaluate can be related to the intended purpose of the evaluation, and the timeframe for the availability of relevant evidence. Many policy evaluations are backward looking only and may take place a significant amount of time after implementation of the policy. This may be necessary to await data on outcomes from the target group for the policy implementation (for example, labour market activation data from participants in a new higher education initiative). Other evaluations focus only on implementation perspectives of the reform, and may take place during implementation or shortly after the policy is embedded.

Different labels are adopted to describe the "when" of evaluation. This paper refers to *ex-ante (evaluation before)*, interim (evaluation during) and ex-post (evaluation after) implementation when discussing different policy evaluation stages.

4.2.1. *Ex-ante evaluation*

When taking the view of evaluation as an assessment of value, the evaluation of policies can begin even before a reform is fully designed. *Ex-ante* evaluation can involve using evaluation as a tool for examining alternative policies and programmes (Sims, Dobbs and Hand, 2002[92]), or to assess whether a chosen programme is correctly designed to meet desired objectives. It can be thought of as the "what if" analysis which occurs before implementation (Bourguignon and Ferreira, 2003[93]). Robust early intervention in the policy process that challenges the assumptions made by policy makers *ex-ante* could help to strengthen the policy and put it on a more positive trajectory to success (Janssens and de Wolf, 2009[94]).

Analysis of a set of recent evaluations reported by OECD countries and economies indicate that it is less common for policy evaluations to be conducted *ex-ante* (i.e. in order to evaluate or estimate the likely impact of a reform before implementation). Instead, evaluations aim to track implementation as it takes place or provide a review of the implementation process and impact after the policy has been fully implemented. Among over 80 policy evaluations reviewed for this analysis, only one was specifically labelled as an evaluation of different policy options identified during the policy design process which took place before implementation of the policy.

At the same time, many policy documents related to the reforms in the Education Policy Outlook reforms database reference a specific period of more informal weighing up of policy options before implementation, as part of the policy design and development process (OECD, 2018[7]). Although it appears that *ex-ante* evaluation of education reforms may take place less formally during the policy development process, there is a case to be made for more institutionalised formal evaluation as an overview of the potential interaction between the policy and the environment before implementation. Many countries face challenges in understanding why there is often not a strong relationship between the level of investment in systemic reforms and how well they are able to achieve their objectives. Part of the difficulty arises from the lack of understanding of, or preparedness for, how the policy is likely to interact with others implemented by another organisation or government department. Through strengthening evaluation of policies *prior to implementation,* a theory of change can be developed which takes into account all available evidence, including the efficacy of related policies in the current and other contexts and system dynamics and interactions between this policy and other policies, to arrive at a solution which is "plausible, doable, testable and meaningful" (Connell and Klem, 2000[95]).

There are some indications of moves towards more robust evaluation *ex-ante*, both for choosing amongst alternative policies and validating the "theory of change" of chosen reforms (Wolpin, 2007[96]). Mandated *ex-ante* evaluations are now required by many funding bodies in advance of providing funding or support for education projects (European Commission, 2014[97]; UNESCO, 2007[98]). Despite this expansion in recent years, evidence suggests that *ex-ante* policy evaluation practices often do not adhere to the same evidentiary standards as evaluations that take place ex-post. For example, it has been argued that *ex-ante* policy evaluations are strongly shaped by the prevailing political context and less likely to lead to learning which may challenge the status quo or assumptions made in the policy process or the problem definition (Hertin et al., 2009[99]).

4.2.2. *Interim evaluation*

A review of the evaluations contained in the Education Policy Outlook policy evaluation digest shows that the interim evaluations carried out tended to focus on monitoring the implementation process, rather than being targeted at tweaking and improving the policy in "real time" as it was being rolled out. However, policies that are being evaluated periodically as they are rolled out they are also providing diagnostic information which can be used to inform future provision or indicate when a reform is not performing as intended. For example:

- The monitoring of the implementation of the **expansion of childcare places in Germany** involves the publication of regular evaluation reports which monitor the attendance rates and identify regional differences in both demand and access. These feed into the continued rollout process.
- The formative evaluation of the **Pasifika Education plan 2013-2017 in New Zealand** consists of monitoring the key targets defined by the policy on an annual basis in order to ascertain whether the policy is on track, while also taking a more in-depth case study approach to selected specific programmes introduced as part of the plan (Ministry of Education, 2014[100]).

Interim evaluations often examine how relevant groups at the school level are dealing with the reform and issues arising. This can take the form of informal consultations and interviews or processes to gather information on policy implementation from schools or classrooms in a structured manner to facilitate understanding on how reforms may impact school or classroom practices. However, practices for gathering information on the implementation and perception of reforms in schools can vary based on the target for the reform and the key actors involved. For example:

- The **Technological Plan for Education (*Plano Tecnológico da Educação*, PTE) in Portugal** aimed to modernise schools' technological infrastructure and improve training and the use of ICT among students and teaching staff. To evaluate the reform, students and relevant school "adults" were surveyed on their attitudes, proficiency and usage patterns of ICT, as well as their perceptions and knowledge of the PTE. Given the nascent nature of the plan and the stage of the process the evaluation was being conducted at, the evaluators also considered it important to listen in depth to the fears, expectations and perceptions of the plan, which led to a number of semi-structured, in-depth interviews being undertaken with relevant stakeholders (CIES, 2011[101]).
- A 2009 evaluation of a new set of **National Professional Standards for Teachers in Australia** focused on measuring attitudes to new standards in the education

system using evidence gathered from case studies, a national survey and a forum for stakeholders and contributors (AITSL, 2016[102]).

- An interim review of the national **Strategy to Promote Literacy and Numeracy in Ireland (2012-2020)** evaluated progress towards meeting strategy objectives through an examination of standardised test results, as well as by surveying every organisation and departmental section named in the plan, holding a consultative forum, and meeting with groups of principals, teachers and students (Department of Education and Skills, 2017[103]).

4.2.3. *Ex-post evaluation*

Ex-post evaluation takes place after the policy has been fully implemented and is a backward looking, often more holistic view of the policy's impact. Ex-post evaluations are important for ascertaining potential impacts of polices.. In the case of one-off measures or policy experiments, ex-post evaluations are a way of providing knowledge and insight into what worked, which can be applied to future policies or as evidence to support a decision on whether to expand the policy.

Box 4.1. Case study: Assessing the impact of the introduction of Individual Student Plans in Denmark.

Individual Mandatory Student Plans (*Elevplaner*) were introduced in Danish primary and lower secondary schools in 2006 in order to monitor student progress against the common objectives, which are the learning objectives set for all students in compulsory education. They include a summary of students' test and evaluation results and can provide the basis for discussions between students, teachers and parents.

In 2008, the Danish Evaluation Institute carried out an evaluation to assess the impact of the rollout of the new student plans and underpin the debate around the introduction of mandatory student plans with empirical evidence. The investigation took a comprehensive approach, seeking to assess the impact from the perspectives of municipalities, school leaders, teachers, students and parents. Evidence to inform the evaluation was gathered using three key sources:

- Case studies at six schools, which included detailed interviews with municipalities, school management, teachers and parents in the school.
- A questionnaire survey covering all municipalities in the country.
- Representative surveys of teachers and parents.

To measure the factors that contributed to impacts, evaluators set up a number of hypotheses, primarily based on the results of an initial feasibility study of six schools carried out in 2007, which described possible factors affecting the assessment of the student plan utility. To ensure a variety of perspectives, the six schools ranged from those that had recently adopted student plans to those that had been using similar tools for a longer period of time. The hypotheses were proposed and put forward before the surveys were conducted, and then tested using statistical regression models based on the survey questionnaire data collected.

The evaluation found an overall positive view among teachers and students about the introduction of the student plans. At the time of the evaluation the introduction of the plans were found to have been well underway in schools, but they had not yet contributed to strengthening teaching differentiation. A key issue identified was around designing the plans to meet the diverse needs and expectations of parents, students and teachers within one document.

Source: EVA (2008[104]), Arbejdet med elevplaner: en national undersøgelse af erfaringer (Working with student plans: a national study of experiences), https://www.eva.dk/grundskole/arbejdet-elevplaner (accessed on 31 May 2018).

4.2.4. *Multiple evaluations of the same policy*

Analysis of recent evaluations shows that key policies are often evaluated in multiple phases or stages, or are assessed multiple times from diverse perspectives by different evaluators. For example:

- the *Folkeskole* reform, implemented in **Denmark** in 2014, has three major objectives for compulsory public schools in Denmark: to support all students to reach their fullest potential, to reduce the importance of social background for academic results, and to strengthen the trust in the Folkeskole and student well-

being. This reform been the object of evaluations at different points in time. Its latest evaluation was published in 2020.

- the third generation of **Portugal's** Education Territories of Priority Intervention Programme (*Territórios Educativos de Intervenção Prioritária*, TEIP) to address educational disadvantage has been evaluated twice by the Ministry of Education and also been the subject of a more in-depth study by the CIES (Centre for Research and Studies in Sociology of the University Institute of Lisbon).
- **Ireland's** educational disadvantage initiative (Delivering Equality of Opportunity in Schools, DEIS) has been evaluated a number of times by the Education Research Centre funded by the Department of Education and Skills and separately by the Economic and Social Research Institute, an independent policy research institute.

Multiple evaluation processes can provide large volumes of formative information throughout the implementation of the policy, which can in turn can provide a number of perspectives from which to evaluate the impact of the policy ex-post.

Box 4.2. Case study: Evaluation of the Canada Student Loans Programme

Evidence gathered during interim evaluations can enhance the ability of evaluators to assess the impacts of a reform after implementation by providing insights into which mechanisms might have contributed to the reform results. The Canada Student Loan's Programme (CSLP) was evaluated by the Employment and Social Development Canada (ESDC) Evaluation Directorate in 2011 in order to determine the validity of the programme's rationale, needs assessment and its success at promoting access to post-secondary education.

This impact evaluation was the culmination of a series of smaller studies that took place between 2006 and 2010. The final report provides a summary of 32 studies that were undertaken over the five-year period, and also takes external reports into account. The smaller evaluation reports covered evidence including literature reviews, international comparisons, surveys, focus groups, key informant interviews, and administrative data. The evaluative activities were organised around a set of evaluation questions that covered programme rationale, programme objective and achievement, impacts and effects, cost-effectiveness, and programme delivery issues and communications.

The programme was considered from multiple different evaluative perspectives, for example, debt repayment prevalence after graduation, how processes were quality assessed, and how the policy promoted equity of access. A series of evaluation questions were used to inform the evaluation process. In addition to synthesising the empirical evidence on the efficacy of the programme, the evaluators also considered whether the programme rationale developed ex-ante was still valid.

Management provided responses to the key recommendations already in the evaluation report. This can be useful in terms of "closing the policy cycle" and showing how the results of the evaluation can be taken into account. In some cases, management further analysed the policy options suggested in the recommendations (for example to associate the amount of the loan with local living costs, which vary dramatically across the country) and were able to respond with further perspectives within the report (e.g. it is accepted that living costs are different, however the complexity and cost of implementing such a programme would likely exceed the financial benefits.)

Source: Human Resources and Skills Development Canada (2011[105]), Summative Evaluation of the Canada Student Loans Program, http://publications.gc.ca/collections/collection_2012/rhdcc-hrsdc/HS28-44-1-2011-eng.pdf (accessed on 31 May 2018).

4.3. Who?

4.3.1. *A taxonomy of evaluators*

There are a range of bodies, institutions and professionals involved in policy evaluation across OECD education systems. This includes specialist education evaluation agencies, academic researchers, private institutes, commercial institutions and international

organisations. Based on the analysis of the Education Policy Outlook digest of policy evaluations, as well as other OECD evidence, evaluators can be loosely classified according to the categories outlined in Table 4.1. Further analysis and some examples of practices for a selection of the categories follow.

Table 4.1. Common evaluator types in OECD countries and economies

Type of evaluator	Key characteristics
Specific institution for educational evaluation	An institute undertaking education evaluation activities, including evaluation of education policies. .
Higher education institution or academically based	Evaluators based in academic institutions, which may conduct an evaluation on request from an education authority.
Ministry or central/state education authority	Evaluation units or personnel based in the relevant central or state education authority.
Bodies with broader research or evaluation responsibilities (not specific to education)	Often run their own research projects or performance audits, not all specific to education. May also conduct evaluations if contracted/requested by an education authority.
External consultant evaluators (engaged by central/state authority or ministry)	Private contractors which specialise in evaluation, engaged by education authority.
Collaboration and committee	Often led by a leading expert in a relevant area, who may then select a further panel of experts to conduct the review. A group or consortium may be centrally selected to be representative of all major stakeholders in the policy reform.
International organisations	May evaluate a policy as part of regular peer review processes, on invitation from a country, or as a condition of programme funding.

Ministry or central/state education authority

Many of the policy evaluations reviewed for this analysis were overseen or steered directly by the ministry of education or another central authority with responsibility for education policy. The arrangements for evaluating policies centrally varies across countries. In Sweden, for example, the Analysis and International Affairs unit of the Department of Education and Research may evaluate programmes and policies, while Slovenia, the central education inspectorate evaluates education programmes, and in Chile the Ministry for Social Development evaluates the social impact of educational programmes (OECD, 2017[12]).

Many education authorities contain internal evaluation units which may routinely review policies. This can create a tension if policy evaluations are carried out within the same environment initially responsible for designing and implementing the policy, but can also ensure that results of evaluations feed more directly into future policy research. Analysis of the Education Policy Outlook evaluations digest shows that education authorities are also responsible for commissioning a wide range of evaluative research from other bodies, which can help to ensure the independence of the evaluation process and raise public trust.

Specific institutes for educational evaluation, and broader research institutes

Evaluators should be independent, credible, transparent, and have the ability to access the data and evidence they need to perform a robust evaluation. Independent evaluation

institutions have a critical role to play in cutting through some of the political and resource related challenges that evaluation faces (Rutter, 2012[55]).

In many OECD countries, there are independent institutions **specifically responsible for educational evaluation and research**, which may include evaluating education policies. Some of these institutes are new creations, while others have been established for decades, and there is a variety of funding, governance structures and research foci in place across the institutes. Table 4.2 shows a selection of institutes for education evaluation and research across OECD countries.

In other cases reforms are evaluated by **institutions that may also evaluate policy in other policy areas**, or conduct policy research. These institutions also tend to be independent in nature, although the evaluations may be instigated initially by the central education authority.

Some independent evaluation institutions have the ability to choose the policies they would like to evaluate, free of any political pressure. The Dutch Central Planning Bureau is a key clearinghouse of public policy options in the Netherlands. Although publicly financed, it is completely independent and free to make its own policy analyses and recommendations. The Bureau is widely considered as the authoritative source on policy evaluation within the Netherlands. It carries out evaluations of policies from across the economic and social spectrum, including evaluations of the efficacy of education policies. Most policy options and policy proposals are first evaluated by the Central Planning Bureau for likely efficacy before proceeding to implementation stage. The Bureau is particularly noted for providing independent evaluations of the electoral manifestos of political candidates, thus helping to inform public opinion and ensure that would be governments are more measured in devising policies.

In other cases, evaluation institutions may be more closely guided by the priorities of the ministry or department of education if they have engaged the institute on a contract basis. For example, The Nordic Institute for Studies in Innovation, Research and Education (NIFU) in Norway conducts evaluations on education policies, as well as reforms in other related policy areas, such as innovation; This institution is funded through research contracts with both public and private sector clients.

Table 4.2. Selected OECD education evaluation and research institutes

Country	Name of institute	Key characteristics
Australia	Australian Council for Education Research (ACER)	The Australian Council for Educational Research (ACER) is an independent, not-for-profit research organisation established in Australia as a company limited by guarantee, whose mission is to create and promote research-based knowledge, products and services that can be used to improve learning across the lifespan. ACER undertakes commissioned research and development and develops and distributes a wide range of products and services.
France	*Conseil Nationale de l'évaluation du système scolaire* (CNESCO)	CNESCO, created by the Law of Orientation and Programming for the Rebuilding of the School of the Republic in 2013, is one of the few institutions in charge of independent evaluation. It aims to enlighten both the actors of the school world (pupils, parents of pupils, professionals of the national education system, local authorities, associations of popular education, etc.) and the general public. CNESCO also ensures the dissemination of the results of evaluations and research.
Korea	Korean Educational Development Institute (KEDI)	The Korean Educational Development Institute (KEDI) has been a leading institution in educational policy development and implementation since it was founded in 1972. KEDI plays a pivotal role as a national think tank in setting the national agenda of Korean education and provides guidelines for innovating the educational system to enhance educational quality. KEDI strengthens global leadership through joint research and international ties and seeks a new educational paradigm to meet the needs of the upcoming fourth industrial revolution.
Mexico	The National Institute for Educational Assessment and Evaluation (INEE)	The National Institute for Educational Assessment and Evaluation (INEE) shared responsibility with the Secretariat of Public Education for evaluation of the education system, and was responsible for evaluating the quality of the national education system from pre-school to upper secondary education. The institute was initially created in 2002 and became an autonomous public body in 2012 (though it ceased operation in 2019).
New Zealand	New Zealand Council for Educational Research (NZCER)	NZCER is Aotearoa New Zealand's independent, statutory education research and development organisation, established in 1934. The NZCER Act 1972 requires the organisation to carry out and disseminate education research and provide information and advice. NZCER conducts research and evaluation work with a range of public and private sector clients, and also produces research-based products such as tests, journals, books and services such as online testing, surveys, test marking and analysis.
United States	National Center for Educational Evaluation and Regional Assistance (NCEE)	The NCEE is responsible for large-scale impact evaluations of education programmes supported by federal funds. It also funds research into improving technical assistance; and supports the development and use of educational research and evaluation throughout the United States.

Collaboration and committee approaches

Many recent OECD policies were evaluated through collaborative efforts, such as expert groups, committees, taskforces or communal efforts by different evaluators. Taking a collaborative approach to evaluation is a well-recognised way of building capacity for evaluation and promoting learning and knowledge transfer at the level of the school and classroom. A committee approach to evaluation can also serve to build capacity among policy makers and ensure that a range of expertise and stakeholder perspectives are reflected in the evaluation process.

Box 4.3. Case study: The Excellence Initiative in Germany

The Excellence Initiative (2005) aims to systemically support higher education and top-level research by awarding additional funding to top-performing universities. On behalf of the federal government and the *Länder*, the Joint Science Conference (*Gemeinsame Wissenschaftskonferenz* – GWK) proposed to appoint an international committee to evaluate the policy. In 2014, the GWK adopted a mandate for the "International Commission of Experts on the Evaluation of the Excellence Initiative" (hereinafter referred to as IEKE) to develop a comprehensive, primarily qualitative, assessment of the Excellence Initiative as a strategic programme and its effects on the German science system. Dieter Imboden was selected as the Chairman of the IEKE and proposed the selection of a further nine members.

The IEKE was supported in its work by an independent office that was selected in an open tender procedure. The office was responsible for the entire organisation of IEKE's work and supported it with work on the content. The office also organised the commission's meetings and took care of the budget. On behalf of the IEKE, the office carried out analysis on specific issues and questions concerning German and international research systems. The office evaluated numerous information sources and reports regarding questions to be answered by the IEKE.

To clarify its mandate, the GWK formulated key questions that covered the effects of the Excellence Initiative on universities not funded as part of the programme. The IEKE was given full flexibility on how to design their work, but was asked to include the end of June 2015 published data protected report by the *Deutschen Forschungsgemeinschaft* (DFG) and the *Wissenschaftsrats* (WR) in their analysis. The IEKE met six times for mainly two-day meetings. During the first meeting the process was discussed and it was decided to conduct interviews as part a first phase of work. The findings of the interviews, the results of the DFG/WR report and other publications, were the basis for the analysis that took place during the second working phase. The IEKE formulated a list of key questions for the conduction of interviews. Overall, more than 100 people were interviewed. Interviewed personnel worked at different hierarchical levels (including students, PhD candidates, Postdocs, professors and university presidents) at different German universities, including universities that did not participate in the Excellence Imitative. In addition, discussions were held with representatives of non-university research institutions and foreign universities. In most cases, two IEKE members were present at each interview, with all members present on a few occasions.

The IEKE was aware from the outset that neither the DFG/WR's quantitative analysis, nor any other research, would make it possible to make stringent statistical statements about the relationship between the Excellence Initiative and any observed changes in quantitative parameters used to characterise university research and the perception of German universities at home and abroad (publications, citations, university rankings, etc.). On the one hand, the observed period since the beginning of the Excellence Initiative is still too short to make any consequences of the Excellence Initiative fully visible. On the other hand, a large number of other national and international programmes and changes are simultaneously influencing the German university and research system, meaning that a clear link between a specific measure (such as the Excellence Initiative) and an observed change is not possible.

Source: IEKE (2016[106]), Internationale Expertenkommission zur Evaluation der Exzellenzinitiative Endbericht (International Commission of Experts for the Evaluation of the Excellence Initiative Final Report), https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/Imboden-Bericht-2016.pdf (accessed on 31 May 2018).

International organisations

An important type of external evaluator in education systems are international organisations such as the OECD, the European Union, the International Monetary Fund, the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the World Bank. International organisations develop and publish wide ranging suites of indicators which give comparative indications of the system level performance of education across countries, although this macro-level approach is often not able to take into account specific contexts at the country or region level that may influence education and training outcomes (Neves, 2008[107]). In their role of evaluators, international organisations may also take the position of peer reviewers (such as the OECD) for specific education reforms, undertake systemic evaluations, or act as monitors when providing aid or financing for education programmes (such as the World Bank).

Box 4.4. Case study: The World Bank evaluation of the Secondary Education Project in Turkey

The Secondary Education Project in Turkey with the World Bank (2006-11) aimed to improve the quality, economic relevance and equity of secondary education and develop lifelong learning. Prior to implementation, the World Bank policy appraisal document (PAD) identified a number of education challenges that Turkey faced that created the imperative to improve secondary education, including low educational attainment, an increasing working age population and income inequality. A number of project development objectives (PDOs) were agreed with a set of associated indicators to form the basis for measuring progress towards achievement of the objectives. The project included creating and rolling out a new curriculum, student assessment instruments and other associated supports, with the aim of improving performance and equity.

The evaluation operated in line with the standard World Bank evaluation framework (World Bank, 1997). It consisted of: identifying objectives, monitoring the change of objectives, following implementation and identifying key strengths and weaknesses of the implementation process, assessing the impact of the programme according to the indicators defined, and rating various aspects of the process according to a set of standard categories predefined by the operations evaluation department (for example, satisfactory or unsatisfactory process, high, moderate or low impact etc.). The ratings were entered into a database to facilitate the wider systemic evaluation process of the World Bank's work.

According to the "Implementation, Completion and Results Report", the project partially achieved the following objectives: revision and implementation of general and vocational curricula, public availability of student achievement results, distribution of materials for teachers, improvement of vocational teachers' skills, introduction of an online career information system, training of school management teams on school development plans, and grant distribution to schools in low enrolment areas.

The report also identified the key strengths of the project which promoted success (including complementarity with an existing European Union programme, good communication with the ministry and beneficiaries of progress, and strong analytical underpinnings). It also identified some of the shortcomings that initially prohibited success (such as limited initial ownership, an overly ambitious PDO, not incorporating lessons from other programmes into the programme design, and not clearly identifying and mitigating against key risks).

Source: Le, P. et al. (2012[108]), Implementation Completion and Results Report (IBRD-47670), http://documents.worldbank.org/curated/en/874001468110647154/pdf/NonAsciiFileName0.pdf (accessed on 31 May 2018).

4.4. What?

OECD education systems measure education reform success in diverse ways during the evaluation process. This section analyses the types of measures and metrics used in recent policy evaluations in OECD countries and economies to ascertain where a reform lies on the success/failure spectrum. The overarching goal of policy evaluation is to make some judgement about the quality and value of a policy. To achieve this, evaluators must take some measures of "effects" of the policy.

Evaluation design is a critical component of the evaluative process, and evaluations should be designed from start to finish with consideration for how the evaluation will be used

(Ramirez et al., 2013[109]). Effectively measuring the impact of a policy begins with defining clear targets at the outset of policy development. This can present particular difficulties with education policies, as targets must be appropriate, measurable, effective and relevant at the system level, as well as translated into objectives at the classroom level. A tension can therefore exist for evaluators between identifying the measures which provide the best evidentiary standards, and the feasibility and cost of getting the information necessary to use these measures.

Resource constraints must also be balanced against the value of the information gained for improving policy. To aid in this balance, existing data should be sought and exploited as much as possible. Evaluation measures should be chosen with consideration of how much relevant data are already available to evaluate the policy, with resources focused on developing measures in those areas where there is the greatest uncertainty or lack of knowledge, or where evaluation is most crucial (Weitzman and Silver, 2013[110]).

The Education Policy Outlook evaluations digest shows that evaluations tend to target measures of output and outcomes, or collect information to measure changes in perceptions, processes and practices. Examples of evaluations focusing on outputs and outcomes include:

- The impact of the Drive to Reduce Dropout Rates in the **Netherlands** is measured by the consistent collection of comparative information on student dropout numbers, as well as the reasons for students leaving school early. This is used to increase the performance of the policy by monitoring progress and targets and allowing comparisons between schools and regions, as well as through providing financial incentives to schools to reduce their dropout rates (Panteia and SEOR, 2016[111]).
- The evaluation of the National Partnership on Youth Attainment and Transitions in **Australia** examined changes in participation and attainment rates of students, their transition outcomes defined within the National Partnership, as well as more detailed indicators across sectors (Dandalo Partners, 2014[112]).

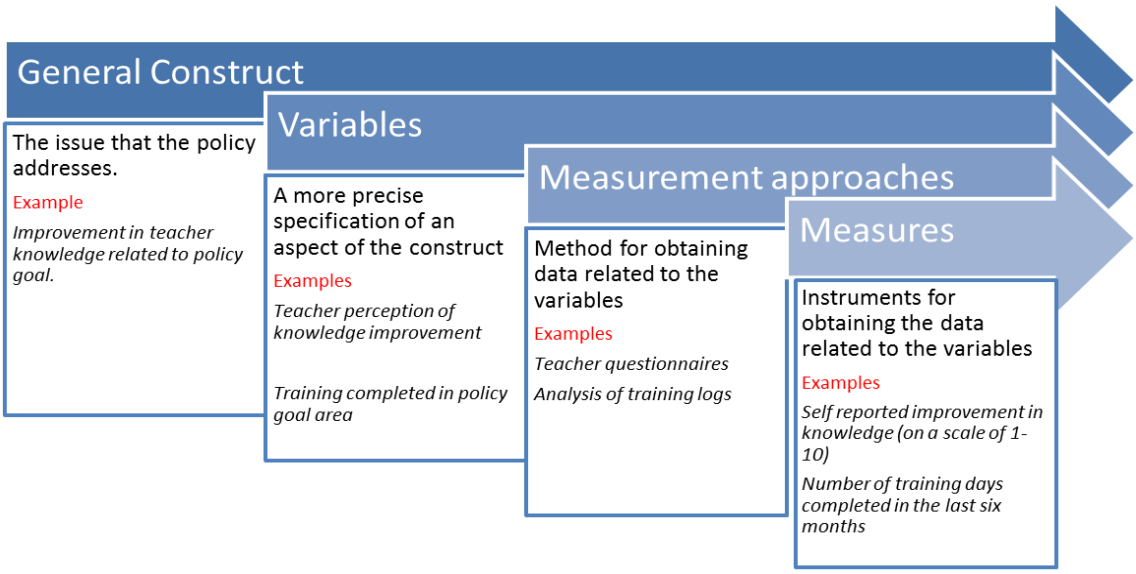
Examples of evaluations focusing on perceptions, processes and practices (such as institutional practice, systemic knowledge or capacity) include:

- **Luxembourg** evaluated its Compulsory Education Reform by capturing how the implementation of the reform was received in schools in order to identify issues arising. Measures captured in surveys and interviews included the levels of satisfaction with and views on new school organisation structures. Issues were also identified and crystallised from the transcripts of detailed free form interviews (Koenig, 2013[113]).
- The **Norwegian** Assessment for Learning programme carried out semi-structured interviews with a range of stakeholders. The interview approach was informed by a range of official and non-official letters, documents and memos from the Norwegian Ministry of Education and Research. These documents mainly served to inform the interviews that were carried out, as well as to clarify to external evaluators how the implementation process was conducted from the central level (Hopfenbeck et al., 2013[114]).

In order for the evaluation to be judged of high quality, and for its conclusions to be well accepted by stakeholders, evaluators must strive to ensure that the measurement decisions made can allow the strongest conclusion possible about the success and impact of the

reform. Quantitative outcome measurements for use in policy evaluation are arguably more accessible. There are already many national and international measures of student outputs and outcomes which can be used as part of a policy evaluation exercise, although they may not be the most suitable measures in all cases. Qualitative measures related to the policy can be more onerous to construct. Figure 4.1 outlines key steps to consider when moving from the theoretical construct to measurement instruments.

Figure 4.1 From construct to measurement instrument (Braverman, 2013[115])



Source: Adapted from (Braverman, 2013[115]) "Negotiating Measurement: Methodological and Interpersonal Considerations in the Choice and Interpretation of Instruments", *American Journal of Evaluation,* pp. 99-114, http://dx.doi.org/10.1177/1098214012460565.

Table 4.3 presents a "menu" of possible qualitative outcomes, adapted from the work of Reisman, Gienapp and Stachowiak (Reisman, Gienapp and Stachowiak, 2007[116]) (2007[116]), that are relevant to education policy evaluation. Further consideration of how evidence can be gathered to gain insight into these outcomes is discussed in the next section.

Table 4.3. Qualitative outcomes menu for education policy evaluation

Relevant general construct	Example variables
Shifts in **social norms** within education providers and institutions	Changes in beliefs, behaviour, values and attitudes towards the policy topic, increased awareness of the topic or increased common understanding of the topic.
Shifts in **organisational capacity** in schools or at lower levels of education governance	Improved organisational and strategic abilities of staff, greater stability in the organisation, more effective communication channels.
Strengthened **alliances** supporting learning	Increased collaboration and knowledge sharing; increased collective support; new or stronger partnerships and alliances, for example, with social partners.
Strengthened **public support** for the policy	Increased parental involvement in education issues, increased public knowledge and awareness, positive media coverage, change in public perception.
Improved **education policy processes**	More transparent policy development processes, stronger and more effective implementation on the ground and policy resilience in the face of challenges.
Changes in **student impact**	Improved educational conditions for students, improved well-being and student empowerment.

Source: Adapted to the education sector from: Reisman, J., A. Gienapp and S. Stachowiak (2007[116]), Measuring Advocacy and Policy, The Annie E. Casey Foundation, Baltimore, https://www.alnap.org/system/files/content/resource/files/main/aecf-aguidetomeasuringpolicyandadvocacy-2007.pdf (accessed on 25 April 2018).

Certain policies can also target improving classroom practice or interactions between students and teachers. Many policy reforms seek to make changes to the organisation of schools, or to impact directly on the knowledge transfer process between teachers and students. Unlike measures for assessing student outcomes and skills, such as standardised tests where the measurement instruments across education systems are aligned, well-accepted measures of classroom and school practices that can provide transferable insight do not yet exist.

At the classroom level, some tools aim to produce measures of teaching practices and student-teacher interactions in a standard way. Examples include standardised classroom observation measures (RTI International, 2016[117]) and structured classroom vignettes (Stecher et al., 2006[118]). The OECD, through the Teaching and Learning International Survey (TALIS), is also piloting methods to capture teaching practices in an internationally comparable manner through a video study, and intends to build a global video library to showcase and disseminate teaching practices (OECD, 2017[119]). The development of such measures, while still at early stages, could in the future feed into the evaluative process.

Box 4.5. Case study: Perceptions and practices in summative evaluation of the New Horizon Programme in Israel

When undertaking summative evaluations, rather than directly measuring student outcomes, evaluators may focus on the changes that the policy has made in perceptions, processes and practices, following a theory that these changes will, if positive, lead to improvements in student outcomes.

The New Horizon Programme (*Ofek Hadash*) in Israel is a national programme which began in 2007 to advance education in Israel in elementary and junior high schools. The reform included four main complementary targets: 1) boosting the status of teachers and raising their salaries; 2) providing equal opportunities to every student and raising student achievements, including through the provision of individual hours for teaching small groups of students; 3) improving the school climate and environment; and 4) empowering and expanding the authority of the school principal.

An evaluation of the programme conducted in 2010 by RAMA (the National Authority for Measurement and Evaluation) focused on understanding the processes and changes that had taken place in schools participating in New Horizon over a long period of time. The intention was to learn what changes the reform had delivered within the school environment, as well as the perceived impact on student learning and achievement through the perspective of those involved in implementing the reform. The approach taken to gather information was primarily through telephone interviews among broad representative samples of teachers and principals, while a small number of face-to-face, more in-depth interviews were also conducted.

The evaluation found that the programme is well implemented in schools and has wide acceptance among teachers and principals, and that the individual hours with students are perceived as most effective for fostering student improvement. At the same time, teachers reported feeling overworked, and teachers and principals are still reporting inadequate physical conditions and a lack of autonomy.

Source: RAMA and Ministry of Education (2010[120]), Evaluation of the New Horizon reform in elementary and junior high education at the end of three years of implementation: Summary of Findings, http://cms.education.gov.il/NR/rdonlyres/45CCC357-4A00-42A6-9E8B-F892A24B202E/163977/OfekChadashsummary.docx.

4.5. How?

In OECD countries, a variety of quantitative and qualitative methodologies are employed in evaluations, including experimental and simulation approaches, such as game theory, behavioural insights, surveys, mixed models and longitudinal analyses of the population targeted by reforms. Unlike other types of evaluation, such as at the school or student level, policy evaluation in education does not appear to have any harmonisation of practice across education systems, as can be seen from the diversity of instruments employed by the evaluations analysed during this review.

There is a vast body of literature covering the application of different evaluation methods to policy evaluation. Table 4.3 summarises Stufflebeam's critical review of evaluation models, identifying 22 distinct evaluation approaches, including two approaches deemed "illegitimate" for attributing value to a particular programme (Stufflebeam, 2001[33]).

Table 4.4. Stufflebeam's methodological typology for education policy evaluation

Evaluation method	Description
Objectives-based studies	Specifying operational objectives and analysing pertinent information to find out how well each objective was achieved. Applicable for tightly focused projects that have clear, supportable objectives.
Accountability/payment by results studies	Typically narrows the evaluative inquiry to questions about outcomes, obtaining an external, impartial perspective compared to the internal perspective often preferred in objective-based studies.
Objective-testing programmes	Using student test results to evaluate the quality of projects, programmes, schools, and even individual educators. Infers that high scores reflect successful efforts and low scores reflect poor efforts.
Outcome evaluation as value-added assessment	Systematic, recurrent outcome/value-added assessment, coupled with gain score analysis, is a special case of the use of standardised testing to evaluate the effects of programmes and policies.
Performance testing	Requires students to demonstrate their achievements by responding to evaluation tasks, such as tests, presentations, portfolios of work products, or group solutions to defined problems.
Experimental studies	Beneficiaries are randomly assigned to experimental and control groups and outcomes are contrasted after the experimental group receive an intervention and the control group do not.
Management information systems	Supplies managers with information needed to conduct and report on the policy. Most management information systems include objectives, specified activities, projected milestones, and a budget.
Benefit-cost analysis approach	A set of largely quantitative procedures used to understand the full costs of a programme and to judge what these investments returned in objectives achieved, as well as broader social benefits.
Clarification hearing	Essentially puts a programme on trial. Role-playing evaluators competitively implement both a damning prosecution of the programme and a defence of the programme before an adjudicator.
Case study evaluations	A holistic, multi-level analysis of a policy in its geographic, cultural, organisational, and historical context, closely examining how it uses inputs and processes to produce outcomes.
Criticism and connoisseurship	Assumes experts in a given substantive area are capable of unique in-depth analysis and evaluation. Methodology includes critics' systematic use of their experiences, insights and abilities.
Programme theory-based evaluation	Begins with either 1) a well-developed and validated theory of how reforms of similar type and settings operate to produce outcomes; or 2) an initial stage to approximate such a theory within the context..
Mixed-methods studies	Employs quantitative or qualitative methods and is preoccupied with using multiple methods rather than whatever methods are needed to comprehensively assess a programme's merit and worth.
Decision/accountability-oriented studies	Emphasises that evaluation should be used proactively for improvement as well as retroactively to judge its merit and worth, to provide a knowledge and value base for making decisions.
Consumer-oriented studies	The evaluator plays the role of the surrogate consumer. The approach regards a consumer's welfare as a programme's primary justification and accords that welfare the same primacy in evaluation.
Accreditation/certification approach	The evaluation's purpose is to determine whether institutions, institutional programmes, and/or personnel should be approved to deliver specified public services and are meeting minimum standards.
Client-centred studies	Embraces local autonomy and helps people who are involved in a programme to evaluate it and use the evaluation for improvement. The evaluator works with diverse clients involved with the policy.
Constructivist evaluation	Rejects the existence of any ultimate reality and employs a subjectivist epistemology. It sees knowledge gained as one or more human constructions, unverifiable, and constantly changing.
Deliberative democratic evaluation	Charges evaluators to uphold democratic principles in reaching conclusions. Proactively promotes the equitable participation of all interested stakeholders throughout the course of the evaluation.
Utilisation focused evaluation	Explicitly geared to ensure that programme evaluations make an impact. All aspects of a utilisation focused evaluation are chosen to help the users apply evaluation findings to their intended uses.

Source: Adapted from Stufflebeam, D. (2001[33]), "Evaluation Models 2", New directions for evaluation 89, https://www.wmich.edu/sites/default/files/attachments/u58/2015/Evaluation_Models.pdf (accessed on 16 March 2018).

The table above shows that the range of possible approaches to policy evaluation is wide, and can be further complicated by the combination or modification of any of the techniques identified. At the same time, policy evaluation methods are often not as rigorous as those used in empirical academic research, and conclusions drawn from evaluation studies can often go beyond what is warranted by the evidence created or examined as part of the evaluation (Lauer, 2004[121]). Given the high stakes of policy decisions and funding which can often rest on the evidence from evaluations, it is important that evaluation methodology is robust, as comprehensive as possible, covers an appropriate time period, and employs a methodology suitable to the reform being evaluated (HM Treasury, 2011[15]). Ensuring that

these standards are met also increases the likelihood that the evaluation process and results are perceived as legitimate by stakeholders.

The specific nature of the policy and policy context can dictate which types of evaluation methods are most appropriate. As education policy makers can introduce a variety of reforms across the system in diverse contexts, this implies that capacity should be available to access a broad range of evaluation methodologies. As discussed in Section 3, although it can be a difficult path for governments to build these wideranging capacities into their system, many countries have been responding to this challenge by enhancing funding, regulation and guidance; harnessing outputs from other types of evaluation and assessment in education systems; and making better use of data sources and other contextual information. These practices help to build a baseline of evidence which can then be used in a multitude of ways that are compatible with the specific evaluation method chosen.

Within the Education Policy Outlook digest of evaluations, most evaluations used either a qualitative approach or combined quantitative and qualitative data in a mixed-methods design. The term "qualitative evaluation" covers a broad range of intents and approaches to the policy evaluation process. It can be characterised by techniques which focus on the processes used to achieve the policy outcome, or which might attempt to get the "interior" perspective of the policy. Qualitative evaluation tends to gather words, images or ideas to feed into the evaluation rather than numerical data (Ritchie et al., 2013[122]). Evidence from the Education Policy Outlook digest of evaluations shows that qualitative and mixed-methods evaluations comprise a variety of designs and techniques, including interviews, observational studies and case studies.

Qualitative methods take a different conceptual approach to experimental or other quantitative methods when evaluating policies. As outlined by Maxwell (2004[123]), while the aim of variable-based research is to study systematically whether there is a causal relationship between defined, observable and comparable inputs or outputs (the "whether" of causality, or the description of the causal effect), qualitative evaluation provides a view of the "how" of causality or the explanation of the causal mechanisms. In evaluation processes, quantitative and qualitative methods can complement each other, with the former generally attempting to estimate causal relationships and the latter generally attempting to explain how and why they occur.

Box 4.6. Qualitative methods in the Early Childhood Education Participation Programme in New Zealand.

The Early Childhood Education (ECE) Participation Programme was set up in 2010 with the aim of increasing participation levels in quality ECE among some groups, largely Māori and Pasifika children, and children from lower socio-economic communities. Specific initiatives included Engaging Priority Families/*whānau* (EPF), Supported Playgroups (SP), Flexible and Responsive Home-based Services (FRHB), Identity, Language, Culture and Community Engagement (ILCCE), the Intensive Community Participation Programme (ICPP) and Targeted Assistance for Participation (TAP).

The initiatives were evaluated in a four stage process by the Wilf Malcolm Institute of Educational Research of the University of Waikato. A multidimensional approach was taken using quantitative and qualitative methods to assess participation and learning outcomes and to examine the transitions to school after the programme.

Stage four of the evaluation consisted of an in-depth qualitative review focused on the EPF initiative. The experiences and outcomes of a cohort of 18 children were studied over a six-month period, over which time evaluators gathered the perspectives of their educators in both the EPF setting and the primary school after transition. Observations within the early childhood education setting were also carried out in order to collect information about learning episodes. Information was analysed on the level and type of educational activities carried out by parents to support the learning of the child. As part of the analysis, the strength of the children's learning foundations across five domains (well-being, contribution, belonging, exploration and communication) was gauged from the perspective and observation activities, and were mapped against the ECE curricular strands (*Te Whariki*).

Through the qualitative analysis, evaluators found that some factors and circumstances were more relevant than others when developing strong learning foundations and transitioning through school. For example, it was found that the EPF co-ordinator role was powerful for engaging parents and ensuring regular attendance in the programme, as was the pedagogical approach taken. Educational activities in the home setting with parents were also associated with developing stronger learning foundations, especially where there was continuity between home and school. On the other hand, learning foundations that were not as strong were associated with less time enrolled in ECE, and more transient family settings with less collaboration between home, school and the pre-school setting were considered a contributing factor to more complex and unpredictable transition outcomes for the children.

Source: Mitchell, L. et al. (2016[124]), "ECE Participation Programme Evaluation Stage 4 Report to the Ministry of Education", https://www.educationcounts.govt.nz/__data/assets/pdf_file/0005/171851/ECE-Participation-Programme-Evaluation-Stage-4.pdf (accessed on 31 May 2018).

4.5.1. *Focus on quantitative methodologies*

Quantitative methodologies in evaluation can be categorised into three broad categories: 1) the analysis of statistical information and data to describe trends relevant to the evaluation; 2) the use of econometric or statistical models to gain empirical knowledge about the impact of a policy; and 3) the use of computer simulation models to gain a deeper insight into, and elaborate on, the policy theory or possible causal mechanisms.

According to the Education Policy Outlook evaluations' digest, the most common quantitatively-focused technique was the compilation and/or analysis of statistical data. It, was used to make inferences about the success or impact of the policy. Statistical data are used in a variety of ways to inform the decision process. Example evaluations include:

- Evaluators of the introduction of free childcare in targeted areas in **Norway** conducted an analysis of detailed data on childcare attendance from the municipality of Oslo, as well as test scores from assessment tests in reading and mathematics in the first grades at school (age 6/7) provided by the municipality's department of education. These data were also merged with information about the children and their parents from Statistics Norway's population registers.
- An evaluation by the Ministry of Education of the **Slovak Republic** for the purposes of deciding on the expansion of childcare facilities used mainly statistical data (from the Statistical Office of the Slovak Republic, Eurostat, Institute of Forecasting of the Slovak Academy of Science and the Slovak Centre of Scientific and Technical Information). From these data, two indices were created to underpin the decision: 1) an investment efficiency index (multiple criteria in line with regional strategies); and 2) an underdevelopment index (number of inhabitants per dwelling/settlement) to decide on how and where expansion should be implemented.
- The Student Loan programme in **Hungary** provides loans to tertiary education students. The state-owned Student Loan Company (*Diakhitel*) compiles annual statistical data on borrowers, new loans and collections in order to monitor trends and identify challenges as they arise, as well as make tweaks to offerings and operations.

While statistical data can provide insights into key trends associated with the implementation of a particular policy, in general changes in trends can only be indirectly associated with reforms (with possible exceptions where the reform may remove or add legal requirements or entitlements to avail of educational services). Ideally, policy evaluation methods would be able to identify a cause-effect relationship between the introduction of a policy, as well as some defined, desirable outcomes, in order to adequately judge whether the policy was valuable.

Experimental evaluations

There have been increasing calls for causal approaches to be used in educational research in order to strengthen the evidentiary base for developing reforms. Policy makers want to find out "what works" (Slavin, 2004[125]). Experimental evaluations can be loosely described as a series of methods that seek to find out whether there is a causal relationship between a reform and a set of outcomes measured in the target population.

When studying policy interventions, causality can never be directly observed. It can be only inferred indirectly, as we can only observe one outcome for each treatment level. Causal inferences about policies can be attempted using econometric or statistical means, although econometric approaches tend to rely on observational data. The "gold standard" for inferring the effect of a policy is often considered to be through randomised controlled experiments (Athey and Imbens, 2017[126]).

Randomised controlled trials, when applied under the correct conditions, can assess the causal impact of a policy (Torgerson and Torgerson, 2001[127]). The trials involve the random assignment of individuals (or other units) into two groups. One group receives the

"treatment" of the reform while the other acts as a control. By measuring the differences in defined "success" outcomes between the groups, a measure of the causal effect of the reform can be obtained. The theoretical benefits of using randomised controlled trials for education evaluation are well recognised as allowing: evaluation without selection bias, for effects of different policy interventions to be disentangled, and for impacts of policies that have not yet been implemented to be investigated in controlled settings.

Experimental approaches have become increasingly prevalent as tools for the evaluation of education interventions in recent years (Sadoff, 2014[128]). However, randomised controlled trials are not a panacea; there are serious challenges to be considered by policy makers when taking the experimental approach. Depending on the nature of the intervention, such experiments in real life may not be deemed ethical. Experiments are also costly as they can only answer one research question at a time. (Sadoff, 2014[128]). Moreover, even where resources are available to conduct such experiments, randomised controlled trials can still be hampered by methodological flaws due to the complexities of designing experiments on reforms that aim to impact entire communities (Hawe, Shiell and Riley, 2004[129]). Further development and normalisation of the process of policy experimentation can help to build a critical mass to overcome some of the challenges and promote learning through replication, verification, and meta-analysis of experimental results (Makel and Plucker, 2014[130]).

Policy maker driven experimental approaches in education that examine causal links between reforms and outcomes have traditionally been rare (Cook, 2002[53]). There have been increasing calls for experimental approaches to evaluating education policies in recent years. However, meeting the strict conditions to conduct randomised experiments is not always feasible in education systems for a variety of reasons, including the difficulty of aligning continuously changing "field" settings in education to meet the textbook assumptions governing experiments (Alexander, 2006[131]).

Other forms of experimental design are commonly used where true experiments are not possible. These techniques involve making causal inferences under quasi-experimental conditions, where the control and treatment groups are not randomly assigned by the researcher. The lack of randomisation means there may be other differences between the groups that contribute to observed variations in outcomes of interest, and the study design must attempt to control for these differences. Examples of such methods include:

- ***Regression discontinuity design*:** This method exploits natural boundaries or discontinuities in some "forcing variable", where individuals on one side of the boundary qualify for the intervention, while those on the other do not. Examples of boundaries in this context could include, for example, youth intervention measures based on geographic areas, age limits, income limits, test scores or academic achievement measures. This is considered close to a natural experiment, as individuals who are close to the boundary on either side are assumed to have similar characteristics. Differences in outcomes between the groups close to the boundary that were subject to the policy and those that were not are then attributed to the causal impact of the policy using this method.
- ***Difference-in-differences*:** This is relevant in cases where some groups (such as schools, local areas, regions) participate in a policy reform and others do not (and act as a *de facto* control group) and the groups can be observed before and after the reform was introduced. In order to infer causality, a counterfactual needs to be estimated for the group which received the policy reform (i.e. an assessment of what their outcome would have been if they were not involved in the policy

reform). The estimation of this counterfactual can be informed by the changes in outcomes of the control group, also taking into account the initial differences between the two groups before treatment.

- Other methodologies seek to draw causal inferences. For example, the ***instrumental variables*** method could be used in cases where the assignment of those to be "treated" by the policy is not random, or where errors or changes occur not at random following the random assignment, creating a bias which will skew the results of the evaluation (where the evaluation is trying to define a statistical relationship between some outcome variable and explanatory variables). The method aims to identify one or more variables which are correlated directly with the explanatory variable, but not directly with the outcome variable. If technical conditions are met, the instrumental variables can be used to counteract bias.

Purely quantitative causal evaluation methods seem less prevalent in OECD education systems, and few examples exist within the Education Policy Outlook evaluations digest. Examples include:

- In **Spain**, an evaluation of the Territorial Co-operation Programme for Reducing Early Dropout was carried out by the University of Valencia to examine the evolution of school abandonment and discern its main features, as well as to analyse the probability of abandonment. Econometric techniques were employed, with determinant variables included related to the programme, along with personal and family characteristics of young people and environmental factors, such as the labour market situation (IVIE, 2014[132]).
- Also in **Spain**, the Programme for Reinforcement, Guidance and Support (PROA) is a school support programme directed at educational centres with students of lower socio-economic status, which includes additional tutoring, support, mentoring and programmes to change school culture and expectations. This programme underwent causal evaluation in 2014. The Universidad Pablo de Olavide (Seville) performed an exercise matching the individuals, centres and students of the treatment group with a control group similar in observable characteristics. Individual characteristics considered included gender, immigrant grade repetition, as well as parental education and occupation. Centre variables included whether the centre was public or private, and the percentage of students whose parents had tertiary education. Results were able to demonstrate that the effect of PROA is positive, with a particularly strong effect on reading. The evaluation also showed that the effects of the programme were cumulative and significant in both the short and long term (García-Pérez and Hidalgo, 2014[133]).

Simulation

In his Nobel lecture, James Heckman identified two conceptually distinct policy evaluation questions which are often confused: 1) "What is the effect of a programme in place on participants and nonparticipants compared to no programme at all or some alternative programme?"; and 2) "What is the likely effect of a new programme or an old programme applied to a new environment?" He considered the second question the more ambitious of the two (Heckman et al., 2001[134]). Simulation is a tool which can address this second question, and which is continuing to increase in popularity for decision support. Although quantitative in nature, simulation methods can often be used to gain insight into the current or possible contexts, processes and mechanisms underlying policies and their impacts, and therefore assist in theory building or revision.

Individual level simulation models are increasingly finding a place in the evaluation of public policies. The most well-established technique is microsimulation, a method which estimates reaction, behaviour and policy effects at the individual level using a realistic synthetic population, often based on real microdata from administrative sources (Spadaro, 2007[135]). More recently, agent-based models, which also incorporate the ability for simulations of interactions between individual population members, are coming to the fore and are being recognised as a necessary tool to realistically model behaviour and changes in large complex social systems (Farmer and Foley, 2009[136]).

Box 4.7. Simulation for the *ex*-ante evaluation of conditional cash transfer programmes

International organisations have long required *ex-ante* evaluations to be carried out on educational programmes as a condition of the grant funding process to make the case for allocating funding based on specific evidence of likely efficacy. One area where *ex-ante* evaluation has been used extensively is conditional cash transfer programmes, such as PROSPERA in Mexico (previously known as PROGRESA, and then *Oportunidades*) or the *Bolsa Escola* programme in Brazil, where governments transfer cash to households where children are at risk of not being in education, conditional on their children enrolling in school. The evaluations of these programmes use simulation methods to estimate the likely impact on enrolment rates and poverty levels, and can be used to generate scenarios showing both the impact of the policy and the counterfactual (i.e. the likely evolution of poverty levels and enrolment rates in the absence of the cash transfer programme).

It is important for the validation and calibration of *ex-ante* methodologies which rely on simulation models that the forecasted impacts are later compared with what actually transpired. For example, the World Bank, having conducted several *ex-ante* evaluations of cash transfer programmes in Mexico, Brazil and Ecuador, subsequently tested the efficacy of various *ex-ante* evaluation methods for these programmes by comparing the results of summative evaluations with the initial estimates and simulations in various *ex-ante* evaluations. The results of the analysis demonstrated the types of models that are most likely to give accurate estimates of impacts for cash transfer programmes and be of most use to policy makers and programme designers.

Source: Leite, P., A. Narayan and E. Skoufias (2011[137]), How do Ex Ante Simulations Compare with Ex Post Evaluations? Evidence from the Impact of Conditional Cash Transfer Programs, The World Bank, http://dx.doi.org/10.1596/1813-9450-5705.

4.5.2. *Longitudinal evaluation*

The impacts of education policies are wide ranging and can take years, even generations, to become fully evident (Oreopoulos, Page and Stevens, 2006[138]). Longitudinal policy evaluation is one method which can help to distinguish between temporary and more lasting effects of policies on their targets as it involves evaluating policy effects on the same units over a longer period of time. This differs from methods where the intervention status and the outcomes are measured at one point in time, as in a cross-sectional study (Setia, 2016[139]).

There are a number of approaches to carrying out longitudinal evaluation. One the one hand it can involve statistical or econometric analysis of panel data by employing some of the quasi-experimental methods discussed in section 4.4.1. This has the advantage of being relatively low cost, although the availability of relevant variables for the analysis may be limited. A specific survey related to the impacts of the policy on the targets can provide more tailored and valuable information, however, repeated follow-up fieldwork can be expensive and requires continued commitment, possibly across political cycles. Other acknowledged issues include dropout of participants and ensuring the consistency of measurements across time periods (Gerken, Bak and Reiterer, 2007[140]).

5. For what? Using evaluations for learning

...the intended purpose of evaluation is to make judgements about a specific programme or programmes at a particular point in time. Implicit in evaluation as decision-oriented inquiry is the presence of a user (or multiple users) for whom it is hoped the evaluation will have relevance.

International handbook of education evaluation (Kellaghan and Stufflebeam, 2003[25])

5.1. Getting value from evaluation

5.1.1. *Learning from evidence*

A long running debate exists as to whether education policy should be evidence based given the imperfect and often conflicting evidence available to decision makers, or evidence informed, which leaves the extent to which evidence is used open to interpretation and perhaps even minimisation (OECD, 2007[38]). High level, generalised policy directions indicated by evidence from large-scale national or international research may have greater scientific rigour, but may not be directly applicable or suitable in all local contexts. Meanwhile, successful localised pilot innovations often fail when applied on a wider scale throughout the system. There is therefore a tension in education between evidence-based and practice-based paradigms when developing policy. Increasing evaluation activity could help to shed light on the sources of these deviations and therefore improve policy quality.

The evaluation of education policy is concerned with both the generation and use of evidence. External evidence may be employed to validate the policy direction chosen before implementation or to assess the impact of the policy. The policy evaluation process itself also generates evidence through the analysis undertaken, which can be used to inform other reform processes. Limiting evaluation therefore also limits the field of evidence on which to base new policy initiatives.

Evaluation is only effective, and its cost can only be justified, if it promotes learning, i.e., if the results are used to inform future policies or modify existing approaches. Transforming education systems into learning systems depends on organisations and actors having the capacity both to evaluate their own work effectively and to learn from other evaluations and evidence. This learning requires two essential precursors to the use of evidence from evaluations: 1) the ability for evaluators to effectively disseminate their research and diffuse key messages into the education system; and 2) the capacity for policy makers to harness the messages and use them in the decision-making process.

Key messages from evaluations may not always be positive. This presents a risk to the institutionalisation of policy evaluation, due to “fear of failure” and the possibility of negative public commentary and political backlash (Bloch and Bugge, 2013[141]). Overcoming this risk will require a further cultural shift in public sectors towards more openness and away from risk aversion. Failure will need to be accepted as a possibility, handled appropriately, and treated as an opportunity for learning and systemic improvement (Wajzer et al., 2016[142]).

Evidence from the Education Policy Outlook digest suggests that evaluators often aim to distil messages from their work to be converted into better policy. Many evaluation reports contain recommendations for future action which can inform the policy going forward and promote improvement.

Box 5.1. Case study: Recommendations from evaluators of a leadership training and development programme in Norway

A leadership training and development programme was introduced in Norway in 2009 to provide training to school leaders, with priority for new leaders. The training focuses on five key areas: 1) the pupils learning outcomes and learning environment; two management and administration; 3) co-operation and organisational development; 4) development and change; and 5) the leadership role.

A series of four evaluation reports were produced on the programme by the Nordic Institute for Studies in Innovation, Research and Education (NIFU) and NTNU Social Research, with the final one published in 2014. The reports each focused on a different topic to capture the complexity of national leadership education through method triangulation and the combining of qualitative and quantitative data. The evaluation of this programme showed the good educational quality of its content and its relevance to the position of head of school. The programme was rated highly by participants in terms of pedagogical and didactical quality. Even with the high perception of success, evaluators identified a number of principles as a foundation for future policy directions. Key recommendations included:

- Future initiatives should continue to aim to integrate the national education programme in other school leadership programmes provided by higher education institutions in Norway.
- Many participants report that the current programme is challenging to attend alongside their regular job. Initiatives that ease the burden on participants should be considered.
- The participants report that the programme has led to a valuable and dynamic social network that stimulates individual and group learning afterwards. These "learning environments" should be supported and further developed as part of the programme.
- The links between the local municipalities (the school owners) and individual schools should be further developed. Change and development in schools following increased leadership competence is dependent on collaboration with the local municipalities.
- The national programme could benefit from being tied to other developmental projects organised by the Directorate of Education and Training.

Source: Hybertsen, et al. (2015[143]), Led to change: The National Leadership Education for School Principals in lower and upper secondary schools in Norway; change in the schools, goal achievement and recommendations, NIFU Nordic Institute for Studies in Innovation, Research and Education, https://www.researchgate.net/profile/Ingunn_Hybertsen/publication/279893842_Led_to_change_The_National_Leadership_Education_for_School_Principals_in_lower_and_upper_secondary_schools_in_Norway_change_in_the_schools_goal_achievement_and_recommendations_Fin.

5.1.2. *Balancing evidence and innovation*

There are continuing ideological debates about the role that evidence should play in devising education policy; in moving towards evidence-based education policy there is a risk of practitioners becoming focused on only choosing policy options for which evidence is available, at the expense of professional judgement and experience (Biesta, 2010[144]).

However, improvements to education can also come from innovations in technology, pedagogy or learning environments for which no prior evidence of efficacy exists. The recognised impact of education and skills on positive economic and social outcomes, combined with the pace of change in wider society, creates an innovation imperative for education systems (OECD, 2015[57]).

Innovation by its nature is the development of a novel approach to solve an existing problem (OECD, 2015[57]). The imperative for innovation can seem to be at odds with the requirement to ensure that policies are based on evidence, and therefore countries need to ensure that in an evidence-based culture there is still room for innovation to flourish. Innovative policies can be designed and improved by evaluations of policy experiments and the generation of evidence to support innovative practices.

Many education funders have begun to develop mechanisms for balancing funding according to the level of evidence available, while still promoting innovation. For example, Results for America, a non-profit organisation which works with the US federal government to shift funding to a more evidence-based footing, developed a tiered funding mechanism which provides grants based on the evidence-level available for the innovation (Results For America, 2015[145]). In Norway, the dedicated Programme for Research and Innovation in the Educational Sector (FINNUT) awards funding for projects which compile knowledge and evidence on current systemic contexts and practices, as well as those that propose innovations in key identified priority areas of education.

Brokering the knowledge from evaluations

The principle that educational change should be grounded in evidence is becoming more prominent (Cooper, Levin and Campbell, 2009[146]). There has been a long standing perception that education research has made less progress and is not as coherent as a discipline when compared with other policy areas, such as health. Nevertheless, progress has been made in recent years in providing a more structured and institutional framework for educational evidence. For example, the Campbell Collaboration for evidence-based social policy, which prepares and disseminates systematic reviews on the effectiveness of various social policy interventions (modelled after the Cochrane Collaboration for health interventions), includes an Education Co-ordinating Group (Odom et al., 2005[147]).

However, there are challenges with using evidence from evaluations effectively. These include a lack of capacity for communication and dissemination of key messages to wide audiences, poor evaluation design, and a lack of linkage between the evaluation outcomes and policy objectives (Patton, 2008[148]). There is a need to ensure that those responsible for policy development have the capacity to mine the volume of information available for meaning, and make judgements as to the quality of available evidence. Striking a balance between the over simplification of a message and providing too much technical detail which cannot be easily digested is a key requirement.

National authorities have begun to respond to calls to distil and present evidence from educational research to policy makers in an easily accessible way by actively developing initiatives such as clearinghouses and other central repositories of education evidence (Table 5.1).

Table 5.1. Distilling evidence effectively: Knowledge brokerage organisations and initiatives

Country	Organisation	Description
United Kingdom	Educational Endowment Foundation	The Educational Endowment Foundation, part of the UK Government "What Works" initiative network of evidence centres, is a research charity focused on building and utilising evidence to improve equity in education. The foundation conducts its own research and extracts information to present as "toolkits", which summarise in dashboard style evidence from various research studies, showing the comparative cost, evidence strength and measured impact on a visual scale for a large range of policy reform options and initiatives. The aim of the toolkits is to allow for policy makers and other users to get a quick overview of the strength of evidence supporting a particular course of action, and identify low cost high impact policy solutions.
Denmark	The Danish Clearinghouse for Educational Research	Established in 2006, the Danish Clearinghouse for Educational Research was one of the first education research clearinghouses to be established. The clearinghouse does not itself conduct research studies, but analyses and attempts to identify meaningful lessons from educational research covering all levels of education from early childhood to higher education. It produces systematic research mappings (which aim to compile the relevant research for a particular policy area) and systematic research reviews (which compile, analyse and synthesise the relevant evidence to tackle a specific research question). The review process takes a systematic and structured approach to identifying, mapping and collating the body of research available to assist the policy maker when considering education policy issues.
Norway	The Norwegian Knowledge Centre for Education	The Norwegian Knowledge Centre for Education was created in 2004 to gather and summarise the results of Norwegian and international evidence on education in a more accessible manner through the creation of an evidence database. The centre conducts systematic evidence reviews and analyses, as well as "state of the field" reviews which summarise major international developments in a given educational field since the beginning of the century. Through its web portal, the centre publishes summary "overviews" of its research, which also explicitly state who the research is primarily aimed at (policy makers, practitioners etc.).
Switzerland	The Swiss Co-ordination Centre for Research in Education	The Swiss Co-ordination Centre for Research in Education (SCCRE) is an institution under the auspices of the Swiss federal government and the Swiss Conference of Cantonal Ministers of Education (EDK). The centre promotes the exchange of information and research results between all stakeholders in the education system. A key ongoing function is to document and summarise education research projects and summarise and add the results and knowledge as an entry to a web-based database, which is available to the public.

At the international level, recent initiatives specifically focusing on the challenges of diffusing education research results include:

- Evidence in Education: a project by the OECD Centre for Educational Research and Innovation which addressed the question of effective brokering between policy makers and researchers (OECD, 2007[38]).
- Evidence-Informed Policy and Practice in Education in Europe (EIPPEE): a project funded by the European Commission which explored the links between research and decision making in education policy across Europe, with a view to developing mechanisms for knowledge brokerage and acting as a capacity-building exercise for utilising research in education systems. Through a survey of countries on their activities and mechanisms for linking research, the project found that the majority of activities were linked to communicating the results of research, with a much smaller focus on the use of research by policy makers later on (Gough et al., 2011[149]).
- Evidence-to-policy: a series of notes issued monthly by the World Bank that highlight the results of evaluations of many social initiatives supported by the World Bank, including education reforms. The notes aim to disseminate non-technical reviews of the growing number of robust evaluations of innovation (World Bank, 2016[150]).

5.2. What next? Reflection

As mentioned earlier in this paper, education policy evaluations still remain far from being systematic practices across OECD countries. However, there is increasing data availability and new methodologies that may provide for an expansion of evaluation activity in the future. In 2007, the OECD Centre for Educational Research and Innovation reflected (OECD, 2007[38]):

> *…in another dozen years we may be noting the same weaknesses in educational research and the same flaws in the communication between research and policy in education* […] *But some progress will also have been made, in all probability we can guess that rigorous research techniques will become more widely understood and applied, and practitioners and perhaps also policy makers will broaden their evidence base; and that the potential for brokering will have been explored in many more countries.*

This survey of the landscape, taking place at around the indicated dozen years later, suggests that at least some of these predictions have come to pass, which will continue to contribute towards a stronger evaluation infrastructure across all education systems.

A key feature of the database of evaluations compiled for this review is the **diversity of strategies employed** for evaluation. In complex education systems with diverse contexts, the responses to change the system, and how they are evaluated, need to be similarly flexible and varied.

Policy makers are increasingly aware that the relevance and applicability of policies between contexts is an important point to consider. **Context is vital** when reviewing the evidence of evaluations in order to decide whether or not a programme or reform is suitable for implementation in a different education system. While policy evaluations can show that a particular programme has been successful in improving outcomes in some contexts, the results are often not transferable when applied in a different context. Thus the question for policy makers when using evidence from policy evaluation to decide between policy options is not just "what works?" but "will it work here?" (Cartwright and Hardie, 2012[151]).

The vast majority of educational evaluations reviewed for this paper could not directly attribute changes in the target of the policy to the policy itself. It is difficult to disentangle the impacts of the policy from other policies which operate in parallel, or which would have occurred as a natural consequence of the contextual situation. Impacts of education policy on individuals compound on each other, they have a longitudinal and even an intergenerational horizon. Caution must therefore be taken in making causal links as this can contribute to a loss of insight or harmful misunderstandings about the policy implementation process.

However, even where causal links may be hard to draw, **carefully thought out evaluation strategies can still produce robust and actionable evidence**. Quantitative performance targets on outputs and outcomes are important measures of success. But not all policy effects can be measured quantitatively. Particularly in the case of education, reflection on the non-economic and wider social or public value of the reform initiative may be relevant (Bozeman and Sarewitz, 2011[152]). Evaluations can also provide insight into how policies change practices and attitudes at lower levels of the system, and therefore provide opportunities to learn more generally how to develop and implement reforms for maximum impact.

Based on the analysis conducted for this paper, three principles for evaluation emerge that can be useful for policymakers:

- ***Systems should "think evaluation", do not just do:*** Evaluation needs to be envisaged as a mind-set that everyone in the system can share; it goes beyond a specific process or activity. Developing it as a broader underlying capacity across the system and supporting it can bring benefits with growing returns to education systems in the longer term. Evidence shows some opportunities to do this, which include: when defining an education innovation, when engaging stakeholders, when designing evaluation processes, and when using evidence.
- ***Diverse contexts require a diverse portfolio:*** There is no one correct way to carry out policy evaluations in education systems. To promote true learning, it is necessary to employ a diversity of approaches, suitable to the context and objectives of the evaluation. Developing portfolios of evaluation instruments that are continuously updated through educational research could also be beneficial. Governments need to combine analysis of the policy context with a strategy and portfolio that ensure the most suitable methodological selection for evaluation is made for each reform process.
- ***Do what you should, not what you can:*** Evaluation should be driven by principles and issues, most relevant to the reform as well as theoretical analysis of the intended change process of the reform, rather than just the available methods. This theoretical step is often overlooked, but it can ensure that the right evaluation questions are posed and improve understanding as to why a reform has impacted in a certain way. In that sense, it can provide invaluable information for future policy decisions.

References

AITSL (2016), *Final Report – Evaluation of the Australian Professional Standards for Teachers*, https://www.aitsl.edu.au/docs/default-source/default-document-library/final-report-of-the-evaluation-of-the-apst.pdf?sfvrsn=428aec3c_0 (accessed on 31 May 2018). [102]

Alexander, H. (2006), "A View from Somewhere: Explaining the Paradigms of Educational Research", *Journal of Philosophy of Education*, Vol. 40/2, pp. 205-221, http://dx.doi.org/10.1111/j.1467-9752.2006.00502.x. [131]

Alhadeff-Jones, M. (2008), "Three Generations of Complexity Theories: Nuances and ambiguities", *Educational Philosophy and Theory*, Vol. 40/1, pp. 66-82, http://dx.doi.org/10.1111/j.1469-5812.2007.00411.x. [62]

Anderson-Levitt, K. (2003), "A World Culture of Schooling?", in *Local Meanings, Global Schooling*, Palgrave Macmillan US, New York, http://dx.doi.org/10.1057/9781403980359_1. [6]

Athey, S. and G. Imbens (2017), "The State of Applied Econometrics: Causality and Policy Evaluation", *Journal of Economic Perspectives—Volume*, Vol. 31/2, pp. 2017-3, http://dx.doi.org/10.1257/jep.31.2.3. [126]

Bakhshi, H. and J. Mateos-Garcia (2016), "New Data for Innovation Policy", NESTA, https://www.oecd.org/sti/106%20-%20Bakhshi%20and%20Mateos-Garcia%202016%20-%20New%20Data%20for%20Innovation%20Policy.pdf (accessed on 19 March 2018). [44]

Bamber, V. and S. Anderson (2012), "Evaluating learning and teaching: institutional needs and individual practices", *International Journal for Academic Development*, Vol. 17/1, pp. 5-18, http://dx.doi.org/10.1080/1360144X.2011.586459. [56]

Bechar, S. and I. Mero-Jaffe (2014), "Who Is Afraid of Evaluation? Ethics in Evaluation Research as a Way to Cope With Excessive Evaluation Anxiety: Insights From a Case Study", http://dx.doi.org/10.1177/1098214013512555. [79]

Biesta, G. (2010), "Why 'What Works' Still Won't Work: From Evidence-Based Education to Value-Based Education", *Stud Philos Educ*, Vol. 29, pp. 491-503, http://dx.doi.org/10.1007/s11217-010-9191-x. [144]

Bloch, C. and M. Bugge (2013), "Public sector innovation—From theory to measurement", *Structural Change and Economic Dynamics*, Vol. 27, pp. 133-145, http://dx.doi.org/10.1016/J.STRUECO.2013.06.008. [141]

Boston, J. and D. Gill (2017), *Social investment : a New Zealand policy experiment*, Bridget Williams Books, https://doi.org/10.1111/1467-8500.12391 (accessed on 19 March 2018). [40]

Bourguignon, F. and F. Ferreira (2003), "Ex-Ante Evaluation of Policy Reforms using Behavioral Models", in *Ex-ante evaluation of Policy reforms*, World Bank, https://pdfs.semanticscholar.org/8176/16daa09d0ec7f860e5cd070e5ab8c4ac2468.pdf (accessed on 3 April 2018). [93]

Bozeman, B. and D. Sarewitz (2011), "Public Value Mapping and Science Policy Evaluation", *Minerva*, http://dx.doi.org/10.1007/s11024-011-9161-7. [152]

Braun, A., M. Maguire and S. Ball (2010), "Policy enactments in the UK secondary school: examining policy, practice and school positioning", *Journal of Education Policy*, Vol. 25/4, pp. 547-560, http://dx.doi.org/10.1080/02680931003698544. [58]

Braverman, M. (2013), "Negotiating Measurement: Methodological and Interpersonal Considerations in the Choice and Interpretation of Instruments", *Americam Journal of Evaluation*, pp. 99-114, http://dx.doi.org/10.1177/1098214012460565. [115]

Bridgman, P. and G. Davis (2003), "What Use is a Policy Cycle? Plenty, if the Aim is Clear", *Australian Journal of Public Administration*, Vol. 62/3, pp. 98-102, http://dx.doi.org/10.1046/j.1467-8500.2003.00342.x. [28]

Buckley, J. et al. (2015), "Defining and Teaching Evaluative Thinking: Insights From Research on Critical Thinking", *American Journal of Evaluation*, http://dx.doi.org/10.1177/1098214015581706. [71]

Burgess, R. (2005), *The Ethics Of Educational Research*, Routledge, http://dx.doi.org/10.2307/3121426 (accessed on 1 April 2018). [80]

Burkhardt, H. and A. Schoenfeld (2003), "Improving Educational Research:Toward a More Useful, More Influential, and Better-Funded Enterprise", *Educational Researcher*, Vol. 32/9, pp. 3-14, http://dx.doi.org/10.3102/0013189X032009003. [75]

Burns, T. and F. Köster (2016), *Governing education in a complex world*. [61]

Cameron, D., A. Mishra and A. Brown (2016), "The growth of impact evaluation for international development: how much have we learned?", *Journal of Development Effectiveness*, Vol. 8/1, pp. 1-21, http://dx.doi.org/10.1080/19439342.2015.1034156. [35]

Cartwright, N. and J. Hardie (2012), *Evidence-based policy : a practical guide to doing it better*, Oxford University Press, https://doi.org/10.1017/S0266267114000091 (accessed on 30 May 2018). [151]

Century, J. and A. Cassata (2016), "Implementation Research", *Review of Research in Education*, Vol. 40/1, pp. 169-215, http://dx.doi.org/10.3102/0091732X16665332. [24]

Chelimsky, E. (2014), "Public-Interest Values and Program Sustainability", *American Journal of Evaluation*, Vol. 35/4, pp. 527-542, http://dx.doi.org/10.1177/1098214014549068. [51]

Chen, H. (2015), *Practical program evaluation : theory-driven evaluation and the integrated evaluation perspective*, Sage Publishing, https://us.sagepub.com/en-us/nam/practical-program-evaluation/book235546 (accessed on 30 May 2018). [91]

CIES (2011), *Estudoteip_sintese TEIP CIES*, http://ment/124676749/Estudoteip-sintese-TEIP-CIES (accessed on 31 May 2018). [101]

Comfort, L. (1982), *Education policy and evaluation : a context for change*, Pergamon Press, https://www.sciencedirect.com/science/book/9780080238562 (accessed on 16 March 2018). [32]

Connell, J. and A. Klem (2000), "You Can Get There From Here: Using a Theory of Change Approach to Plan Urban Education Reform", *Journal of Educational and Psychological Consultation*, Vol. 11/1, pp. 93-120, http://dx.doi.org/10.1207/s1532768Xjepc1101_06. [95]

Cook, T. (2002), "Randomized Experiments in Education: Why are they so rare?", Institute for Policy Research, http://www.ipr.northwestern.edu/publications/docs/workingpapers/2002/IPR-WP-02-19.pdf (accessed on 21 March 2018). [53]

Cooper, A., B. Levin and C. Campbell (2009), "The growing (but still limited) importance of evidence in education policy and practice", *Journal of Educational Change*, Vol. 10/2-3, pp. 159-171, http://dx.doi.org/10.1007/s10833-009-9107-0. [146]

Dandalo Partners (2014), *Evaluation of the National Partnership on Youth Attainment and Transitions*, http://www.dandolo.com.au (accessed on 31 May 2018). [112]

Datnow, A. (2005), "The Sustainability of Comprehensive School Reform Models in Changing District and State Contexts", *Educational Administration Quarterly*, Vol. 41/1, pp. 121-153, http://dx.doi.org/10.1177/0013161X04269578. [8]

Davis, B. and D. Sumara (2008), "Complexity as a theory of education", *Transnational Curriculum Inquiry 5 (2)*, http://nitinat.library.ubc.ca/ojs/index.php/tci (accessed on 22 March 2018). [59]

De Mauro, A., M. Greco and M. Grimaldi (2016), "A formal definition of Big Data based on its essential features", *Library Review*, Vol. 65/3, pp. 122-135, http://dx.doi.org/10.1108/LR-06-2015-0061. [153]

Department of Education and Skills (2017), *Literacy and Numeracy for Learning and Life 2011-2020*, https://www.education.ie/en/Publications/Education-Reports/pub_ed_interim_review_literacy_numeracy_2011_2020.PDF (accessed on 31 May 2018). [103]

DFØ (2016), *Guidance notes on the Instructions for Official Studies*, https://dfo.no/filer/Fagomr%C3%A5der/Utredninger/Guidance_notes_on_the_Instructions_for_Official_Studies_-_Instructions_for_the_Preparation_of_Central_Government_Official_Studies_V2.pdf. [78]

Earl, L. and H. Timperley (2015), "Evaluative thinking for successful educational innovation", *OECD Education Working Papers*, No. 122, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jrxtk1jtdwf-en. [73]

European Commission (2017), *Highlights of a year of activity*, http://publications.jrc.ec.europa.eu/repository/bitstream/JRC107563/kjna28715enn.pdf (accessed on 30 May 2018). [67]

European Commission (2014), *Internal Guidance on Ex Ante Conditionalities for the European Structural and Investment Funds PART I*, http://ec.europa.eu/regional_policy/sources/docgener/informat/2014/eac_guidance_esif_part1_en.pdf (accessed on 3 April 2018). [97]

European Commission (2010), *Europe 2020: A European strategy for smart, sustainable and inclusive growth*, http://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf (accessed on 21 March 2018). [43]

EVA (2008), *Arbejdet med elevplaner : en national undersøgelse af erfaringer (Working with student plans: a national study of experiences)*, https://www.eva.dk/grundskole/arbejdet-elevplaner (accessed on 31 May 2018). [104]

Farmer, J. and D. Foley (2009), "The economy needs agent-based modelling", *Nature*, Vol. 460/7256, pp. 685-686, http://dx.doi.org/10.1038/460685a. [136]

Fischer, F. and G. Miller (2007), *Handbook of Public Policy Analysis : Theory, Politics, and Methods*, Taylor and Francis Group, Boca Raton, https://doi.org/10.4324/9781315093192 (accessed on 12 March 2018). [27]

García-Pérez, J. and M. Hidalgo (2014), "Evaluación de PROA: su efecto sobre el rendimiento de los estudiantes *", http://www.mecd.gob.es/dctm/inee/evaluacionpct/pctproajigpmhupo.pdf?documentId=0901e72b81a1ab05 (accessed on 31 May 2018). [133]

Gerken, J., P. Bak and H. Reiterer (2007), *Longitudinal Evaluation Methods in Human-Computer Studies and Visual Analytics*, http://www.cs.umd.edu/hcil/InfoVisworkshop/papers/VIS_Workshop_gerken_bak.pdf (accessed on 17 May 2018). [140]

Goldin, C. and L. Katz (2008), *The race between education and technology*, Belknap Press of Harvard University Press. [2]

Gough, D. et al. (2011), "Evidence Informed Policymaking in Education in Europe EIPEE Final Project Report". [149]

Griñó, L. et al. (2014), *Embracing Evaluative Thinking for Better Outcomes:Four NGO Case Studies*, https://www.interaction.org/sites/default/files/EvaluativeThinkingReport_FINAL_online.pdf (accessed on 28 March 2018). [70]

Guskey, T. (2000), *Evaluating professional development*, Corwin Press. [52]

Hawe, P., A. Shiell and T. Riley (2004), "Complex interventions: how "out of control" quot; can a randomised controlled trial be?", *BMJ (Clinical research ed.)*, Vol. 328/7455, pp. 1561-3, http://dx.doi.org/10.1136/bmj.328.7455.1561. [129]

Heckman, J. (2008), "Econometric Causality", *International Statistical Review*, Vol. 76/1, pp. 1-27, http://dx.doi.org/10.1111/j.1751-5823.2007.00024.x. [54]

Heckman, J. et al. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture", *Journal of Political Economy*, Vol. 109/4, https://www.journals.uchicago.edu/doi/pdfplus/10.1086/322086 (accessed on 17 May 2018). [134]

Hertin, J. et al. (2009), "Rationalising the policy mess? Ex ante policy assessment and the utilisation of knowledge in the policy process", *Environment and Planning A*, http://dx.doi.org/10.1068/a40266. [99]

HM Treasury (2011), *The Magenta Book Guidance for evaluation*, HM Treasury, London, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220542/magenta_book_combined.pdf (accessed on 7 March 2018). [15]

Hopfenbeck, T. et al. (2013), *Balancing Trust and Accountability? The Assessment for Learning Programme in Norway A Governing Complex Education Systems Case Study*, http://www.oecd.org/education/ceri/Norwegian%20GCES%20case%20study%20OECD.pdf (accessed on 30 May 2018). [114]

Howlett, M. and M. Ramesh (1995), *Studying public policy : policy cycles and policy subsystems / Michael Howlett and M. Ramesh - Details - Trove*, Oxford University Press, https://trove.nla.gov.au/work/16820227 (accessed on 15 March 2018). [26]

Hua, H. and J. Herstein (2003), "Education Management Information System (EMIS): Integrated Data and Information Systems and Their Implications In Educational Management 1", http://www.infodev.org/sites/default/files/resource/InfodevDocuments_188.pdf (accessed on 1 April 2018). [77]

Human Resources and Skills Development Canada (2011), *Summative Evaluation of the Canada Student Loans Program*, http://publications.gc.ca/collections/collection_2012/rhdcc-hrsdc/HS28-44-1-2011-eng.pdf (accessed on 31 May 2018). [105]

Hybertsen, I. (2015), *Led to change: The National Leadership Education for School Principals in lower and upper secondary schools in Norway; change in the schools, goal achievement and recommendations*, NIFU Nordic Institute for Studies in Innovation, Research and Education, https://www.researchgate.net/profile/Ingunn_Hybertsen/publication/279893842_Led_to_change_The_National_Leadership_Education_for_School_Principals_in_lower_and_upper_secondary_schools_in_Norway_change_in_the_schools_goal_achievement_and_recommendations_Fin. [143]

IEKE (2016), *Internationale Expertenkommission zur Evaluation der Exzellenzinitiative Endbericht (International Commission of Experts for the Evaluation of the Excellence Initiative Final Report)*, https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/Imboden-Bericht-2016.pdf (accessed on 31 May 2018). [106]

IVIE (2014), *Evaluación del Programa de Cooperación Territorial para la reducción del abandono temprano de la educación (Evaluation of the Programme of territorial cooperation for reducing early school leaving)*, http://www.mecd.gob.es/dctm/inee/evaluacionpct/pctabandonoivie.pdf?documentId=0901e72b81a1ab03 (accessed on 31 May 2018). [132]

Jakobi, A. and J. Teltemann (2011), "Convergence in education policy? A quantitative analysis of policy change and stability in OECD countries", *Compare: A Journal of Comparative and International Education*, Vol. 41/5, pp. 579-595, http://dx.doi.org/10.1080/03057925.2011.566442. [5]

Janssens, F. and I. de Wolf (2009), "Analyzing the Assumptions of a Policy Program", *American Journal of Evaluation*, Vol. 30/3, pp. 411-425, http://dx.doi.org/10.1177/1098214009341016. [94]

Kearney, M. and R. Yelland (2010), "Higher Education in a World Changed Utterly Doing More with Less", http://www.oecd.org/site/eduimhe10/45925685.pdf (accessed on 7 March 2018). [9]

Kellaghan, T. and D. Stufflebeam (2003), *International Handbook of Educational Evaluation*, Springer Netherlands. [25]

Kitamura, Y. (2009), "Education Indicators to Examine the Policy-Making Process in the Education Sector of Developing Countries", https://www.gsid.nagoya-u.ac.jp/bpub/research/public/paper/article/170.pdf (accessed on 16 March 2018). [36]

Koenig, S. (2013), *La réforme de l'école fondamentale : rapport sur le premier bilan - Éducation nationale / Enfance / Jeunesse / Luxembourg*, http://www.men.public.lu/fr/actualites/publications/fondamental/statistiques-analyses/autres-themes/reforme-ef/index.html. [113]

Lagarde, F., J. Kassirer and L. Lotenberg (2012), "Budgeting for Evaluation: Beyond the 10% Rule of Thumb", http://dx.doi.org/10.1177/1524500412460635. [76]

Lauer, P. (2004), *A Policymaker's Primer on Education Research: How to Understand, Evaluate and Use It*, https://files.eric.ed.gov/fulltext/ED518626.pdf (accessed on 20 March 2018). [121]

Leite, P., A. Narayan and E. Skoufias (2011), *How do Ex Ante Simulations Compare with Ex Post Evaluations? Evidence from the Impact of Conditional Cash Transfer Programs*, The World Bank, http://dx.doi.org/10.1596/1813-9450-5705. [137]

Le, P. et al. (2012), *Implementation Completion and Results Report (IBRD-47670)*, http://documents.worldbank.org/curated/en/874001468110647154/pdf/NonAsciiFileName0.pdf (accessed on 31 May 2018). [108]

Levin, B. (2001), *Conceptualizing the process of education reform from an international perspective*, http://dx.doi.org/10.14507/epaa.v9n14.2001. [49]

Madaus, G., D. Stufflebeam and M. Scriven (1983), "Program Evaluation", in *Evaluation Models*, Springer Netherlands, Dordrecht, http://dx.doi.org/10.1007/978-94-009-6669-7_1. [30]

Ma, J., M. Pender and M. Welch (2016), "Trends in Higher Education Series The Benefits of Higher Education for Individuals and Society About the Authors", https://files.eric.ed.gov/fulltext/ED572548.pdf (accessed on 5 March 2018). [3]

Makel, M. and J. Plucker (2014), "Facts Are More Important Than Novelty: Replication in the Education Sciences", *Educational Researcher*, Vol. 43/6, pp. 304-316, http://dx.doi.org/10.3102/0013189X14545513. [130]

Martin, S. and I. Sanderson (1999), "Evaluating Public Policy Experiments", *Evaluation*, Vol. 5/3, pp. 245-258, http://dx.doi.org/10.1177/13563899922208977. [88]

Mason, M. (2008), "Complexity Theory and the Philosophy of Education", *Educational Philosophy and Theory*, Vol. 40/1, pp. 4-18, http://dx.doi.org/10.1111/j.1469-5812.2007.00412.x. [60]

Maxwell, J. (2004), "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education", *Educational Researcher*, Vol. 33/2, pp. 3-11, http://journals.sagepub.com/doi/pdf/10.3102/0013189X033002003 (accessed on 16 May 2018). [123]

Mayne, J. (2008), *Building an evaluative culture for effective evaluation and results management*, https://www.betterevaluation.org/sites/default/files/ILAC_Brief20_Evaluative_Culture.pdf (accessed on 21 October 2018). [69]

McConnell, A. (2010), "Policy Success, Policy Failure and Grey Areas In-Between", *Journal of Public Policy*, Vol. 30/03, pp. 345-362, http://dx.doi.org/10.1017/S0143814X10000152. [19]

Mertens, D. (2015), *Research and evaluation in education and psychology : integrating diversity with quantitative, qualitative, and mixed methods*, Sage Publications. [84]

Ministry of Education (2014), *Pasifika Education Plan Monitoring Report 2013 Pasifika Education Plan*, https://www.educationcounts.govt.nz/__data/assets/pdf_file/0005/164048/Pasifika-Education-Monitoring-Report-2013.pdf (accessed on 30 May 2018). [100]

Mitchell, L. et al. (2016), "ECE Participation Programme Evaluation Stage 4 Report to the Ministry of Education", https://www.educationcounts.govt.nz/__data/assets/pdf_file/0005/171851/ECE-Participation-Programme-Evaluation-Stage-4.pdf (accessed on 31 May 2018). [124]

National Audit Office (2013), *Evaluation in government*, https://www.nao.org.uk/wp-content/uploads/2013/12/10331-001-Evaluation-in-government_NEW.pdf (accessed on 16 March 2018). [37]

Neves, C. (2008), "International Organisations and the Evaluation of Education Systems: A Critical Comparative Analysis, European Journal of Vocational Training, 2008", *European Journal of Vocational Training*, Vol. 45/3, pp. 72-89, https://eric.ed.gov/?id=EJ836657 (accessed on 21 October 2018). [107]

New South Wales Department of Education and Communities (2014), *Evaluation Framework.*, https://education.nsw.gov.au/policy-library/associated-documents/evaluationframework.pdf (accessed on 16 March 2018). [82]

Odom, S. et al. (2005), *Exceptional Children Research in Special Education: Scientific Methods and Evidence-Based Practices*, https://pdfs.semanticscholar.org/6329/7d39c0c7a2503878705359f2743ed07b4af4.pdf (accessed on 21 October 2018). [147]

OECD (2019), *Education Policy Outlook 2019: Working Together to Help Students Achieve their Potential*, OECD Publishing, Paris, https://dx.doi.org/10.1787/2b8ad56e-en. [17]

OECD (2018), *Education Policy Outlook 2018: Putting Student Learning at the Centre*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264301528-en. [7]

OECD (2017), *Education at a Glance 2017: OECD Indicators*, OECD Publishing, Paris, http://dx.doi.org/10.1787/eag-2017-en. [4]

OECD (2017), *TALIS 2018 Video Study and Global Video Library on Teaching Practices: A Tool for Teacher Peer Learning*, http://www.oecd.org/education/school/TALIS-2018-video-study-brochure-ENG.pdf (accessed on 31 May 2018). [119]

OECD (2017), *The Funding of School Education: Connecting Resources and Learning*, OECD Reviews of School Resources, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264276147-en. [12]

OECD (2017), *The OECD Handbook for Innovative Learning Environments*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264277274-en. [74]

OECD (2016), *Supreme Audit Institutions and Good Governance: Oversight, Insight and Foresight*, OECD Public Governance Reviews, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264263871-en. [14]

OECD (2016), *Trends Shaping Education 2016*, OECD Publishing, Paris, http://dx.doi.org/10.1787/trends_edu-2016-en. [1]

OECD (2015), *Building on Basics*, Value for Money in Government, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264235052-en. [39]

OECD (2015), *Education Policy Outlook 2015: Making Reforms Happen*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264225442-en. [16]

OECD (2015), *Final NAEC Synthesis New Approaches to Economic Challenges*, http://www.oecd.org/naec/Final-NAEC-Synthesis-Report-CMIN2015-2.pdf (accessed on 7 March 2018). [10]

OECD (2015), *Government at a Glance 2015*, OECD Publishing, Paris, http://dx.doi.org/10.1787/gov_glance-2015-en. [11]

OECD (2015), *Social Impact Investment: Building the Evidence Base*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264233430-en. [42]

OECD (2015), *The Innovation Imperative in the Public Sector: Setting an Agenda for Action*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264236561-en. [57]

OECD (2014), *Measuring Science, Technology and Innovation*, http://www.oecd.org/sti (accessed on 21 October 2018). [66]

OECD (2014), "Mini Survey on Supreme Audit Institutions and Performance-Related Budgeting", http://www.oecd.org/governance/budgeting/Hand-Out%20-%20Mini%20survey%20Supreme%20Audit%20Institutions%20and%20Performance-relate%20%20%20.pdf (accessed on 22 May 2018). [13]

OECD (2013), *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264190658-en. [18]

OECD (2007), *Evidence in Education: Linking Research and Policy*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264033672-en. [38]

OECD (2005), *Modernising Government: The Way Forward*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264010505-en. [23]

OECD (1994), *Making Education Count: Developing and using international indicators*, https://files.eric.ed.gov/fulltext/ED411322.pdf (accessed on 16 March 2018). [34]

Olsen, K. and S. Reilly (2011), "Evaluation Methodologies A brief review of Meta-evaluation, Systematic Review and Synthesis Evaluation methodologies and their applicability to complex evaluations within the context of international development", http://www.iodparc.com. [46]

Open Knowledge International (n.d.), *Open Knowledge: What is Open?*, https://okfn.org/opendata/ (accessed on 21 October 2018). [154]

Oreopoulos, P., M. Page and A. Stevens (2006), "The Intergenerational Effects of Compulsory Schooling", *Journal of Labor Economics*, Vol. 24/4, https://www.journals.uchicago.edu/doi/pdfplus/10.1086/506484 (accessed on 17 May 2018). [138]

Panteia and SEOR (2016), *Monitoring and evaluation of the VSF policy 2012-2015*, https://www.rijksoverheid.nl/onderwerpen/vsv/documenten/rapporten/2016/02/15/monitoring-en-evaluatie-vsv-beleid-2012-2015 (accessed on 30 April 2018). [111]

Patton, M. (2008), *Utilization-focused evaluation*, Sage Publications, https://doi.org/10.1177/1098214010373646 (accessed on 21 October 2018). [148]

Pollitt, C. and G. Bouckaert (2017), *Public management reform : a comparative analysis - into the age of austerity*, Oxford University Press, https://global.oup.com/academic/product/public-management-reform-9780198795179?cc=fr&lang=en& (accessed on 21 March 2018). [50]

Preskill, H. and S. Gopal (2013), *Evaluating Complexity: Propositions doe Improving Practice*, FSG, http://www.pointk.org/resources/files/Evaluating_Complexity.pdf. [64]

RAMA and Ministry of Education (2010), *Evaluation of the New Horizon reform in elementary and junior high education at the end of three years of implementation: Summary of Findings*, http://cms.education.gov.il/NR/rdonlyres/45CCC357-4A00-42A6-9E8B-F892A24B202E/163977/OfekChadashsummary.docx. [120]

Ramirez, R. et al. (2013), *Utilization focused evaluation : a primer for evaluators*, Southbound, https://idl-bnc-idrc.dspacedirect.org/handle/10625/53020 (accessed on 21 October 2018). [109]

RAND Corporation (n.d.), *Education Reform | RAND*, https://www.rand.org/topics/education-reform.html (accessed on 14 March 2018). [20]

Reisman, J., A. Gienapp and S. Stachowiak (2007), *Measuring Advocacy and Policy*, The Annie E. Casey Foundation, Baltimore, https://www.alnap.org/system/files/content/resource/files/main/aecf-aguidetomeasuringpolicyandadvocacy-2007.pdf (accessed on 25 April 2018). [116]

Results For America (2015), *Invest in What Works Fact Sheet: Evidence-Based Innovation Programs - Results for America*, https://results4america.org/tools/invest-works-fact-sheet-federal-evidence-based-innovation-programs/ (accessed on 24 May 2018). [145]

Ritchie, J. et al. (2013), *Qualitative research practice : a guide for social science students and researchers*, Sage Publications. [122]

RTI International (2016), *Measures of quality through classroom observation for the Sustainable Development Goals: lessons ...; Background paper prepared for the 2016 Global education monitoring report, Education for people and planet: creating sustainable futures for all; 2016*, http://unesdoc.unesco.org/images/0024/002458/245841E.pdf (accessed on 4 April 2018). [117]

Rutter, J. (2012), *Evidence and Evaluation in Policy Making*, https://www.instituteforgovernment.org.uk/sites/default/files/publications/evidence%20and%20evaluation%20in%20template_final_0.pdf. [55]

Sadoff, S. (2014), "The role of experimentation in education policy", *Oxford Review of Economic Policy*, Vol. 30/4, pp. 597-620, http://dx.doi.org/10.1093/oxrep/grv001. [128]

Sanderson, I. (2002), "Evaluation, Policy Learning and Evidence-Based Policy Making", *Public Administration*, http://dx.doi.org/10.1111/1467-9299.00292. [83]

Scheerens, J., C. Glas and S. Thomas (2007), *Educational evaluation, assessment, and monitoring : a systemic approach*, Taylor & Francis, https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.4324%2F9780203971055 (accessed on 3 April 2018). [89]

Schlotter, M., G. Schwerdt and L. Woessmann (2011), "Econometric methods for causal evaluation of education policies and practices: a non-technical guide", *Education Economics*, Vol. 19/2, pp. 109-137, http://dx.doi.org/10.1080/09645292.2010.511821. [45]

Scriven, M. (1991), *Evaluation thesaurus, 4th ed.*, https://psycnet.apa.org/record/1991-98719-000 (accessed on 17 November 2019). [87]

Setia, M. (2016), "Methodology Series Module 3: Cross-sectional Studies.", *Indian journal of dermatology*, Vol. 61/3, pp. 261-4, http://dx.doi.org/10.4103/0019-5154.182410. [139]

Shlonsky, A. and R. Mildon (2014), "Methodological pluralism in the age of evidence-informed practice and policy", *Scandinavian Journal of Public Health*, Vol. 42/13_suppl, pp. 18-27, http://dx.doi.org/10.1177/1403494813516716. [86]

Sims, R., G. Dobbs and T. Hand (2002), "Enhancing Quality in Online Learning: Scaffolding Planning and Design Through Proactive Evaluation", *Distance Education*, Vol. 23/2, pp. 135-148, http://dx.doi.org/10.1080/0158791022000009169. [92]

Slavin, R. (2004), "Education Research Can and Must Address "What Works" Questions", *Educational Researcher*, Vol. 33/1, pp. 27-28, http://dx.doi.org/10.3102/0013189X033001027. [125]

Slavin, R. (1995), "Best evidence synthesis: an intelligent alternative to meta-analysis.", *Journal of clinical epidemiology*, Vol. 48/1, pp. 9-18, http://www.ncbi.nlm.nih.gov/pubmed/7853053 (accessed on 21 March 2018). [48]

Smismans, S. (2015), "Policy Evaluation in the EU: The Challenges of Linking Ex Ante and Ex Post Appraisal", *European Journal of Risk Regulation*, Vol. 6/01, pp. 6-26, http://dx.doi.org/10.1017/S1867299X00004244. [29]

Spadaro, A. (2007), *Microsimulation as a tool for the evaluation of public policies : methods and applications*, Fundación BBVA, https://www.fbbva.es/en/publicaciones/microsimulation-as-a-tool-for-the-evaluation-of-public-policies-methods-and-applications-2/ (accessed on 17 May 2018). [135]

Stecher, B. et al. (2006), "Using Structured Classroom Vignettes to Measure Instructional Practices in Mathematics", https://www.rand.org/pubs/working_papers/WR336.html (accessed on 4 April 2018). [118]

Stufflebeam, D. (2001), "Evaluation Models 2", *New directions for evaluation* 89, https://www.wmich.edu/sites/default/files/attachments/u58/2015/Evaluation_Models.pdf (accessed on 16 March 2018). [33]

Stufflebeam, D. (2001), "The Metaevaluation Imperative", *American Journal of Evaluation*, Vol. 22/2, pp. 183-209, http://dx.doi.org/10.1177/109821400102200204. [47]

Taut, S. and D. Brauns (2003), "Resistance to Evaluation A Psychological Perspective", *Evaluation*, Vol. 9/3, pp. 247-264, http://journals.sagepub.com/doi/pdf/10.1177/13563890030093002 (accessed on 29 May 2018). [81]

Torgerson, C. and D. Torgerson (2001), "The Need for Randomised Controlled Trials in Educational Research", *British Journal of Educational Studies*, Vol. 49/3, pp. 316-328, http://dx.doi.org/10.1111/1467-8527.t01-1-00178. [127]

Trochim, W. (2009), "Evaluation policy and evaluation practice", *New Directions for Evaluation*, Vol. 2009/123, pp. 13-32, http://dx.doi.org/10.1002/ev.303. [68]

Trombly, C. (2014), "Schools and Complexity", *Complicity: An International Journal of Complexity and Education*, Vol. 11/1, https://journals.library.ualberta.ca/complicity/index.php/complicity/article/view/19017 (accessed on 22 March 2018). [63]

UNESCO (2007), *Standard-setting at UNESCO. Vol. 1, Normative action in education, science and culture : essays in commemoration of the sixtieth anniversary of UNESCO*, Martinus Nijhoff, https://doi.org/10.1163/ej.9789004164505.1-430 (accessed on 3 April 2018). [98]

Van Den Akker, J., B. Bannan and A. Kelly (2013), *Educational Design Research Part A: An introduction*, SLO • Netherlands institute for curriculum development, http://international.slo.nl/publications/edr/ (accessed on 4 April 2018). [90]

Van Geert, P. and H. Steenbeek (2014), "The Good, the Bad and the Ugly? The Dynamic Interplay Between Educational Practice, Policy and Research", *Complicity: An International Journal of Complexity and Education*, Vol. 11/2, https://journals.library.ualberta.ca/complicity/index.php/complicity/article/view/22962/17093 (accessed on 22 March 2018). [65]

Vedung, E. (2000), *Public policy and program evaluation*, Transaction Publishers. [22]

Viennet, R. and B. Pont (2017), "Education policy implementation: A literature review and proposed framework", *OECD Education Working Papers*, No. 162, OECD Publishing, Paris, http://dx.doi.org/10.1787/fc467a64-en. [21]

Wajzer, C. et al. (2016), *Failing Well - Insights on dealing with failure and turnaround from four critical areas of public service delivery*, Institute for Government, https://www.instituteforgovernment.org.uk/sites/default/files/publications/IFGJ4331_Failing-Well_25.07.16_WEBc.pdf (accessed on 29 May 2018). [142]

Weitzman, B. and D. Silver (2013), "Good Evaluation Measures: More Than Their Psychometric Properties", *American Journal Of Evaluation*, http://dx.doi.org/10.1177/1098214012461628. [110]

White, H. (2013), "An introduction to the use of randomised control trials to evaluate development interventions", *Journal of Development Effectiveness*, Vol. 5/1, pp. 30-49, http://dx.doi.org/10.1080/19439342.2013.764652. [85]

Wilson, K. (2014), "New Investment Approaches for Addressing Social and Economic Challenges", *OECD Science, Technology and Industry Policy Papers*, No. 15, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz2bz8g00jj-en. [41]

Wolpin, K. (2007), *Model Validation and Model Comparison*, https://pubs.aeaweb.org/doi/pdf/10.1257/aer.97.2.48 (accessed on 3 April 2018). [96]

World Bank (2016), *Brief- Evidence to Policy*, http://www.worldbank.org/en/programs/sief-trust-fund/brief/evidence-to-policy (accessed on 24 May 2018). [150]

Worthen, B. and J. Sanders (1987), *Educational evaluation : alternative approaches and practical guidelines*, Longman. [31]

Wyatt, T. (2017), "Developing evaluative thinking and evidence- based practice: A synthetic case study", https://research.acer.edu.au/cgi/viewcontent.cgi?article=1316&context=research_conference (accessed on 29 March 2018). [72]

Annex 5.A. Analytical questions

The following questions were employed to develop a comparative overview of the characteristics of reform evaluation policies in OECD countries, using the key lines of analysis presented in Section 3.

Who?

1) What is the position of the evaluator within the system?
2) Do the evaluators have independence of choice to undertake an evaluation, and to publish evaluation reports?
3) Do evaluators have access to the data and evidence they need?
4) Do the evaluators adhere to a set of professional guidelines or standards?

When?

1) Are policies routinely evaluated? When are policies most likely to be evaluated?
2) Is evaluation integrated into the policy development process?
3) Does evaluation take place to assess "fitness" before implementation?
4) Is there evidence of evaluative thinking throughout the cycle of a policy?
5) How are *ex-ante* and summative evaluations linked?

How?

1) How is the method of evaluation design validated and quality assured?
2) Which methods are used, and what is the basis for choosing these methods?
3) Are evaluation methods chosen with context in mind?
4) How are uncertainty and unreliability dealt with in evaluations?

What?

1) Are measures which can be evaluated defined at the beginning of the policy process? What is the basis for decision on measurements of impact?
2) Are measures relevant and suitable to assess impact of the reform? Are measures tied to wider national and international policy objectives?
3) How are evaluation questions composed?

For what?

1) Are the objectives for utilisation of the evaluation clearly set out? Is there some foresight or perception in advance of how the evaluation will be used?
2) How is the evaluation disseminated and shared with relevant stakeholders?
3) How does the evaluation feed into future policy initiatives?
4) How do policy makers use evaluation when devising reforms?
5)
6)