1. Introduction: Evaluation as an instrument in social policy

Bent Greve

1.1 INTRODUCTION

The purpose of this book is to theoretically and empirically describe issues related to the evaluation of social policy and the use of best evidence, or knowledge, in a variety of fields related to social policy. The aim is to look into types of evaluation, methodological issues related to evaluation, and provide concrete examples of evaluation, including how they relate to social policy decisions. The book is therefore divided into three parts, elaborated on below.

Social policy is understood here in a very broad sense, including in terms of health, labour market policy, long-term care, care and support for the disabled, and education. Furthermore, the chapters encompass both in-kind and cash benefits in the analysis, therefore covering large parts of welfare state spending, how this can have an impact on quality of life for individuals, and how best to use available resources.

In many countries, development in evidence-based policies in recent years has meant that there is a greater need to know what is working and what is not (Peters, 2015). For a long time, evaluation of policies was done *ex post* with the aim of trying to ensure that policy decisions were carried out in the best way. Today, evaluation is not only made to verify whether decisions and implementation have attained their goals, but also whether, before implementing new initiatives, there is evidence indicating that one choice will be better than another. Naturally, there are pitfalls and dilemmas in using and collecting evidence, and these are also described in the book.

Economists have learned that resources are scarce. For them the debate on how best to use them and prioritize them among different options is thus not new. However, even while acknowledging the need to prioritize among different options, and being aware that money spent on one thing cannot be spent on another (opportunity costs), this in itself does not show how to make an evaluation and also how to use evidence in policy-making. As this book reveals, in many different ways, there are methodological issues to knowing what is the best instrument to use, the problems with

ensuring the right data are used, and, with so many different issues being interrelated, it is always possible (as least for scientists) to criticize one analysis and argue that more knowledge is needed to be sure of what is the best evidence.

Given the pressure on and changes in welfare states (Greve, 2013) the developments around how to use this best evidence might not come as a surprise. Many of the changes around the world in welfare states have been related to income transfers, such as social assistance, unemployment benefits and pensions. Here the need for knowledge on what works best might be less strong in relation to impact of income transfers on incentives and quality of life,¹ including happiness (Easterlin, 2001). Criteria for receiving benefits and incentives to find a job have been analysed by looking into the impact of active labour market policy and the ability to enter or re-enter the labour market. In relation to welfare services there is a need to know what works and what gives the best return on spending and investment in the different areas. This is difficult, given that quality of welfare services can in certain cases be difficult to depict and define.

In recent years, however, there has been increased interest in issues other than money as metrics for development and the need to know more about relations between monetary issues and non-monetary issues (Stiglitz et al., 2009; Helliwell et al., 2015), including how to evaluate impact on happiness and well-being in comparison with other more traditional economic metrics. Bringing non-monetary issues into decision-making and evaluation, including quality of life, does not make it easier to find the best evidence. However, there is still an indication that there is a need to know more about why and how different kinds of decisions influence both the use of money in the welfare state and the broader impact on the level of welfare in different countries. By this it is also clear that implementation within countries and transfer of knowledge from one country to another can and will always be difficult as context and policy priorities (and ideologies) differ.

This book will not be an odyssey in theory and philosophy of science. However, methodological issues will be debated, as this is part of the controversy about how and to what extent it is possible to use evaluation in the development of social policy. These debates range from – at least theoretically – the argument that (1) only concrete evidence should be used, to (2) that particularly when it comes to social services, finding evidence on social interaction and social relations is difficult or impossible. By focussing on methodological issues and how evaluation is used in policy-making, and probing into different areas of the welfare state, this book tries to depict and describe these many and varied angles.

This is also evidenced by the variety of researchers who have contributed

to the book, as they come from a range of disciplines and social policy areas. In healthcare, for example, use of evidence has been central for many years; see, for example, Clark et al. (2013) on obesity. This is despite the fact that what was once thought of as best evidence might be different today given our ability to use new types of analysis, better instruments of measurement and better calculations. The existence of and impact of big data makes evaluation more feasible today. And there are clear rules regarding how and when new medicines can be introduced, whereas there is not the same clarity in many areas of social care, education or other social policy, where the lines are more blurred. As another example, there are conflicting arguments as to whether and to what degree, frontline social workers should have the discretion, based upon their own professional judgment, to decide how to help their clients in the best way. Trust in social workers' discretion has thus been seen as better or at least as important as following specific and detailed approaches set by administration. Trust in social workers, in contrast to using best evidence, is thus an example of a possible conflict.

The overall impact of using best evidence in preventative initiatives is a further issue. In relation to preventative instruments, there is also a possible conflict between spending now and saving later, including how this can be done when budgets are tight, and, for example, there are rules regarding public sector deficit. Spending more on the promise of saving later might not persuade policy-makers or those who will have to pay or lend money for the purpose. Policy-makers' use (or misuse) of evidence is also included in the presentations.

1.2 WHAT IS EVALUATION AND WHAT IS BEST EVIDENCE?

There is no clear and single definition of evaluation, even though research in the field began in the 1930s (Palfrey et al., 2012). Naturally, decision-makers, having always been interested in what they can get for taxpayers' money, also have to be sure that they know what is going on, especially in the wake of the growth of public sector spending in many welfare states. Performance indicators have included, for example, elements such as how many patients to treat, how many students to educate, and how many children to take care of and so on. Focus on output has always been a central aspect in the development of indicators and collection of data.

Further, there is no clear and simple definition of best evidence. It is often seen as 'rooted in the concept of "experience", relating to what is manifest and obvious' (Billings, 2013, p. 299). Methodologically, therefore, there is

also a constant debate between quantitative and qualitative approaches to find out whether there is evidence for one thing or another, including how, if possible, to transfer approaches from one context to another, and even more so between countries.

There have also been developments in the areas of auditing and accreditation. With regard to the principal-agent dilemma – that the principal is not sure the agent who is making decisions on his behalf is doing what is expected of her and is motivated by her own interests – by setting standards as to what is to be delivered there are at least some verifiable elements. Accreditation may be seen as an instrument where the provider, if accredited, might be allowed to undergo less evaluation, due to the trust that the provider delivers high-quality services.

Historically, there have also been ideas around trade-offs. One idea relates to equality and efficiency (Okun, 1975). Simple rules are easier to administer, but this might be at the cost of not achieving goals of equality. Paying for welfare might have an impact on the efficiency of the overall economy, although how strong this link is depends, for example, on whether one looks at the welfare state from a social investment perspective or a more traditional economic perspective.

Another issue influencing development has been that although indicators have been set and used, they have not always revealed the effectiveness of different interventions – as argued already in the mid-1980s (Pollitt, 1985). Reaching a target might be at the expense of other goals. For example, a planned economy might demand that manufacturers should produce so many tons of furniture, and they do so by using very heavy materials to reach the target. However, when the furniture is made it is too heavy to move! Naturally, this anecdote indicates that one needs to be cautious of how to set targets, and then how to achieve these, including how to measure the development. Again the principal-agent relationship is seemingly important, but there is also the risk that you only get what you measure, and a risk of changing perspective if the indicators for performance are not used in a reliable way.

In evaluation there also needs to be a clear programme theory and an approach linking research and practice (Olsson, 2007), otherwise the usefulness of an evaluation will be reduced and there might be a risk that not all of the important evidence is included in the analysis. Programme theory explains how an intervention is understood to contribute to a chain of results that produce the intended or actual impacts, positive or negative. It can be a very useful way of bringing together existing evidence about a programme, and clarifying where there is agreement and disagreement about how the programme is understood to work, and where there are gaps in the evidence.

The issue of how to measure things is not only one of looking into the number of operations, children cared for, or activities completed. It is also an issue of how to analyse change in societies and political decisions. For example, when one wants to depict incremental changes in welfare states the measurement needs observable and reliable indicators (Rocco and Thurston, 2014). Further, there is also a need to look into how the stakeholders in an evaluation (such as funders, managers, project workers, target population, other practitioners, policy-makers, politicians, academics, etc.) are behaving and integrated in the measurement and analysis (Green and South, 2006).

There has been an increasing focus on the fact that targets or changes need to be seen from a qualitative perspective. Targets might be reached in the number of medical procedures carried out for example, but if many patients get the wrong treatment the number of operations carried out is not a good indicator of success. At the same time, it is often more difficult to measure quality in an evaluation, for example, in care for the elderly. Perceived quality by users is important, but might not reflect the quality of an intervention.

Evaluation can be used to judge the outcome of activities, and whether the set goals have been achieved. However, evidence from interventions can also be used as part of an argument for preventative initiatives, if it can show that an intervention helps to reduce future costs, and that social investment in an area can prevent negative developments. A classic example is Heckman and Masterov (2007), who argue that early interventions for children help long-term productivity, and that preschool training is better than later schooling and job training when looking at the return of investment in human capital.

The issue of social investment is an important aspect of evaluation, given that it will always be difficult to measure issues related to quality of life, well-being and happiness. These issues, however, can also be an important outcome of an intervention, and integrating these into an evaluation provides a more holistic approach (Leck et al., 2014). Thus, evaluation is not always about controlling performance, but about how to use knowledge in order to make decisions. Whether this is always the case is also discussed in this book.

There are thus many and different types and models of evaluation (goal-oriented, the decision model, evaluation research, goal-free evaluation, evaluation of alternative explanations for impacts, utilization-oriented, etc.), but one can also look at results, process, system, economic, actor and programme theory (Palfrey et al., 2012). Thus, when making decisions there might be interests at stake that point to different types and ways of evaluation, and, how to use them. This can be seen as an argument

Figure 1.1 The evidence hierarchy

for trying to know what the best evidence is. However, what the best evidence is varies across disciplines, although the evidence hierarchy seems schematic and very stylized in form (Figure 1.1), albeit with variation in what there is overall agreement about.

This also raises the issue of when a generalization is possible. Naturally, as in the famous discussion around a Popperian view, observing a single black swan contradicts the generalization that all swans are white (Fuller, 2004; Taleb, 2007), so in this sense a single case study can provide new knowledge, meaning that context matters when interpreting and understanding evaluation in different areas and subjects.

Nevertheless, there is now a gold standard to aim for. However, it is not, as will be discussed and shown in several places in this book, possible to achieve in all areas and all topics. This means that sometimes when we are unable to obtain the best knowledge, we might need to use the second-best solution and second-best level of knowledge. Still, some knowledge, most of the time, is better than no knowledge at all. Naturally, quantitative analysis is useful if there are theoretical and solid foundations for the expected causal relations – for example, a causal relation might exist, but, if not grounded in theoretical knowledge and also informed by institutional structures in a society, this might be misleading. McCambridge et al. (2014), for example, discuss problems related to bias in randomized controlled trials (RCTs).

Thus, it is the concrete empirical analysis that needs to point towards which type of intervention is the best. Naturally, in the analysis one needs

to find out how best to achieve and get the necessary information in order to conduct a proper evaluation.

1.3 OVERVIEW OF THE BOOK

The book is divided into three parts. Below is a short overview of the individual chapters in each part after this introductory chapter.

Part I: What Evaluation is and Examples of Methods

Part I focuses on the definition of evaluation and gives examples of different methods, thus presenting an overview of the central aspects and variety of approaches. This also includes differences in the availability of data and use of data.

The use of RCTs in social policy fields is the focus of Chapter 2 by Boruch et al. The chapter includes examples of how to use them and how they have been used in social policy fields, including crime and education. The chapter also discusses the strong and weak points of using RCTs. A central methodological problem is how to ensure that participating individuals are randomly chosen when making the intervention and, further, that the comparison group is suitable.

In Chapter 3, by Mueller and Gaus, the focus is on the important issues of whether the implementation of social policy interventions has the intended consequences, and/or if there are unintended outcomes. This is mainly due to the difficulties related to estimating unbiased causal intervention effects, and although RCTs are thought to serve this purpose, RCTs are not always feasible in evaluation practice, and, further, there may be a conflict between internal and external validity.

Public and private investments are often intended to give a return – one way or another. Due to this, money spent on one project cannot be spent on another. The social return on investment (SROI) perspective, including general elements of cost–benefit analysis is presented by Costa in Chapter 4. The chapter highlights that social issues, although difficult to measure, should also be included in an analysis of social policy, and major appraisal tools for public projects are discussed.

One specific focus, central to Chapter 5 by Boardman and Vining, is on how to make an economic evaluation of social policy. The chapter provides an overview of the methods used to evaluate social policies, which focus on the criterion of allocative efficiency, including both comprehensive methods (cost–benefit analysis) and non-comprehensive economic evaluation methods. It also discusses some of the important

issues associated with the economic evaluation of social policies, including some of the pitfalls.

Naturally, there is a need for critical assessment of the use of systematic reviews. This is the focus of Chapter 6 by Sundberg. The chapter contains an overview of different types of reviews, but also strengths and weaknesses of the different methodologies, and puts forward realist synthesis as a possible approach when trying to evaluate social policy interventions.

An often overlooked perspective in evaluation is how to integrate participants and their knowledge into the evaluation. This is the focus of Chapter 7 by Richardson, who argues that the user's perspective and knowledge can be important not only as a source of knowledge itself, but also as a way of gathering data for the analysis. This might be a future avenue to explore for the field of healthcare.

Part II: Evaluation and Policy

In Part II the focus is on evaluation and policy-making. Policy-makers can use evaluation for change – or they can decide not to use it. They might use it when it suits their purpose, but in other instances decide that for political reasons they prefer another solution. Thus, the intersection of policy and evaluation is important. Once the evidence is obtained, the question then is how to use it.

Chapter 8 by Smith and Haux has exactly this focus: when and how evidence is used in policy-making and its possible impact. They explore the idea of evidence-based policy-making (EBPM) and consider the absence of agreed definitions. They discuss various models of policy-making and their criticisms and take a more practical focus to consider the issues that arose when the UK's New Labour government tried to take an evidence-based approach to policy-making.

Not only is there the use of evidence, but also its possible misuse, which is the focus of Chapter 9 by Deeming. This may occur when policy-makers only use the evidence that is in line with their own ideas and perceptions as to how society should develop. Thus, there is a risk that policy-makers will not use evidence and evaluation that point in the opposite direction. Another, classical, issue is whether policy-makers have the right to, given their preferences, take the decisions they prefer.

This theme is pursued in another way in Chapter 10 by Liket, who discusses how policy-makers, given evaluation and knowledge of what works, combine accountability for decisions with ensuring cost-effectiveness. Liket also points to some main challenges to evaluations, including poor distinctions between performance, monitoring and evaluation, low-quality evaluation questions, and finally problems in data collection.

Chapter 11 by Warren and Wistow on policy, evidence and difference within welfare regimes focuses on the UK through the example of benefits for those unable to work due to disability and/or ill health. Despite being set within one regime, there are strong regional and local differences in this particular area of welfare state policy and practice. Finally, they present a cross-national comparison of different countries with different regimes.

In Chapter 12 by Hansen the focus is on performance management. As argued in the chapter, performance management is necessary, furthermore, that it will continue to be used, but also that there are advantages and disadvantages. Thus, politicians, managers and employees in the public sector need to familiarize themselves with performance management in order to increase the possible benefits of using performance management and minimize its possible disadvantages.

Chapter 13 on critical perspectives by Dahler-Larsen presents arguments and counterarguments for and against using evidence in social policy. The chapter shows that positions range from those who point to practical implementation problems, to those who question the practical and political relevance of evidence, to those who discuss whether evidence serves specific interests. Thus, the chapter also looks into the dilemmas and pitfalls involved when using evidence.

Social impact bonds are the focus of Chapter 14 by Roy, McHugh and Sinclair. They argue that there is a lack of engagement with ethical issues in the use of these bonds, including the possible neglect of non-economic issues in policy-making. Therefore, the use of social impact bonds might have negative implications for the implementation of social policy.

Part III: Evaluation of Concrete Social Policy Areas

Part III looks into concrete social policy areas, in order to see what the present state of the art is within different central welfare state policies, as there have been strong differences in the development and use of evaluation within different sectors. These chapters provide a good overview of traditions and knowledge in central welfare state areas.

Chapter 15 on healthcare by Knapp and Evans-Lacko focuses on how to understand and evaluate the overall healthcare system, including how to look into prevention, equity in access to healthcare, and also how to monitor the economic resources within the healthcare system.

A central question in social policy is on what basis one should accept and use new types of medicines. This is the focus of Chapter 16 by Calnan and Douglass. Medicine is an area where there has long been a focus on what works, and if there are possible side-effects. Therefore, there is a need for strong criteria for when and how to use new types of medicine.

Long-term care is a specific policy area of growing importance in social policy. Chapter 17 by Rodrigues addresses challenges and problems in evaluating long-term care. A very central issue is the quality of life in long-term care – as well as a precise definition of what long-term care actually is. Measuring effectiveness is also very difficult given that those in need of long-term care might increase their ability to live independently, but their health condition may not improve.

Labour market policy has been a central policy area for a long time, especially with a need to know what works. Chapter 18 by Eichhorst and Konle-Seidl presents an overview of its possible impact. It is argued that this is an area where we have more knowledge, but also a need for good solid data and a methodological approach ensuring that we measure what we want to measure. There is also a difference in evaluation at the micro- and macro-levels.

Social policy intervention for people with disabilities is the focus of Chapter 19 by Etherington. Here the focus is on the possible impact on the lives of people with disabilities due to the increased focus on work first and activation within the paradigm of active labour market policy. Seemingly, ideology is more important than knowledge about what works best within the area.

Integration in society, especially for migrants and refugees, is central in social policy, including knowing what will have the best impact. This is the core issue in Chapter 20 by Hagelund and Kavli. After a critical assessment of the concept they also point to the difficulty in actually measuring integration, and that lack of data and understanding of what integration is has a strong impact. One will presumably always be in favour of integration, but one needs to define and know what kind of intervention can lead to integration.

Children's and young people's well-being is a central aspect in many welfare states. Chapter 21 by Edbrooke-Childs et al. explores the many and varied methodological issues related to the ability to analyse interventions and support for this population. Besides methodological reflections, there is a discussion of what has been learnt, and how this should be able to inform policy-makers on what to do in relation to policies.

Chapter 22 by Van Lancker focuses on child and family policy, and also that there might be intended as well as unintended impacts of different kind of family policies. He argues that there might be a tendency towards a 'Matthew effect' in the field – those who already have are given more.

Chapter 23 by Hansen et al. looks at what to be aware of when implementing large-scale policy reforms. They argue in favour of eight crucial attention points based upon evaluation of public sector reforms:

indicators of how to use evaluation as an instrument for improving future implementation of large-scale reforms.

In Chapter 24 by Hussain different ways to measure poverty are discussed, but also, and even more important, how evaluation of poverty interventions can be used as a way of ensuring the best and most efficient approach when it comes to trying to reduce poverty. Further, concrete examples of interventions with the aim of reducing poverty are presented.

Chapter 25 by Brown reflects on evaluation within the field of education, including use of the OECD's Programme for International Student Assessment (PISA) survey as part of the evaluation. Focus is also on the relation between profession, evaluation and policy in an area where strong professional interest groups are working and thus use of evidence-based interventions might work against professionals' understanding in the given context of what works best.

The book ends with some conclusions and perspectives related to evidence-based policies in Chapter 26 by Greve.

1.4 LIMITATIONS

Naturally a book like this is not able to cover all fields and all areas of evaluation in social policy, given that it is such a large area. This means, for example, that only more limited technical approaches are included, and that these relate more particularly to the more mature welfare states – we could have included even more specific areas and examples of more concrete evaluations. However, the idea has been to present an overview of central elements and ideas within the broader topics of social policy and welfare state analysis, thereby enabling the reader to obtain a comprehensive overview and also an idea of how evidence-based analysis is carried out.

NOTE

1. For an overview for many countries and different indicators see the OECD Better Life Index, accessed 12 October 2016 at <http://www.oecdbetterlifeindex.org/#/1111111111>.

REFERENCES

- Billings, J. (2013), 'Improving the evidence base', in S. Leichsenring, J. Billings and H. Nies (eds), *Long-Term Care in Europe: Improving Policy and Practice*, London: Palgrave Macmillan.
- Clark, R., E. Waters, R. Armstrong, R. Conning, S. Allender and B. Swinburn (2013), 'Evidence and obesity prevention: Developing evidence summaries to support decision making', *Evidence and Policy*, **9**(4), 547–56.
- Easterlin, R.A. (2001), 'Income and happiness: Towards a unified theory', *The Economic Journal*, **111**(473), 465–84.
- Fuller, S. (2004), *Kuhn vs. Popper*, New York: Columbia University Press.
- Green, J. and J. South (2006), *Evaluation*, Maidenhead, UK: Open University Press.
- Greve, B. (ed.) (2013), *The Routledge Handbook of the Welfare State*, Abingdon, UK: Routledge.
- Heckmann, J. and D. Masterov (2007), 'The productivity argument for investing in young children', *Review of Agricultural Economics*, **29**(3), 446–93.
- Helliwell, J., R. Layard and J. Sachs (eds) (2015), *World Happiness Report 2015*, accessed 12 October 2016 at <http://worldhappiness.report/wp-content/uploads/sites/2/2015/04/WHR15.pdf>.
- Leck, C., D. Upton and N. Evans (2014), 'Social return on investment: Valuing health outcomes or promoting economic values?' *Journal of Health Psychology*, **21**(7), 1–10.
- McCambridge, J., K. Kypri and D. Elbourne (2014), 'In randomization we trust? There are overlooked problems in experimenting with people in behavioral intervention trials', *Journal of Clinical Epidemiology*, **67**(3), 247–53.
- Okun, A. (1975), *Equality and Efficiency: The Big Trade-Off*, Washington, DC: Brookings Institution Press.
- Olsson, T. (2007), 'Reconstructing evidence-based practice: An investigation of three conceptualisations of EBP', *Evidence and Policy*, **3**(2), 271–85.
- Palfrey, C., P. Thomas and C. Phillips (2012), *Evaluation for the Real World: The Impact of Evidence in Policy Making*, Bristol: Policy Press.
- Peters, B.G. (2015), *Advanced Introduction to Public Policy*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Pollitt, C. (1985), 'Measuring performance: A new system for the NHS', *Policy & Politics*, **13**(1), 1–15.
- Rocco, P. and C. Thurston (2014), 'From metaphors to measures: Observable indicators of gradual institutional change', *Journal of Public Policy*, **34**(11), 35–62.
- Stiglitz, J., A. Sen and J.P. Fitoussi (eds) (2009), *The Measurement of Economic Performance and Social Progress. Revisited, Reflections and Overview*, Paris: Commission on the Measurement of Economic Performance and Social Progress.
- Taleb, N.N. (2007), *The Black Swan: The Impact of the Highly Improbable*, New York: Random House.