Evidence-based policy or policy-based evidence?

The effect of policy commitment on government-sponsored evaluation in Britain (1997-2010)

Arnaud Vaganay

A thesis submitted to
the Department of Methodology
of the London School of Economics
for the degree of Doctor of Philosophy

London, December 2014

[This page was intentionally left blank]

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I consider the work submitted to be a complete thesis fit for examination.

I authorise that, if a degree is awarded, an electronic copy of my thesis will be deposited in LSE Theses Online held by the British Library of Political and Economic Science and that, except as provided for in regulation 41 it will be made available for public reference.

I authorise the School to supply a copy of the abstract of my thesis for inclusion in any published list of theses offered for higher degrees in British universities or in any supplement thereto, or for consultation in any central file of abstracts of such theses.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 61,053 words.

Arnaud Vaganay

London, 12 December 2014

[This page was intentionally left blank]

Acknowledgements

I would like to express my very great appreciation to Dr Jouni Kuha, my supervisor, for his patient guidance, availability, encouragement and useful critiques of this research project.

My grateful thanks are also extended to:

– Pr Christopher Hood and Pr Edward Page for examining the final version of my thesis;

– Pr Martin Lodge and Pr Kenneth Benoit for their feedback on my 'Upgrading Material';

– Dr Paul Mitchell and Pr Jonathan Jackson for their feedback on my 'Aims and Methods Paper';

– Pr Richard Jackman, for his advice during the first year of my thesis;

– Dr Philippe Warin and Dr Stéphanie Abrial for supporting my decision to start a PhD;

– The London School of Economics and Political Science and the Department of Methodology for their financial support;

– All the people whom I have interviewed in the framework of this project and those who have provided information and guidance.

I would also like to thank my colleagues and friends for their professional and moral support in this long and difficult journey. I am particularly grateful to: Dr Katrin Hohl; Dr Claudia Abreu Lopes; Ms Ioanna Gouseti; Dr Carolyn Côté Lussier; Dr Stavroula Tsirogianni; Dr Aude Bicquelet; Mr John Fyson; Ms Esther Heyhoe.

Finally, I wish to thank my parents, my family and my friends for their support and encouragement throughout my PhD.

[This page was intentionally left blank]

Abstract

In most mature welfare states, policy evaluations are sponsored by the very organisations that designed and implemented the intervention in the first place. Research in the area of clinical trials has consistently shown that this type of arrangement creates a moral hazard and may lead to overestimates of the effect of the treatment. Yet, no one so far has investigated whether social interventions were subject to such 'confirmation bias'.

The objective of this study was twofold. Firstly, it assessed the scientific credibility of a sample of government-sponsored pilot evaluations. Three common research prescriptions were considered: (a) the proportionality of timescales, (b) the representativeness of pilot sites; and (c) the completeness of outcome reporting. Secondly, it examined whether the known commitment of the government to a reform was associated with less credible evaluations.

These questions were answered using a 'meta-research' methodology, which departs from the traditional interviews and surveys of agents that have dominated the literature so far. I developed the new PILOT dataset for that specific purpose. PILOT includes data systematically collected from over 230 pilot and experimental evaluations spanning 13 years of government-commissioned research in the UK (1997-2010) and four government departments (Department for Work and Pensions, Department for Education, Home Office and Ministry of Justice). PILOT was instrumental in (a) modeling pilot duration using event history analysis; (b) modeling pilot site selection using logistic regression; and (c) the systematic selection of six evaluation reports for qualitative content analysis. A total of 17 interviews with policy researchers were also conducted to inform the case study and the overall research design.

The results show little overt evidence of crude bias or 'bad' design. On average, government-sponsored pilots (a) were based on timescales that were proportional to the scope of the research; (b) were not primarily designed with the aim of warranting representativeness; and (c) were rather comprehensively analysed in evaluation reports. In addition, the results indicate that the known commitment of the government to a reform had no significant effect on the selection of pilot sites and on the reporting of outcomes. However, it was associated with significantly shorter pilots.

In conclusion, there is some evidence that the known commitment of a government to a reform is associated with less credible evaluations; however this effect is only tangible in the earlier stages of the research cycle. In this respect, sponsorship bias would appear to be more limited than in the context of industry-sponsored clinical trials. Policy recommendations are provided, as this project was severely hindered by important 'black box' issues and by the poor quality of evaluation reports.

[This page was intentionally left blank]

Table of contents

[This page was intentionally left blank]

Table of exhibits

List of acronyms

CDG	Child Development Grant
Cm	Command Paper
CRC	Cooperative Research Centre
DfE	Department for Education
DWP	Department for Work and Pensions
EASG	Evidence and Analysis Steering Group
ESRC	Economic and Social Research Council
FDA	Food and Drug Administration
FoI	Freedom of Information
GSE	Government Science and Engineering
GSE	Government Science and Engineering
GSR	Government Social Research
HC	House of Commons
HEO	Higher Executive Officer
HMT	Her Majesty's Treasury
HO	Home Office
IB	Incapacity Benefit
IfG	Institute for Government
IS	Income Support
JCP	Jobcentre Plus
JCPD	Jobcentre Plus District
JO	Job Outcome
JRRP	Job Retention and Rehabilitation Pilot
JSA	Jobseeker's Allowance
LA	Local Authority
LSE	London School of Economics and Political Science
LSHTM	London School of Hygiene and Tropical Medicine

MoJ	Ministry of Justice
MP	Member of Parliament
MTFC	Multi-Treatment Foster Care
NAO	National Audit Office
NDA	New Drug Application
NDPB	Non-Departmental Public Body
NUTS	Nomenclature of Territorial Units for Statistics
OECD	Organisation for Economic Cooperation and Development
ONS	Office for National Statistics
ORB	Outcome Reporting Bias
PID	Project Initiation Document
PSB	Public Service Bargain
PtW	Pathways to Work
R&D	Research and Development
RCT	Randomised Controlled Trial
WWEG	Work, Welfare and Equality Group

1. Introduction

> *"We need social scientists to help determine what works and why, and what policy initiatives are likely to be most effective, and we need better ways of ensuring that those who need such information can get it quickly and easily"* – David Blunkett MP[1].
>
> *"It is my experience that in terms of studies you can probably get an academic to do anything you want"* – Eric Pickles MP[2].

1.1. Background

1.1.1. Evidence-based policy *vs.* policy-based evidence

Governments the world over increasingly monitor and evaluate their policies. Are these evaluations driven by policy-makers' will to learn about the effectiveness of their reforms and make 'better' policies – as suggested by David Blunkett? Or are they conducted to legitimate decisions that have already been made – as suggested by Eric Pickles? This is the question I will be addressing in the following pages.

Theoretically, the case is unclear. Some, pointing to the rhetorical continuity between the 1997 New Labour Manifesto, pledging to implement 'what works' and the recent What Works Initiative launched by the Coalition government, have concluded to a rationalisation and a depoliticisation of policy-making (Winner, 1997). This policy framework is underpinned by the idea of 'instrumental rationality' (Dryzek, 1990) whereby more and better evidence leads to more and better policy. In the UK, the belief in this framework was entrenched by a number of core New Labour documents in the late 1990s and early 2000s including *Modernising Government* White Paper (Cabinet Office, 1999a), and *Professional Policy Making for the Twenty First Century* (Cabinet Office, 1999b) and institutionalised through the creation of a number of new units including the Performance and Innovation Unit, Social Exclusion Unit and Centre for Management and Policy Studies. However, some have argued that this framework was not just a development of New Labour but was deeply entrenched in the positivist worldview of many policy professionals (Morçöl, 2001).

Others also argued that, in a political system still structured by elections, evaluation was the "continuation of politics by other means" (Bovens, 't Hart, & Kuipers, 2008). Through the evaluation of policies, so the argument

[1] Secretary of State for Education (1997-2001), speech to the ESRC, 2 February 2000.
[2] Secretary of State for Communities and Local Government (2010-), oral evidence given to the Communities and Local Government Committee, 12 September 2011.

goes, politicians have found a way to maintain their legitimacy in times of ambiguity, risk and insecurity, coupled with descriptions of government as no longer in control or even entirely legitimate. Doubt and uncertainty increasingly permeate relationships between citizens and politicians, citizens and state professionals and even citizens and the market (Beck, 1992). This is recorded in opinion surveys that report a loss of trust in government and politicians in the latter part of the 20th century, reductions in voter turnout at UK general elections in the early 21st century, as well as media reports about the diminishing authority of professionals among service users, for example the lack of respect shown to teachers in the classroom (Keat, Whiteley, & Abercrombie, 1994; Pharr & Putnam, 2000). In the UK, New Labour sought to deal with this creeping doubt by explicitly acknowledging the conditions of the 'new governance' and its challenges, but presenting them as opportunities to 'modernise' public services and key institutions and 'renew' democratic practice.

1.1.2. Overarching Research Question

Individual opinions, however, often lean strongly towards the latter theory. Examples of 'misuses' of scientific advice by British governments abound and often make for popular news stories and scholarly case studies. They include decisions such as the multiple reclassifications of cannabis, the culling of badgers, the introduction of a minimum pricing for alcohol, the abolition of the Education Maintenance Allowance or the increase of the number of police-on-the-beat. Far fewer have been the examples of 'good' uses of scientific advice during the same period.

In mentioning these examples, my point is not to take a position but to warn against negativity bias and hasty conclusions. First, these accounts are mainly based on single case studies and case studies are not meant to be widely generalizable. Second, these examples tend to be referred to specifically because they fit the cynical and widely held view that politicians put their interests before those of the people they serve.

The idea that underpins this thesis is that we have now accumulated a large number of case studies and individual accounts of the role of research in public policy. Many of them are discussed in this thesis. However, we are still missing 'the big picture', i.e. the typical use of research across a wide range of conditions. To paraphrase Brint, we need to shift the discussion "from an analysis of variations to a general characterisation" of the role of research in policy (Brint 1990). To get to this point, as I will show throughout this thesis, we need different theories and methods that those which have been used so far.

Thus, this thesis contributes to the following Overarching Research Question: *To what extent do political institutions influence policy evaluation?* (See Exhibit 1).

1.1.3. Audiences

I believe that this question is relevant to at least three audiences. Readers with an interest in research methods (including statistics, economics, and epidemiology) might find it interesting to see how the scientific norms withstand the test of 'real-world research'. Colleagues working in the broad fields of political science, economics and management will be offered a new approach to study research decisions in an organisational context. In addition, policy analysts and evaluators will hopefully get a better understanding of some of the implications of contract research.

This thesis focuses on the UK, therefore British researchers and all researchers interested in Britain constitute its primary audience. However, I believe that the appeal of this dissertation goes beyond these borders. There is a growing – if recent – interest in policy evaluation at all levels of government and in all corners of the world. This thesis has some relevance for other countries and scientific disciplines as well.

1.2. Theoretical contribution

This thesis makes a number of contributions to the literature; some of them are direct, others are indirect. The first contribution is theoretical and concerns the way the effect of institutions on policy research should be approached.

Our collective knowledge of the effect of institutions on policy evaluation is scattered across different disciplines. Useful contributions have been made in philosophy, sociology, research methods, political science (including public policy and public administration) as well as in the professional literature (mainly education and nursing). However, none of them is in itself sufficient to understand the phenomena at hand. There is still no dominant theory of the effect of institutions on policy research, which might explain why progress has been slow so far, as shown in section 2.

This thesis approaches the question as an example of 'confirmation bias'. Confirmation bias is the tendency of individuals and organisations to favour information that supports their prior hypotheses, beliefs or commitments (Plous, 1993). This theory, which was developed in social psychology, is an important addition to the political science literature. Indeed, it brings together a myriad of existing concepts – agency, blame, reputation, utilisation – in a more parsimonious and widely applicable concept.

Thus, the central question in this thesis can be formulated as follows: *To what extent is the research conducted or commissioned by political institutions subject to confirmation bias?*

The notion of confirmation bias illustrates the trade-off that researchers must resolve between their commitment to the scientific method and their commitment to the intervention they are asked to evaluate (which can be spontaneous or imposed). This tension can be observed in at least three research decisions; each of which forms a separate empirical chapter in this thesis (see below and Exhibit 1). These questions are:

- The time afforded to research;
- The sampling of units;
- The reporting of evaluation outcomes.

1.3. Methodological contribution

This thesis offers a new tool for the analysis of policy evaluation activities: the PILOT dataset.

1.3.1. Approach

Studies into the role of research in policy-making have essentially relied on small-N designs, looking at cases or pairs of cases in isolation but with a view to foster a deep understanding of the underlying social phenomena. Besides, the data has mainly been based on subjective data, including interviews and surveys of policy-makers as well as self-reported anecdotes from researchers. To be sure, small-N studies and subjective data have significantly fostered our understanding of policy research decisions by exploring the subject and identifying key variables.

This being said, the existing methodological toolbox is not adapted to find out the extent to which policy research decisions accommodate institutional constraints. There are two reasons behind this. Firstly, a limited number of cases does not allow analysing the prevalence of the phenomenon. What is needed instead is a quantitative analysis accompanied by relevant inferences. Secondly, I need reliable and factual data that is comparable across a maximum of cases. Interviews and surveys are not the most appropriate methods.

Here again, the medical research literature provides a useful example of how this can be achieved. Research on scientific integrity has relied on structured data (studies) to see if, for example, clinical trials were more favourable to test drugs when those trials were funded by the drug manufacturer rather than by the regulator.

Noting that an ever larger number of policy studies have been conducted over the past 20 years and that the vast majority of these studies are now available online, this thesis tries, for the first time on this issue, to replicate this research design with the PILOT dataset.

As a preliminary step, I also conducted a series of 17 interviews with policy researchers, with the aim of better understanding the decision-making process in British ministerial departments and to identify the key variables to be included in the dataset as well the availability of data.

1.3.2. The PILOT dataset

I developed the PILOT dataset for that specific purpose. It focuses on the pilot schemes launched in the UK between May 1997 and May 2010, which corresponds to the Labour governments. Pilot schemes are policies trialled for a limited period on a fraction of the territory on which they are meant to be rolled out. They were chosen as unit of analysis for two reasons. First, the availability of a control group not receiving the new intervention means that more methodologies can be used for the evaluation of outcomes. Second, *ex post* evaluations of national programmes often take place too late in the policy cycle, when new policies are already in place and the interest in the old policy has vanished. PILOT includes 233 pilot schemes systematically identified in three policy areas: (1) employment and welfare; (2) education and parenting; and (3) crime and justice.

PILOT includes three categories of variables. Firstly, variables pertaining to the research design of each study. The duration of the research is one of them, as well as the locations and the reported outcomes of each pilot. Secondly, the dataset provides information related to the type of policy intervention being piloted. Those include the target group of the intervention and the type of policy instrument tested (spend vs. regulatory interventions). It also contains policy-specific variables. For example, labour market programmes have been classified as mandatory or voluntary. Thirdly, the dataset offers data related to the political context of each pilot. It includes both straightforward facts, such as the time between the start of the pilot and the next general election and whether or not the pilot implements a manifesto pledge.

PILOT was instrumental in (a) modeling pilot duration using event history analysis; (b) modeling pilot site selection using logistic regression; and (c) the systematic selection of six evaluation reports for qualitative content analysis.

1.4. Empirical contribution

This new theoretical and methodological framework helped me answer a series of specific questions, which can be put under two ‘empirical strands’.

1.4.1. Scientific credibility of government-sponsored evaluations (Empirical Strand 1)

A key assumptions underpinning this thesis is that research is subject to strict professional norms. Thus, we have an idea of the type of research decision that an organisation committed to scientific norms would make.

The first empirical contribution of this thesis is thus fairly descriptive; but I believe that it has a substantive interest. Bearing in mind the different prescriptions associated with the three research decisions mentioned in section 1.2 and presented in Exhibit 1, my thesis brings an answer to the following questions:

- Is the duration of pilots proportional to the complexity of the intervention and the evaluation? (Specific Question 1a);
- Are pilot sites representative of the population? (Specific Question 1b);
- Are the intervention outcomes comprehensively reported in pilot evaluation reports? (Specific Question 1c).

1.4.2. Effect of policy commitments on the scientific credibility of evaluations (Empirical Strand 2)

The second empirical contribution of my thesis is to analyse the effect of policy commitments on the scientific credibility of the research decisions of interest. An observable effect would bring evidence of confirmation bias.

The first of these decisions is the time afforded to an evaluation. Time is a precious resource for both the evaluator and the policy-maker, but for opposite reasons. Whereas the former tend to press for longer evaluations, which will allow her to collect more and more robust data, the latter often advocate shorter evaluations, most of them imposed by the political agenda. Against this background, I formulated the following specific question: *Are interventions to which the government is strongly committed subject to shorter evaluations?* (Specific Question 2a).

Previous studies have shown that research duration could be affected by the salience of the intervention (Carpenter, 2002, 2004; Dranove & Meltzer, 1994; Olson, 1997). However, the evidence base is limited to clinical trials and drug approval processes.

The contribution of this thesis is to provide a new model capturing the effect of policy commitment on the duration of evaluations.

The second research decision likely to reflect confirmation bias is the sampling of the units to be included in the evaluation. Whereas a commitment to scientific norms will lead evaluators to select samples that

are representative of the wider population, a commitment to the intervention is more likely to favour samples perceived as exemplary. To address this question, I had to focus on a specific policy area, namely employment and welfare programmes. The following question was formulated: *Are high-performing Jobcentre Plus districts more likely to be selected as pilot sites?* (Specific Question 2b).

Previous studies have pointed to the limited external validity of some investigations (Keitner, Posternak, & Ryan, 2003; Pratt & Moyé, 1995; Rothwell, 2005). However, the evidence base is limited to clinical trials.

This thesis provides a new model capturing the effect of a district's performance on its probability of being selected as pilot site.

The third research decision which might indicate confirmation bias if the reporting of the findings of the evaluation. It is expected that an evaluation influenced by scientific norms will report outcomes fully and completely, whereas an evaluation influenced by policy commitments will report these findings selectively. Against this background, I formulated the following specific question: *Are interventions to which the government is strongly committed subject to more spin?* (Specific Question 2c).

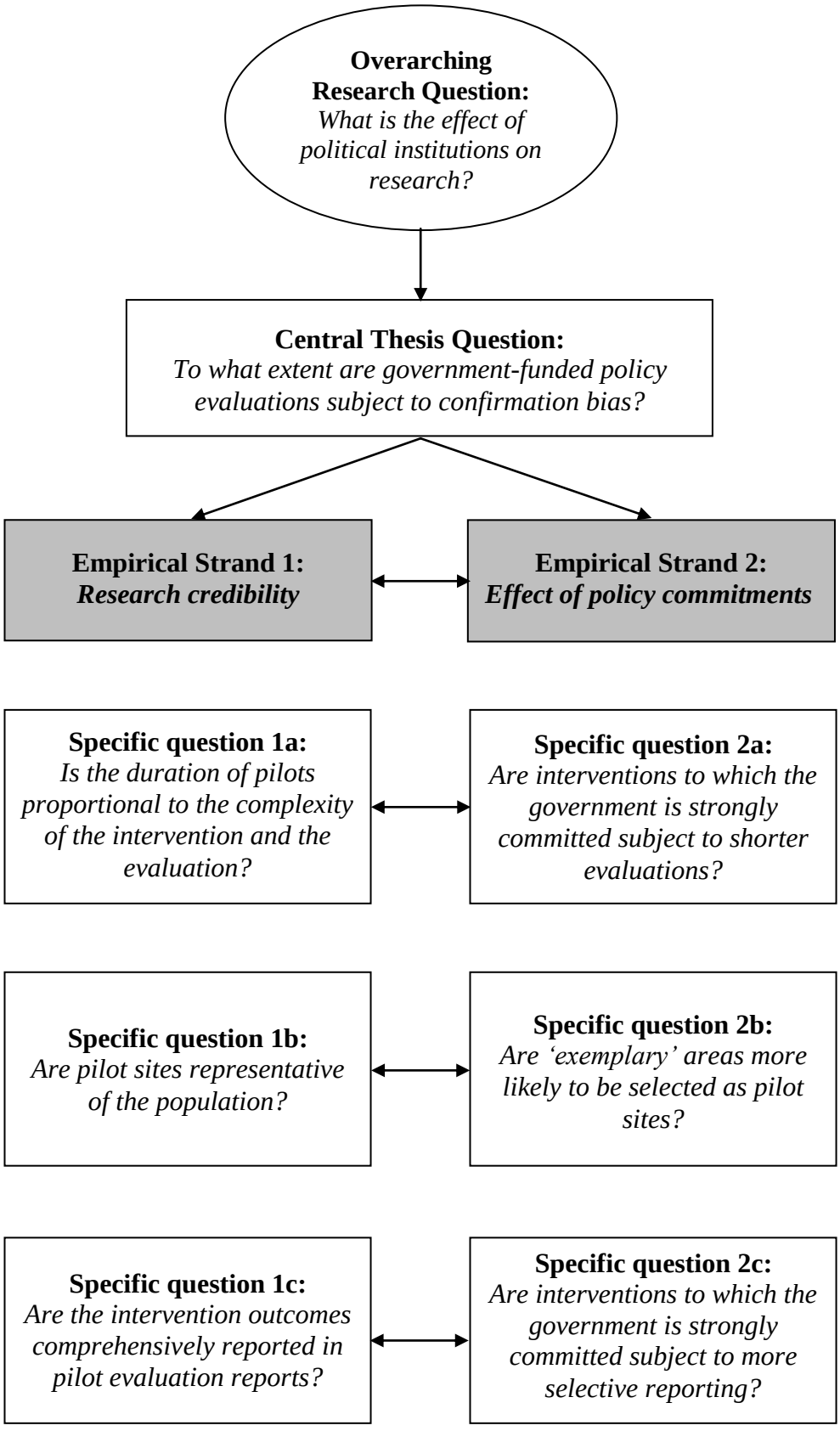
Previous studies have attempted to define and operationalise the notion of spin (Boutron, Dutton, Ravaud & Altman, 2010). Some of them have shown that studies sponsored by an organisation having a vested interest in the intervention were more spun than others (Bourgeois, Murthy, & Mandl, 2010). A recent study has shown that policy-makers could be at times guilty of 'leaning' on evaluators to influence the reporting of results (The LSE GV314 Group, 2014). However, it was based on a survey, which means that these accounts can be subject to desirability bias.

This thesis contributes to the literature by providing a qualitative content analysis of six evaluation reports.

Exhibit 1 – Expected effects of commitments on research decisions

← Commitment to the intervention	**Research decisions**	Commitment to the → scientific method
Shorter	**Duration**	Longer
Exemplary	**Sampling**	Representative
Selective	**Outcome reporting**	Complete

Exhibit 2 – Operationalisation



1.5. Relevance

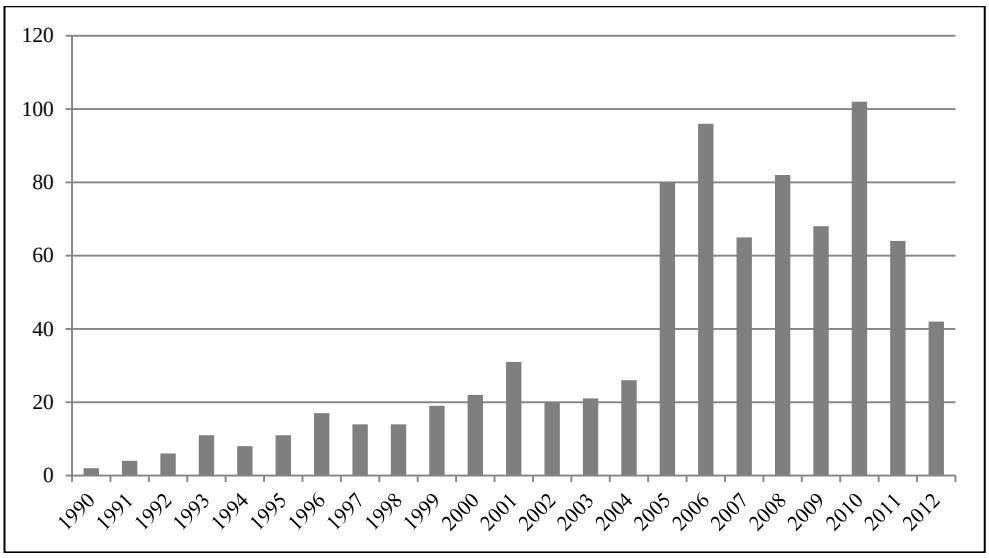
The findings of this thesis matter for several reasons. First, policy research absorbs an ever larger amount of public resources across the world. Yet, as discussed earlier, very few studies have looked at the prevalence and severity of confirmation bias in this research. Second, biased policy research decisions have social implications that need to be publicly discussed.

1.5.1. The growing influence of research in government: the British case

The social relevance of this thesis can be best understood in the light of the recent evolution of the role of science in policy-making. This role can be seen through a myriad of indicators. The following section focuses on the UK only.

The government's commitment to research between has been most obvious in the number of research outputs. Exhibit 3 shows the evolution of the number of policy studies published by the DWP between 1990 and 2012.

Exhibit 3 – Number of DWP research publications per year (N=825)*



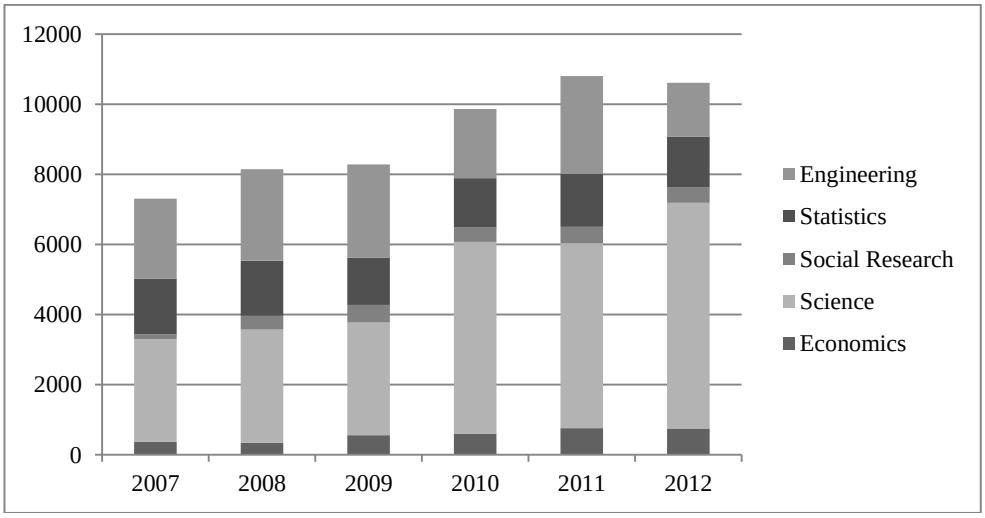
(*)Number of reports published on the DWP website under the series 'Research Reports'

This upward trend in output was made possible by increasing financial resources for research activities. It is very difficult to give accurate and comparable estimates of how much each government department in the UK spends on policy research. This is due to gaps between sums allocated to research and sums actually spent, with large variations due to accounting

and programme reasons. Furthermore, these budgets are typically not consolidated. For example, between 2005 and 2010, the Department for Education spent £11-12 million each year on core external social and economic research. In addition, the policy directorates have directed significant expenditure over the same period, around £15-20 million each year to policy evaluations, which is higher than the core research expenditure. Overall, the Department's spending on policy evaluation typically represents about 0,05% of total departmental spend" (*Science and Analysis Review of the Department for Children, Schools & Families (now Department for Education)*, 2010).

A final indicator of the influence of science in government is the number of staff belonging to a 'research profession' within the civil service. However, the definition of who belongs to this profession and who does not is not always consistent and thus, results vary from one source to another. According to the network of Heads of Science and Engineering Profession, there were about 12,000 specialist science or engineering posts across the Civil Service. According to the ONS Annual Civil Service Employment Survey, there are about 10,000 people who identify science and engineering as their primary profession. 3600 of them are members of the Government Science and Engineering (GSE) network (*The future of the Civil Service: Making the most of scientists and engineers in government*, 2013). Importantly, whichever indicator is used, the proportion of scientists and engineers within the Civil Service has grown between 2007 and 2012, due to both a growth in their recruitment and a drop in the total number of staff (see Exhibits 4 and 5).

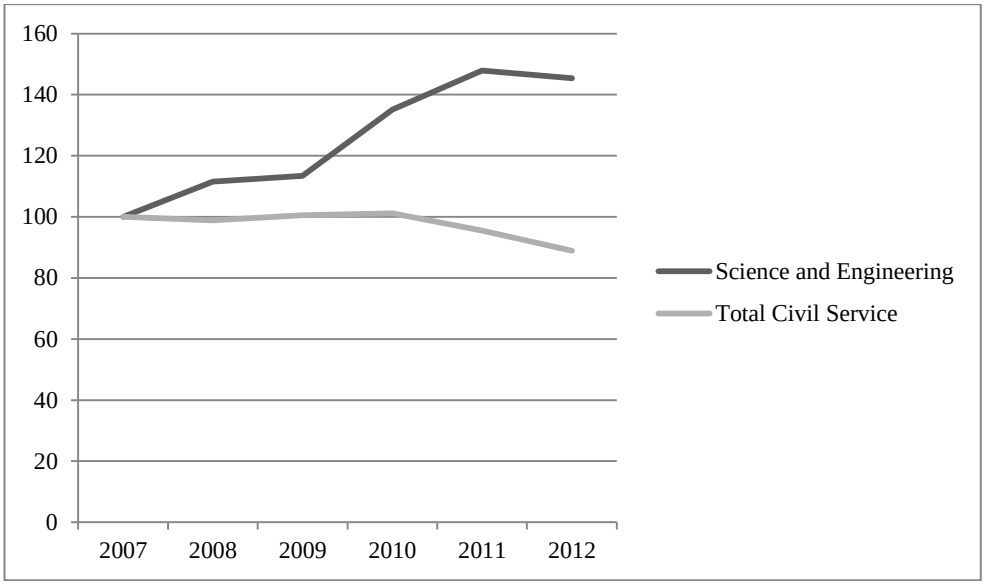
Exhibit 4 – Evolution of the number of the different scientific professions in the Civil Service



Source: ONS (Annual Civil Service Employment Survey)

Not only are scientists and engineers increasingly numerous within the Civil Service, they also tend to get more top jobs. In 1968, the Fulton Report on the Management of the Civil Service found that many scientists, engineers and other professional specialists were not given the responsibility or authority they deserved. The Committee therefore recommended that these specialists be given more policy-making and management opportunities, and training to equip them for their new work. Four decades later, that recommendation seems to have largely been taken on board. In 2011 for example, over half (56%) of the members of the GSE were in the grade/range HEO to grade 7, which is the highest in the civil service (*The future of the Civil Service: Making the most of scientists and engineers in government*, 2013). Furthermore, three of the four Cabinet Secretaries since 1998 have been trained economists, and the last two entered the civil service through the Government Economic Service.

Exhibit 5 – Evolution of the number of scientists and engineers in the Civil Service and of the total staff in the Civil Service (100 = year 2007)



Source: ONS (Annual Civil Service Employment Survey)

1.5.2. Trustworthiness of scientific results

The findings of this thesis have also important implications for the scientific profession.

On one level, confirmation bias leads researchers to make decisions that are sub-standard from a scientific viewpoint and thus are likely to be contested by the rest of the scientific community. On another level, confirmation bias will lead to overestimate the effect of the intervention and lift any doubt regarding its possible inefficacy. Confirmation bias has been found in very

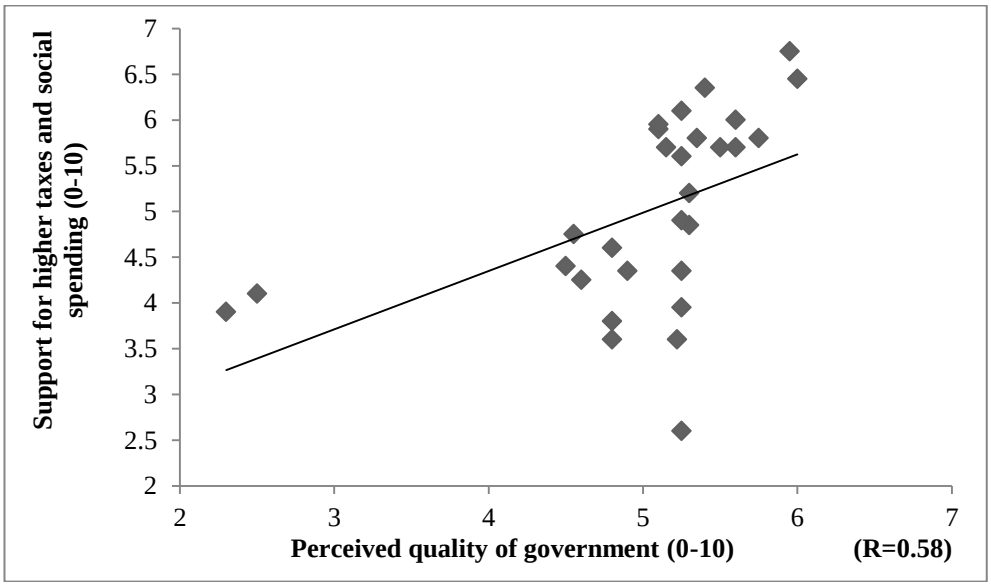
different disciplines and research areas, including in studies on the effects of nicotine (Turner & Spilich, 1997); antidepressants (Becker-Brüser W, 2010; Bruce Baker, Johnsrud, Crismon, Rosenheck, & Woods, 2003); and cell phone use (Huss, Egger, Hug, Huwiler-Müntener, & Röösli, 2007). Ultimately, confirmation bias might lead to poor policy decisions.

In the long run, biased policy research could damage the authority of science. This authority rests on two important aspects. First, it rests on the belief that science can provide true and useful accounts of the 'real world' (Bocking, 2004). Related to this view has been the notion that science is most authoritative when it speaks with unanimity (Bocking, 2004).

1.5.3. Legitimacy of public institutions

In addition to damaging the authority of scientists, biased policy research could also undermine the legitimacy of democratic governments. Evidence (presented in Exhibit 6) shows that there is a strong positive correlation between the perceived quality of government (i.e. the perceived fairness and efficiency of the implementing agencies) and attitudes to taxes and social spending (Svallfors, 2012). This result emerges in spite of the fact that countries with a high quality of government are also countries that already spend more on the welfare state than countries with lower quality of government (Rothstein, Samanni, & Teorell, 2012).

Exhibit 6 – Perceived quality of government and attitudes to social spending across European countries (1 unit = 1 country)



Source: European Social Survey Round 4, 2008

1.6. Thesis outline

The rest of this thesis is divided into three parts. Chapters 2 to 4 provide background, summing up what we have learned about the factors shaping research decisions and how similar questions have been addressed in the past. Chapters 5 to 7 are the empirical part of this thesis and present my findings using different approaches to the research question. Chapter 8 concludes by highlighting the contribution of this thesis and examining its broader implications for the research and policy-making communities.

Chapter 2 presents the theoretical framework of this thesis. It starts with a review of the literature on the effect of political institutions on policy research, emphasising the key contributions of philosophy, sociology, political science, research methods and of the professional literature. It will then offer an alternative approach using the confirmation bias theory and justify such an approach. The rest of this chapter is devoted to the operationalisation of this approach with a special focus on research design and variables.

Chapter 3 presents the institutional context of this study. I will first describe the decision-making process that led me to focus on the pilots and experiments conducted by the British government between May 1997 and May 2010 in four ministerial departments. I will then introduce the main actors, organisations and processes that underpin this study with a view to familiarise the reader. I will finally discuss the implications of the case study in relation to the research questions.

Chapter 4 approaches the research question from a methodological angle and presents the new PILOT dataset developed for the purpose of this study. I will first review how research questions similar to mine have been addressed in the past as well as the relative merits of these methodologies. The rest of the chapter will define the scope of the study, the sources of information used for the building of the dataset. It will also present the main variables available and the kind of questions the dataset can potentially answer.

Chapter 5 is the first in a series of three empirical chapters. It focuses on the duration of pilots. In this chapter, I will first analyse whether, on average, the duration of pilots is proportional to the scope of the research (Specific Question 1a). I will then examine whether the known commitment of the government to a reform is associated with shorter pilots, based on the data provided by PILOT (Specific Question 2a).

Chapter 6 looks at the selection of pilot sites in employment programmes. As in the previous chapter, I will first analyse whether all regions have the same chance of being selected as pilot site – which would be expected from a study claiming external validity (Specific Question 1b). I will then analyse

the influence of policy commitments on the selection of pilot sites (Specific Question 2b).

Chapter 7 is concerned with the reporting of evaluation outcomes. In this chapter, I will review the content of six evaluation reports systematically selected from the PILOT dataset. These six reports were selected with a view to maximise the contrast between policy interventions showing a strong/weak commitment. I will first assess the prevalence of outcome reporting bias (or 'spin') in these reports, using seven criteria developed by clinicians (Specific Question 1c). I will then examine possible 'associations' between the level of spin and the existence of a commitment to the policy (Specific Question 2c).

The conclusion (chapter 8) will develop three points. First, it will review how the evidence presented in earlier chapters supports my overall argument, and how this evidence fits with the rest of the literature. Second, it will take stock of the PILOT dataset and discuss the strengths and weakness of the research design used in this study. Thirdly, it will discuss the broader implications of my findings, both from an academic and policy point of view. A future research agenda will be proposed.

2. Theoretical framework

2.1. Introduction

This chapter investigates a little-understood question in political science: to what extent do political institutions influence policy evaluation?

A large literature has flourished in recent years to analyse how institutions affect each stage of the policy cycle, which commonly includes (1) agenda setting, (2) policy formulation, (3) implementation, and (4) evaluation. Thanks to these contributions, we have learned much about the reasons and the ways in which institutions produce sub-optimal policies or 'policy bias' by favouring one group or issue to the detriment of others (Ehrlich, 2011; Schattschneider, 1960). For instance, agenda-setting theories have shown how these institutions compete for turning private issues into public policy. Other theories have emphasised the key role played by these different institutions in the formulation of policy. We have also learned much about how institutions can affect the implementation of reforms.

Comparatively, our knowledge of what influences the way policy evaluations are conducted appears limited. Binder, Rhodes, Rockman and colleagues (2008) barely address the issue in *The Oxford Handbook of Political Institutions*. Other textbooks are equally succinct (Greenwood, Oliver, Suddaby, & Sahlin-Andersson, 2008; Lowndes & Roberts, 2013; Peters, 2011). Yet, policy evaluations raise important questions, mainly because they tend to be conducted or commissioned by the very organisations that designed and implemented the intervention in the first place. The assumption of independence, which underpins scientific research, is often violated.

Political institutions typically include formal democratic bodies (parliaments, governments, bureaucracies, political parties, presidents, etc.), however, institutional theories have also scrutinised the role groups and organisations without a constitutional mandate but nevertheless influential (interest groups, media, pollsters). The institutions I will be referring to in this chapter are essentially government departments and agencies (see chapter 3 for a more detailed justification and description).

I define policy evaluation as "the ex post assessment of the strengths and weaknesses of public programs and projects" (Bovens et al., 2008). The emphasis on ex post means that this chapter does not address the literature on ex ante analysis, where methods to evaluate policy alternatives are used as decision-making aids (Bovens et al., 2008; Dunn, 2004; Nagel, 2002). Policy evaluation is akin to research and development (R&D) in the social sphere; with the difference that policy research is non-proprietary and can

be conducted and replicated outside of government. Both policy evaluation and R&D differ from basic research, which is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particular application or use in view (OECD Glossary of Statistical Terms).

It should be said right away that political institutions influence research in two capacities. First, as regulators of scientific activities: throughout the 20th century, political institutions have increasingly influenced research notably through public funding, the sanctioning of research misconduct and the definition of the research agenda. Second, institutions influence research as clients of scientific organisations. This chapter focuses on this second aspect only.

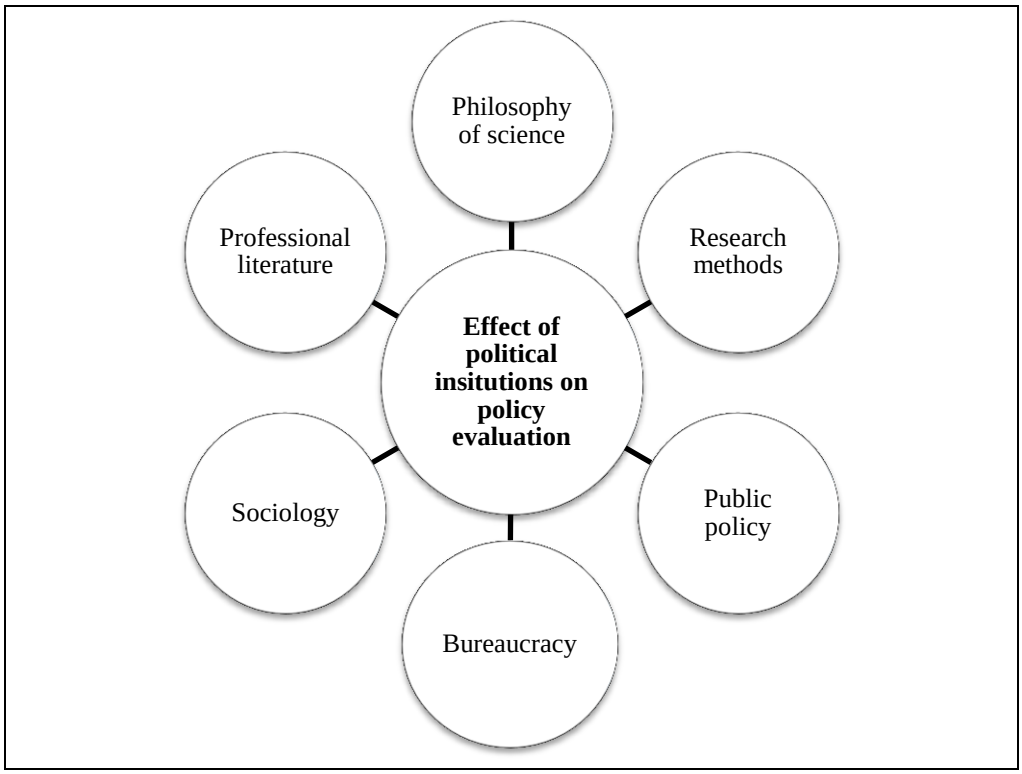
The goal of this chapter is to lay the foundations for an empirical research agenda assessing the effect of political institutions on policy evaluation. Three specific objectives have been assigned to it. First, this chapter reviews the theoretical and empirical literature on the interaction between science and institutions – the last review dating back from 1998 (Weiss, 1998). This review will in turn help me identify the type of issues that would make a significant contribution to this scholarship. The second objective of this chapter is to come up with a 'better' theory. The third objective is to operationalise the research question.

The rest of this chapter is structured as follows. Section 2.2 synthesises what we know about the effect of political institutions on research and identifies gaps in the literature. Section 2.3 offers an alternative approach to this scholarship based on the idea of 'confirmation bias' and briefly introduces this idea. Section 2.4 operationalises this approach. Section 2.5 considers different independent variables whereas section 2.6 focuses on dependent variables. Section 2.7 discusses the specificities of social research compared with other types of applied research such as clinical trials. Section 2.8 concludes.

2.2. Review of the literature

The literature on the effect of political institutions on policy research spans several disciplines, including philosophy, research methods, political science, sociology and diverse 'professional' literatures including education and nursing (see Exhibit 7). A systematic review of the literature would be a difficult exercise. Rather, the following section reviews the literature in a *narrative* fashion, focusing on what I considered to be the most significant contributions and highlighting points of consensus and disagreement.

Exhibit 7 – Map of the literature



2.2.1. Research without institutions

Assessing the effect of political institutions on policy research requires a thought experiment, namely the identification of the principles guiding research in a state of nature, or more realistically, in a context where scientists would work with minimum constraints. Other things remaining equal, any deviation from these principles occurring in an institutional context can be attributed to these very institutions.

The history and philosophy of science argue that these principles have been defined in two phases. Until the Enlightenment, science was primarily defined by its *purpose*, namely the advancement of knowledge. As such, it was virtually undistinguishable from philosophy. In a state of nature, research would be conducted by free individuals pursuing neither private gain nor political ideology, but simply the truth (Bocking, 2004). The advent of the 'scientific revolution' – between the Renaissance and the 18th century – has led to a redefinition of science based on its *methods*. The Oxford English Dictionary defines the scientific method as "a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses"[3]. The emphasis here is on the *procedure*: what makes a claim scientific is not its substance but the

[3] (Oxford English Dictionary, definition for 'scientific')

way the information was gathered, analysed and interpreted. Thus, theoretically, disagreements among scientists are not concerned with the relevance of the findings but with the credibility of the research process and the assumptions underpinning it (Bocking, 2004).

In addition to these 'technical prescriptions', science is based on a number of 'moral prescriptions', which are equally binding, not because they are procedurally efficient, but because they are believed right and good (Merton, 1942). These moral norms all relate to scientists' attitudes and behaviours in relation to each other and their research (Zuckerman, 1988). According to Merton, these norms include:

- *Communality* ("communism" in the original text), i.e. the common ownership of scientific results and methods and the consequent imperative to share both freely.
- *Universalism* specifies that scientific work and findings should be evaluated on the basis of "pre-established impersonal criteria: consonance with observation and with previously confirmed knowledge", and not on the personal, social or cultural attributes of the scientists involved.
- The principle of *organised scepticism* refers to the "detached scrutiny of beliefs in terms of empirical and logical criteria". This principle has implications for both producers and consumers of scientific findings: the former need to present their findings and methods transparently so that their value can be assessed, and the latter need to suspend judgement until they have examined findings and methods according to accepted standards and criteria.
- Finally, *disinterestedness* demands that scientists' work remain uncorrupted by self-interested motivations. It precludes the pursuit of science for the sake of riches, though Merton recognised the powerful influence of competition for scientific priority.

Thus, the assumption in much of the literature is that, in a state of nature, or in a state of minimum constraints, science would be independent and pursued for the sole purpose of human enlightenment. It would also scrupulously apply the scientific method and follow the moral norms of science. It is also the assumption that underpins the rest of this thesis.

2.2.2. Political institutions as 'consumers' of scientific advice

The fundamental difference between research undertaken in a state of nature and research undertaken in an institutional or professional context is that, in the latter situation, scientists are employed (or commissioned or compensated) and that their research is actually *utilised*.

The literature on 'research utilisation' has shed light on the three properties that make evidence a highly sought after resource (Beyer & Trice, 1982; Innvaer, Vist, & Trommald, 2002; Lavis, Robertson, Woodside, & et al., 2003; Weiss, 1980). The *conceptual* property of research is the closest to the purpose of science in a state of nature. It emphasises its capacity to enlighten individuals and organisations by articulating concepts and changing their understanding of natural and social phenomena. The *instrumental* property of research makes it capable of assisting decision-making by bringing answers to clearly predefined problems. Lastly, the *symbolic* property of research involves using research results and processes to legitimate and sustain pre-determined positions. This typology is now widely accepted in the literature and has been applied to many different policy areas including drug policy (Ritter & Lancaster, 2013), urban health (Murphy & Fafard, 2012) and education (Luke, 2011).

Which form of utilization is most prevalent is difficult to establish given the lack of commonly agreed indicators as well as the normative aspect of the question which might bias survey responses. Though very limited, the evidence would suggest that the conceptual use of research is more prevalent in the day-to-day professional activity of professionals and managers in government agencies than symbolic utilization, which, in turn, is more important than instrumental utilization (Amara, Ouimet, & Landry, 2004). Importantly, research shows that the three types of research utilization are not mutually exclusive and are in fact frequently combined (Amara et al., 2004).

The type of research utilization (or the type of combination) depends on a number of factors. One of them is the policy area. For example, Carpenter has shown that bureaucratic agencies of state were more involved in the provision and regulation of health policy than in other policy areas (Carpenter, 2012). The type of research utilisation depends also on the reputation that a given agency wants to enhance. Indeed, agency reputation shapes administrative decisions (Carpenter & Krause, 2012; Carpenter, 2001, 2010). The literature has also convincingly shown that policy-makers are more likely to make an instrumental use of scientific advice when there is a consensus among experts on a causal theory (Boehmer-Christiansen, 1994; Hood & Jones, 1996; Lavertu & Weimer, 2011).

Lastly, the literature has shed light on the aspects of research that can be used (Weiss, 1998). The findings are at the core of the instrumental type. In the conceptual type, utilization extends to the general ideas and generalizations from evaluation, even if they do not serve a specific and immediate purpose. In addition to the above, a policymaker using research symbolically can also take advantage of other parts of the research process. The sheer fact that an evaluation is being conducted can be used to demonstrate policy-makers' rationality and sound management (Feldman & March, 1981). Likewise, the definition of the scope of the research can be an indicator of symbolic use of evidence. Excluding inconvenient questions,

areas or stakeholders from a study can help producing 'congenial' results. There is some evidence that governments can 'play it safe' when they commission an evaluation (The LSE GV314 Group, 2014). Similarly, the design of the study and the choice of measures can also be the subject of political struggles among different agencies (Breslau, 1997; Weiss, 1998).

2.2.3. Research as a "shadow institution"

The public policy literature argues that research primarily benefits elected policy-makers (e.g. ministers). Thus, these policy-makers will seek to influence the course of research. This literature makes extensive references to 'politicians', 'political actors' and 'interest groups'. According to this scholarship, political actors tend to frame evidence in a way that supports their agenda. This is evident is statements such as "evaluation is the continuation of politics by other means" (Bovens et al., 2008).

Public policy scholars argue that democratic institutions provide incentives to successfully pass reforms. When the enactment of legislation or the implementation of a decision looks difficult, the authority of the government gets undermined, which in turn jeopardizes future reforms. The moral authority of science can, on occasions, facilitate reforms. Some have mentioned that in the US, pilot evaluations were used as "shadow institutions" used to legitimate contentious reforms (Brodkin & Kaufman, 2000; Rogers-Dillon, 2004). Thus, many have empirically sought associations between the degree of salience of an issue and the way research will be used to support policy. Some have argued that "on the small issues, evidence sometimes counts"; however, when it comes to the big issues, "politics is the order of the day" (Tonry, 2004). Heavily politicised policy areas are characterised by more *ad-hoc* or muddled-through policy-making (Lindblom, 1959). In such instances, there is intense media scrutiny of decision-making and prolonged conflict between competing interest groups and a permeating sense of crisis. Typical example would be drugs, where evidence is used symbolically (Monaghan, 2010) and schools (Henig, 2008, 2009).

Others have shown that political institutions reward politicians not so much for what they have achieved but rather for 'winning the argument'. A less demanding version of this theory suggests that policy-makers are driven less by the desire to get credit for what they have done than by the desire to avoid blame (Hood, 2011; Weaver, 1986). The ominous label of 'failure' or 'fiasco' that hovers over policies that failed to deliver entails a political statement (Bovens et al., 2008). Thus 'cherry-picked' information can be used as 'political ammunition' in the political debate.

Regardless of the initial motivation, political institutions will occasionally lead politicians to use research symbolically, or to 'frame' it in a way that suits their aims. "They will produce – or engage others to produce –

accounts of policy episodes that are, however subtly, framed and timed to convey certain ideas about what happened, why and how to judge this, and to obscure or downplay others" (Bovens et al., 2008). Methodologically, these accounts have mainly relied on individual case studies. Some of these case studies have considerably improved our understanding of the interaction between science and political institutions in areas such as drugs (Monaghan, 2010), education (Henig, 2008, 2009). There have been occasional reports of ministers stepping in the middle of a research project and 'leaning on' researchers (The LSE GV314 Group, 2014).

2.2.4. Research and bureaucracy

Public administration specialists diverge from their public policy colleagues in four important ways.

First, they object the idea of a direct interest of elected policy-makers in research. According to them, elected policy-makers frame public policy, but do not carry it out. Their hypothesis is that the effect of political institutions on research is in fact mediated by government agencies and their employees. This is supported by numerous accounts of the role of expertise in the development of national agencies including the Forest Service, the Department of Agriculture (Bocking, 2004) and the FDA (Carpenter, 2001, 2002, 2010). In this context, civil servants sometimes invoke ministers to deflect blame. Some evaluators reported that opposition to evaluation would typically come from civil servants, even though they might have pretended there was opposition from ministers (Ettelt & Mays, 2013).

Second, public administration scholars contend that research is not used with a view to secure a reform or for argumentative purposes, but to enhance the power and the legitimacy of the agency. As demonstrated by Weber, bureaucracies assert power through specialised expertise and control of information, justified by their claim to be the only means by which the complexities of modern society can be managed (Bocking, 2004; Weber, 1946). According to Carpenter, the technical-scientific reputation of an agency is one of four reputational strategies used by public administrators to face the challenges of modern governance (Carpenter & Krause, 2012). The result has been an ever-expanding application of administrative rationalism: seeking, with the guidance of technical expertise, rational and efficient solutions to the problems of society (Bocking, 2004).

Third, the public administration literature is more specific when it comes to defining the effect of political institutions. At a 'macro' level, reputation gives agencies more autonomy, in the sense of being able to sway the wishes of elected officials on particular matters of policy and to secure deference from these elected officials (Carpenter, 2001; MacDonald & Franko, 2007; MacDonald, 2010). Political institutions also impact the research process in very specific ways. For example, several studies have

shown the timing of new drug approval by the U.S. Food and Drug Administration was influenced by the type of reputation of the agency (Carpenter, 2002, 2004) as well as the relative importance of the drug, measured in terms of therapeutic novelty and expected return on investment (Dranove & Meltzer, 1994). Political institutions also influence the methodologies used in programme evaluations (Breslau, 1997).

The fourth difference of the public administration literature is on methods. Whereas interviews, survey and desk reviews dominate the public policy literature, a more systematic approach based on administrative data has been used in public administration (Carpenter, 2001, 2002, 2004; Krause & Douglas, 2005; MacDonald, 2010). This is a key difference, given that only a systematic approach can reliably assess the long-term effect of political institutions on policy research.

2.2.5. The norms of science

It would be incomplete to review the literature on the effect of institutions on research without mentioning the response of the scientific community to the uses and misuses of science. Sociologists of science and sociologists of professions contend that the professionalization of science has led to the creation of scientific institutions. These institutions include universities, academic journals and professional societies, which all play an important role in diffusing and enforcing scientific norms.

The 'optimistic' view is that these institutions play a key role in repressing research misconduct and questionable research practices (Steneck, 2003). Research is a professional activity. As such it is subject to norms, i.e. prescriptions commonly known and used by practitioners (Andersen, 2007; Ostrom, 1986). These prescriptions refer to which actions are required, prohibited or permitted in specific situation. The existence of such norms is a vital part being a profession. Their enforcement depends on the provision of incentives, which are reflected in the criteria used to appoint, evaluate, and promote individual faculty members. Today, the rewards of a successful academic career typically include the personal gratification derived from scholarship and discovery, recognition by peers, and academic promotion and tenure, as well as enhanced responsibility and outside financial opportunities.

A more pessimistic view is that these institutions defend the vested interests of researchers. According to sociology of scientific knowledge, science is neither exceptional nor immune from the forces that affect other human activities. Mitroff (1974) for example, showed that for each of the Mertonian norms there exists counternorms that play equally important roles in the practice of science. Social and historic studies demonstrate science to be an enterprise consisting of individuals who passionately engage in value-laden activities to demonstrate their correctness and depend

upon the socio-cultural context from which their work emanates (Hull, 1988; Pinch, 1986; Sapp, 1990). The practice of science also includes fraudulent activity, sometimes involving even mainstream scientists (Sapp, 1990). Eventually, sociologists of science concluded that Merton's case cannot be convincingly made and that his norms might be better viewed as an ideology of science (Mulkay, 1976). The studies of scientific practice cited above corroborate Mulkay's assertion that the Mertonian norms are an ideology that serves the interests of the scientific community in at least three ways. The norms (1) enhance the epistemic status of scientific knowledge; (2) increase the political power of scientists; and (3) elevate the social status of scientists. Functionally, they work at the interface between the scientific community and the general population and provide justification for the continued support of science in society.

2.3. Confirmation bias

Despite the above-mentioned merits, the current theoretical framework available to explain the effect of institutions on policy research is too fragmented to allow progress. There is scope for a more parsimonious and 'universal' theory.

A promising way of bridging the above-mentioned gap can be found in two related literatures. The research methods literature has approached the question as an example of 'experimenter's bias'. The experimenter's bias – also known as research bias – has been defined as "a subjective bias towards a result expected by the human experimenter" (Sackett, 1979). The social psychology literature has developed the related concept of 'confirmation bias' (also called confirmatory bias or 'myside' bias) which is attributed to English psychologist Peter Wason (Gale & Ball, 2002) and describes the tendency of people to favour information that confirms prior beliefs or hypotheses, regardless of whether the information is true (Plous, 1993). The following section outlines the causes and the consequences of confirmation bias, which is the term I will use throughout this thesis. The variables mediating this effect are also presented.

2.3.1. Causes

Confirmation bias is often described as a result of automatic, unintentional strategies rather than deliberate deception (Hergovich, Schott, & Burger, 2010; Oswald & Grosjean, 2004). It results from the supposed inability of a human being to be objective, and more specifically from (1) the desire to appear consistent and/or to fulfil public commitments; and (2) reciprocation.

A first cause of bias is commitment. A commitment is a public engagement or obligation to take a specific course of action. Its normative power is such

that individuals will often stick to the original deal even though it has changed for the worse. The reason people stick to their commitment is that they want to maintain a positive self-image. People strengthen their original commitment by the addition of supportive new thoughts and feelings (Cialdini, Cacioppo, Bassett, & Miller, 1978). This is particularly relevant in political contexts, where reasoning can be subconsciously biased, favouring conclusions that governments have already committed to. A two-decade study of political pundits by Tetlock found that, on the whole, their predictions were not much better than chance. Tetlock divided experts into 'foxes' who maintained multiple hypotheses, and 'hedgehogs' who were more dogmatic. In general, the hedgehogs were much less accurate. Tetlock blamed their failure on confirmation bias – specifically, their inability to make use of new information that contradicted their existing theories (Tetlock, 2005). Evidence of confirmation bias has also been found in scientific decisions (Hergovich et al., 2010; Koehler, 1993; Mahoney, 1977; Nickerson, 1998).

Another cause of bias is reciprocity. As a social construct, reciprocity means that in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model. Conversely, in response to hostile actions they are frequently much more nasty and even brutal (Cialdini, 2003; Fehr & Gächter, 2000). Reciprocity is so strong that people tend to reciprocate regardless of whether they like the person who originally gave the favour and even if they did not want the favour, as was demonstrated in Regan's experiment (Regan, 1971). Reciprocation can be genuine and unconscious (Cialdini, 2003). The problem is a growing concern in the medical research community, where a vast majority of pharmaceutical companies resort to 'friendly actions'. Those include free drug samples, meals, continued medical education, financial incentives to participate in clinical trials, honoraria for delivering lectures, leisure trips, expensive text books and items of low monetary value such as pens and notepads. Reciprocity can extend to any action perceived as a 'favour' such as a job offer, a promotion, a bonus, professional honours and the sponsorship of research projects and scientific meetings (Institute of Medicine, 2009).

There is no direct evidence of reciprocation; however there is strong evidence that scientists attitude towards industry-funded research becomes more positive as the amount of interactions between the two spheres increases (Austad, Avorn, & Kesselheim, 2011). A review of 17 surveys on the attitudes of researchers to financial ties in research revealed that investigators are concerned about the impact of financial ties on choice of research topic, research conduct and publication, but this concern is less among investigators already involved with industry. Researchers approve of industry collaboration and financial ties when the ties are indirectly related to the research, disclosure is up front, and results and ideas are freely publicized. However, their trust in disclosure as a way to manage conflicts

may reveal a lack of awareness of the actual impact of financial incentives on themselves and other researchers (Glaser & Bero, 2005).

2.3.2. Consequences

Confirmation bias leads to unconventional or sub-standard decisions. In a research context, this means decisions – and thus, potentially, findings – that are likely to be controversial within the scientific community and portrayed as not credible (Bocking, 2004). Opposing interests will highlight uncertainties in the evidence, discrepancies and ambiguities in the interpretation, ties between researchers and business or political interests, and any other technical aspects of the problem that can provide an opportunity to question the credibility of the research process. In other words, confirmation bias can damage scientific credibility, understood as the extent to which science in general is recognized as a source of reliable information about the world (Bocking, 2004). Empirically, confirmation bias has been found to affect the way we process information, report results and interpret findings.

First, confirmation bias impairs the way we process information. Experiments have found repeatedly that people tend to test hypotheses in a one-sided way, by searching for evidence consistent with their current hypothesis (Kunda, 1999; Nickerson, 1998). More specifically, confirmation bias has been invoked to explain 'illusory correlations', which is the tendency to see non-existent correlations in a set of data (Fine, 2006). For example, a study recorded the symptoms experienced by arthritic patients, along with weather conditions over a 15-month period. Nearly all the patients reported that their pains were correlated with weather conditions, although the real correlation was zero (Redelmeier & Tversky, 1996).

Second, confirmation bias skews analyses towards an outcome that is favourable to the experimenter. The most striking case of confirmation bias is when two opposing experimenters find themselves at odds with the published findings of research they sponsor. But the Experimenter's bias is not always that spectacular. Often, it will lead to overestimate the effect of the intervention and lift any doubt regarding its possible inefficacy. Experimenter's bias has been found in very different disciplines and research areas, including in studies on the effects of nicotine (C. Turner & Spilich, 1997); antidepressants (Becker-Brüser W, 2010; Bruce Baker et al., 2003).

Confirmation biases are not limited to the search and collection of evidence. Even when two individuals are given the same information, the way they interpret it can be biased. This has been recently demonstrated in a study on the neural responses of 30 committed partisans during the U.S. Presidential election of 2004. The authors presented subjects with reasoning tasks

involving judgments about information threatening to their own candidate, the opposing candidate, or neutral control targets (Westen, Blagov, Harenski, Kilts, & Hamann, 2006).

2.3.3. Mediating variables

Research shows that the strength of confirmation bias depends on the issue being considered, but not on personal differences such as cognitive abilities.

First, confirmation bias is stronger for emotionally charged issues and for deeply entrenched beliefs. This was exemplified in the *Stanford Biased Interpretation Experiment* in which participants with strong opinions about the death penalty read about mixed experimental evidence. Twenty-three percent of the participants reported that their views had become more extreme, and this self-reported shift correlated strongly with their initial attitudes (Lord, Ross, & Lepper, 1979). More recently, Taber and Lodge conducted a similar study using the emotionally charged topics of gun control and affirmative action. They measured the attitudes of their participants towards these issues before and after reading arguments on each side of the debate. Two groups of participants showed attitude polarization: those with strong prior opinions and those who were politically knowledgeable (Taber & Lodge, 2006).

Second, individual characteristics do not seem to have an effect on the severity of confirmation bias as previously thought. Empirical research has consistently shown that confirmation bias is persistent, regardless of intelligence level. In two experiments involving a total of over 1400 university students and eight different comparisons, the authors found very little evidence that participants of higher cognitive ability displayed less confirmation bias (Stanovich & West, 2007). There is moderate correlations between cognitive ability and the ability to avoid such biases (Gilinsky & Judd, 1994; Handley, Capon, Beveridge, Dennis, & Evans, 2004; Kokis, Macpherson, Toplak, West, & Stanovich, 2002).

2.3.4. Confirmation bias in policy research

The above shows that the question of the effect of institutions on policy research would be more effectively addressed by analysing the prevalence and severity of confirmation bias in policy research. This entails two interesting questions. The first question is normative and relates to the ideal of research. It could be formulated as follows: *To what extent is institutional policy research scientifically credible?* The second question is positivist and relates to the effect of confirmation bias: *Are evaluations of interventions to which institutions are committed less credible?*

Framing the problem in such a way would provide a number of benefits. First, confirmation bias offers a more parsimonious explanation of the effect of institutions on policy research than existing theories. It simply implies that commitments and reciprocation lead to substandard or unconventional research decisions. Second, confirmation bias works with a broad range of institutions (scientific institutions, government institutions, political institutions, private-sector companies, etc.) and policy areas. Third, as will become evident in the rest of this thesis, it allows more accurate explanations and predictions than the idea of 'research utilisation' which many authors have struggled to operationalise.

Conveniently, the concept of confirmation bias sits comfortably with existing political science theories. First, confirmation bias leads to use research *symbolically*, i.e. for confirming the idea that the experimenter wishes to promote (Beyer & Trice, 1982; Knorr, 1977; Lavis et al., 2003; Pelz, 1978; Weiss, 1979). However, the concept of confirmation bias is more complete than that of research utilisation, since it offers an entire causal theory regarding the effect of political institutions on policy research. Second, confirmation bias can be seen as a type of agency cost. Governments evaluate policies on behalf of the citizenry (the principal). However, because the two parties have different interests and the government has more information (policy and research expertise), citizens cannot directly ensure that their agent is always acting their best interests (Fama, 1980). Third, confirmation bias is motivated by the desire to avoid blame (Hood, 2011). Were governments not sanctioned for their performance in office (in terms of curbing crime, unemployment, illiteracy, etc.), it can be argued that 'inconvenient' evaluation findings would not be a problem. Likewise, confirmation bias can be seen as the expected behaviour of organisations seeking to enhance their performative reputation (Carpenter, 2010).

2.4. Policy commitment and confirmation bias

Estimating the risk of confirmation bias can be done in different ways. The following section first describes what the sponsorship of a study can tell us about the objectivity of the researcher. It argues that the strength of the government's commitment to the intervention might be a better option.

2.4.1. Specifications for a correlational study

Identifying causal mechanisms – provided they exist – is methodologically challenging as it requires an experimental setup. In an ideal experiment, a sample of researchers would be selected from the population and a fraction of this group – randomly selected – would be placed in political institutions. Conversely, the rest of the group would conduct the same research

independently. Because the two groups are comparable by virtue of the random assignment, any systematic difference in the ways both sub-groups make policy research decisions could be attributable to political institutions. No significant difference between the incentivised group and the non-incentivised one would confirm the dominance of the professional logic, whereas a significant difference would disprove it. Whereas controlled experiments are hardly ever possible, comparable circumstances sometimes occur naturally. This is the case for example when very similar research projects are carried out by different teams, one working in conditions close to those provided by the treatment, the other not.

The second best design implies thus to observe and record the partial effect of political institutions on research decisions as they appear to the researcher. In the absence of formal 'treatment' – the presence of political institutions cannot be contrasted with their absence – I am left with comparing situations where the effect of the performative logic is relatively stronger or weaker. Such variations have in the past provided an interesting setting for the study of the relations between democratic institutions and the enforcement of air pollution legislation in the US (Wood, 1988).

2.4.2. Study sponsorship

The first strategy consists in identifying the sponsor of the study. Confirmation bias estimated in this way is better known in the literature as 'funding bias' or 'sponsorship bias', however these terms are synonymous (Lexchin, 2012). Empirical studies of funding bias have mainly been undertaken in the area of biomedical research, where drug manufacturers, regulators and patient groups often perform similar studies.

The results of this research are rather unambiguous and consistent: research findings are influenced by the logic of the sponsoring organisation. In one study, for example, researchers looked into every trial of psychiatric drugs in four academic journals over a ten-year period, finding 542 trial outcomes in total. Industry sponsors got favourable outcomes for their own drugs 78% of the time, while independently-funded trials only gave a positive result in 48% of cases. Competing drugs put up against the sponsor's drug in a trial were more effective only in 28% of cases (Kelly, Cohen, Semple, & et al., 2006).

The underlying assumption is that studies funded by organisations that do not have a vested interest in the outcome of the trial show a lower risk of confirmation bias. The credibility of this assumption rests on the idea that regulators and patient groups have no vested interest in the drug, which is highly questionable. Government funding can result in bias if the aim is to minimise the cost of therapy that it pays for. Likewise, patient organisations may want what they see as the newest and best medications made available to their membership (Lexchin, 2012).

In the case of policy research, this design would involve comparing similar studies, some sponsored by governmental organisations, others sponsored by non-governmental organisations. This is possible provided two conditions are met. The first condition is that there are enough studies sponsored by non-governmental organisations to warrant statistical power. This is possible in medical research, which is a highly regulated market – and thus subjected to multiple controls and investigations (from regulators and patient groups). A similar design might prove difficult with social interventions for the opposite reason: policy evaluations are rarely conducted outside government. The second condition is that the studies funded by governmental and non-governmental organisations be reasonably comparable, not only in terms of the intervention, but also in terms of scope, design, timing, etc. This could also be challenging.

2.4.3. Commitment to the intervention

The second strategy consists in contrasting interventions to which the experimenter is strongly committed with interventions to which the experimenter is weakly committed. In political economy, the notion of commitment has mainly been applied in relation to central banks and monetary policy, where predictability and stability are key performance indicators (Nakazono & Ueda, 2013). A number of empirical studies exist regarding the effects of monetary policy commitment (Baba et al., 2005). The notion has been more rarely applied to the executive branch. Uses have been limited so far to the notion of compliance to international commitments (Kelley, 2007) and to issues of fiscal policy. Yet, the inability of governments to tolerate an open outcome and accept genuine uncertainty as stipulated in the idea of experimentation resonates with earlier observations by Campbell (1969) who had noted that governments tend to commit to policy politically and thus find it difficult to be seen at fault.

Going back to the area of medical research, a drug manufacturer could be committed to a drug because it represents a radical new breakthrough in treatment. A drug manufacturer is also more likely to be committed to drugs that are expected to generate high economic returns. Thus, confirmation bias could be estimated by comparing the effect of drugs with different levels of FDA ranking of therapeutic novelty or drugs with different sales prospects (Dranove & Meltzer, 1994). Likewise, FDA review times were found to be decreasing in (a) the wealth of the richest organisation representing the disease treated by the drug; and (b) media coverage given to this disease. These results suggest that ‘political influence over drug approval operates primarily through ‘salience signals’ transmitted by groups and media (Carpenter, 2002).

In the policy context, the financial cost of an intervention is certainly a factor, however it can be difficult to get the information in a reliable and consistent way. The cost can be measured in terms of political capital.

Reforms to which the government is strongly committed would be subject to a higher risk of bias than reforms to which the government is weakly committed.

2.5. Independent variables

The rest of this chapter is dedicated to the operationalization of the research question. I will start with the independent variable, namely the government's commitment to a given reform. Government is largely a black box, which means that the strength of policy commitments can only be *estimated*. The following section offers three possible strategies.

2.5.1. Pilots and phased introductions

The first strategy consists in comparing the decisions made in two different research contexts. The first context is that of a *policy pilot*, i.e. an intervention trialled for a limited period on a fraction of the territory on which it is meant to be rolled out. In principle, the probability that a pilot will be rolled out nationally is unknown at the time of its launch and contingent on the results of the evaluation. The second context is that of the *phased introduction of a reform*, i.e. a reform for which the probability to be fully implemented is known (and close to 1), but for which setup is similar to that of a pilot. Like pilots, phased introductions – which are known in the UK as *pathfinder pilots* – are evaluated on a small scale and over a limited period. They can be evaluated using the exact same designs and methodology. The only difference is in the government's intention and this intention is usually clearly stated.

The extent to which formal pilots and pathfinder pilots are strictly comparable has been debated. Some have mentioned that the term 'pilot' encompassed vastly different projects (Ettelt & Mays, 2013; Jowell, 2003). According to Ettelt and Mays, pilots can be used for experimentation, for early implementation, for demonstration and for learning how to operationalise a policy (Ettelt & Mays, 2013). I would argue that these categories are difficult to apply and not mutually exclusive. Indeed, many of the pilots launched by the Labour government tested new ways of delivering social services. It was assumed that the effectiveness of social policy was hampered by inefficient agencies and organisations. In other words, these were public service reforms as much as they were social policy reforms. Furthermore, these categories indistinctly apply to both formal and pathfinder pilots and can be controlled for.

2.5.2. Manifesto pledges

Comparing pilots and phased introductions is a compelling way of estimating policy commitment; however it is not the most frequent type of policy-making. Another, more widely applicable, way of assessing a government's commitment to a specific reform is to check whether it was announced in the ruling party's manifesto for the previous election (Rose, 1980). Manifestos may not be something voters care about, however the media have specialised in this activity. The multiplication of 'pledge trackers' (such as *The Guardian*'s) shows how crucial it is for a party to implement its pledges once in government. Against this background, it is unlikely that an office-seeking government will commit large resources to pilot a measure that contributed to its electoral success.

Previous research on electoral pledges finds that politicians fulfil most of their electoral promises when they are in power. Pomper and Lederman (1980) find that from 1944–1976, 79% of the pledges proposed by the winning party in the US were fulfilled. Rallings (1987) concludes that 64 of the British pledges from 1945 to 1979 were implemented. Royed (1996) studied British and US electoral pledges during the 1980s. She finds that the British Conservative party implemented more than 80% of its electoral pledges while in government. She also studied British parties in opposition and finds that they obtain much lower fulfilment rates. Only 15% and 32% of the pledges proposed by the Labour party in Britain in 1979 and 1983, respectively, were fulfilled. Royed (1996) also studied the US case during the 1980 and 1984 electoral cycles and found that even though the Democrats had a majority of seats in at least one of the houses during these years, the Reagan administration was able to act upon 60% of its electoral pledges. See also Artes (2013) and Chaney (2013).

In looking at the effect that parties have on policy, manifestos offer a good prediction of what parties will do when in office. This claim is supported by Klingemann, Hofferbert and Budge (1994), who find that government party programmes are remarkably well reflected in post-election priorities, measured as percentages of central government spending in major areas, that is to say that their expenditure reflects the differential issue saliency written into their party manifestos. Based on data from 1970-1979, Rose found that contrary to popular belief governments implemented a large proportion of their manifesto pledges, noting that Labour governments 'acted upon' 55% of their manifesto pledges whilst Conservative governments 'acted upon' 80% of theirs.

2.5.3. Seniority of the 'reform champion'

The problem with manifesto pledges is that they are better suited for studies comparing a large number of heterogeneous policies.

An alternative consists in identifying the 'champion' of the reform. Policy reforms are often introduced by a member of the government; and the seniority of the endorser can be taken as an indication of the salience of the reform. The announcement of a reform can be seen as a delegation issue, whereby each principal, from the Prime minister to the mid-level bureaucrat can decide whether to be the 'manager', taking direct responsibility for the outcome, or the 'chair of the board' overlooking operations (Hood, 2011). Given politicians' propensity to avoid blame even when that implies not getting credit (Hood, 2011; Weaver, 1986), a reform announced by the Prime minister can be considered more salient than a reform announced by a senior minister. Likewise, a reform announced by a senior minister is considered more salient than a reform introduced by a junior minister. It follows from this that the reforms for which no public announcement is made (which might occur when the reform can be implemented through secondary legislation or statutes) are the least salient.

Using the champion of the reform as independent variable has an additional benefit: it allows more contrast and thus greater measurement validity than dichotomous variables on phased introductions and manifesto pledges. Indeed, there is no reason to believe that such a commitment be so clear-cut. Ordinal variables measuring government policy preferences have been used in studies analysing governments' responsiveness to public opinion (Hobolt & Klemmemsen, 2005).

2.6. Dependent variables

The following section is concerned with the research decisions that are most likely to reflect confirmation bias. According to Sackett (1979), confirmation bias can occur in any one of seven stages of the research cycle, from the formulation of the research question to the reporting of findings. Three of these stages are reviewed below as they offer an interesting window for the study of confirmation bias.

2.6.1. Research duration

The first such window is the duration of the research project. Time is a precious resource for both the researcher and the policymaker, but for opposite reasons.

On the one hand, researchers committed to scientific norms will often push for longer research projects. First, repeated measurements are recommended to reduce the statistical phenomenon of regression to the mean which happens when unusually large or small measurements are followed by measurements that are closer to the mean (Barnett, Van der Pols, & Dobson, 2005; Stiegler, 1997). Second, the psychological literature shows that

individuals display different behaviours to novel and usual signals, for example in an experimental context (Gillespie, 1991). Thus an individual's response to a new policy is likely to be different as he gets used to it. Third, setting up a research project often takes time; especially when it involves the training of the policy implementers. Researchers eager to generate high-quality evidence are thus more likely to report on the long-term effect of the treatment.

On the other hand, a commitment to the intervention will lead to speedier research. Waiting is politically costly. People suffering from unemployment or crime want a solution to their problems, and incentivised policy-makers want to deliver it as early as possible. There is ample evidence that politicians and business leaders are hard-pressed to deliver before the next election or generate a rapid return on investment. Research within the finance and accounting literatures finds that managers do sacrifice (at least some) long-term investments in response to pressure from the capital markets (Graham et al's (2005)). Similarly Bartov (1993), Bushee (1998), Dechow & Sloan (1991), and Penman & Zjang (2002), all report evidence consistent with the idea that managers sell assets, cut R&D or reduce earnings to meet earnings targets. More recently Benner (2007, 2010) has suggested that firms going through significant technological transitions face particularly intense pressure, causing them to reduce capital investment and investment in R&D (Repenning & Henderson, 2010).

In light of the above, I argue that, other things being equal, shorter research projects denote confirmation bias. The literature on industry-sponsored clinical trials provides evidence to support this claim (Carpenter, 2002; Dranove & Meltzer, 1994; Olson, 1997). A 2010 review compared around a hundred truncated clinical trials and four hundred matched trials that ran their natural course to the end: the truncated trials reported much bigger benefits, overstating the usefulness of the treatments they were testing by about a quarter (Bassler, Briel, Montori, & et al., 2010) (Montori, Devereaux, Adhikari, & et al., 2005), (Trotta, Apolone, Garattini S, & Tafuri, 2008). Evidence from the policy area is thinner but highlights a similar phenomenon. Anecdotes and interviews have concurred to stress that the greatest source of incompatibility between research and policy rested on the conflict between their respective cycles (Boa, Johnson, & King, 2010; Coleman, 1979; Hallsworth, Parker, & Rutter, 2011; Jowell, 2003).

2.6.2. Sampling decisions

Another important decision likely to be affected by confirmation bias is the sampling of the units who will be part of the study.

From a scientific viewpoint, this decision is dominated by the need to have a sample that is as representative of the population at large as possible. A study which conclusions hold over variations in persons, settings, treatments

and outcomes is said to have *external validity*. The method most often recommended for achieving this close fit is the use of formal probability sampling (Rossi, Wright, & Anderson, 1983; Shadish, Cook, & Campbell, 2002). Regardless of the method used, sampling is usually seen as a difficult decision to make, with uncertain results. Using the examples of the *Bangladesh Integrated Nutrition Project* and the *California Class-Size Reduction Program* which both failed in replicating effective interventions evaluated with RCTs, (Cartwright & Hardie, 2012) have argued that experiments could not alone support the expectation that a policy will work outside the testbed, given the importance of logistical and contextual factors in the success of a social policy. Whether or not a 'sufficient' level of external validity can be achieved, it is safe to say that researchers committed to the scientific logic will choose their samples in a way that guarantees generalizability.

A sample put together with a view to favour the intervention is expected to be *exemplary* rather than random, heterogeneous or typical. The first reason has to do with the fact that research in a political or market context is skewed towards application. Thus, when the research is carried out *ex post*, it can be tempting for the principal to focus the evaluation on the individuals or groups who seem to have better responded to the intervention. When the research is carried out *ex ante*, the principal may also have an interest in testing the intervention on atypical individuals or groups, for example with a view to increase the probability of generating flattering results. The other reason is basic risk aversion. Research has shown that politicians are motivated primarily by the desire to avoid blame rather than by seeking to claim credit for their decisions (Weaver, 1986).

Against this background, I claim that, other things being equal, the representativeness of research samples can be seen as a test for the relation between the researcher and the policy-maker. Here again, the literature on clinical trials suggests that this claim is not unfounded. There is a fairly large – and growing – number of studies pointing to the flimsiness of medical trials' external validity (Keitner et al., 2003; Pratt & Moyé, 1995; Rothwell, 2005; van Staa, Leufkens, Zhang B, & et al., 2009; Zimmerman, Chelminski, & Posternak, 2004). For example, one such study took 179 representative asthma patients from the general population and looked at how many would have been eligible to participate in a selection of asthma treatment trials (Travers, Marsh, Williams, & et al., 2007). The answer was 6% on average. Flimsy external validity means that a trial is irrelevant to real-world populations.

2.6.3. Outcome reporting

The third decision that is most likely to reflect confirmation bias is the reporting of evaluation outcomes.

A researcher committed to scientific norms is expected to report findings *in full*, according to pre-specified research questions, theories and variables. Specifying the method from the outset of the research process means that outcomes cannot be manipulated, for example, in order to present positive outcomes. Therefore, provided they apply similar methods, different researchers are likely to report the same results, whether these results are positive, negative or null. The recent years have witnessed the multiplication of initiatives meant to standardise reporting such as CONSORT or COMET.

Conversely, a researcher committed to the intervention is expected to report outcomes *selectively*. Research findings are anything but neutral. In highly regulated industries such as pharmaceutics, an inconclusive trial means that a new drug will not be approved by the regulator. The medical literature highlights a number of recurrent strategies to present these findings in accordance with the interest of the principal. One of them consists in measuring uninformative surrogate outcomes (such as blood pressure or cholesterol rather than the prevalence of specific events such as heart attack or death) or in changing the outcome once the trial is finished (Chan, Hróbjartsson, Haahr, & et al., 2004; Jureidini, McHenry, & Mansfield, 2008; Vedula, Bero, & Scherer, 2009). Another strategy consists in bundling outcomes in a way that changes the presentation of results from negative to positive or from insignificant to significant, for example through the use of composite health indicators (Montori, Jaeschke, Schünemann, & et al., 2004; Shaughnessy, 2003). A third strategy implies ignoring the drop-outs that inevitably occur during a trial, which can result in dramatically overstating the benefit of a treatment (Melander, Ahlqvist-Rastad, Meijer, & et al., 2003).

As with the two previous outcomes discussed in this chapter, evidence of the effect of the performative logic on the reporting of outcomes is scarce. Although there is no evidence that a stronger performative logic is positively correlated with a more selective reporting of outcomes, the literature suggests that it will create pressure on evaluators. A recent web survey of some 200 academics having done policy research for the British government since 2005 indicates that government officials were more likely to propose changes affecting the interpretation of findings or their weight than not. However, it is less clear from the survey whether the requested changes did help produce supportive reports (The LSE GV314 Group, 2014). Beyond this survey, the evidence base consists mostly of some anecdotes, such as Metcalf's report of the pressure exerted by the US Department of Agriculture during the evaluation of the National School Lunch Program (Metcalf, 2008).

2.7. Relevance of the analogy between medical research and policy evaluation

Throughout this chapter, I have shown that the literature on the influence of industry sponsorship on the scientific credibility of clinical trials could serve as a useful guide for the study of policy evaluations. Indeed, this literature has very effectively analysed the conundrum which researchers face when they become agents. Many more references to the medical literature will be made in the remainder of this thesis. However, the comparison has also important limitations, which need to be fully understood before we move on to the empirical part of this work. The following section briefly discusses the similarities and differences between medical research and social policy research.

The main similarity between clinical trials and social policy evaluation is that they are both a type of *applied research*. In other words, neither is conducted with the primary purpose of advancing knowledge. Rather, they are meant to inform important decisions about the development of a product or policy, which the organisation is already committed to launch (to varying degrees). This change makes the conduct of research somehow more complex. Each decision not only needs to satisfy the norms of science, it also needs to support the aims of the organisation. Thus, both clinicians and policy researchers have to find the right balance between *professionalism* and *loyalty* to their employer – or reciprocation of 'favours' in the case of contract research (Hood & Lodge 2006).

However, there are also important differences between these two types of research. The first difference is in their purpose. Clinical research is essentially *confirmatory*, i.e. it quantifies the extent to which deviations from a model could be expected to occur by chance (Gelman 2004). This is due to the fact that (i) health-related variables are easily quantifiable and (ii) medical treatments entail a risk. Medical treatments can not only fail to cure life-threatening diseases, they can also create other diseases and even kill. This is why new drugs have to undergo a series of four consecutive clinical trials[4], all using randomised controlled trials (RCTs). It is commonly admitted in the medical community that RCTs are the most robust way of evaluating the efficacy of a new treatment.

Conversely, social policy research is essentially *exploratory*, i.e. meant to isolate patterns and features of the data (Hoaglin, Mosteller & Tukey 1983). There is no restriction regarding the type of research that can be used to evaluate the effect of a social intervention. Impact evaluations can be conducted using any kind of design (experimental, quasi-experimental or

[4] Phase 1: Screening for safety; Phase 2: testing the efficacy of the drug, usually against a placebo; Phase 3: confirmatory study; Phase 4: post-marketing studies delineating additional information on the drug's benefits, risks and optimal uses.

non-experimental) and any kind of data. Indeed, the idea of a 'gold standard' in social policy evaluation is a highly contested one (Hollister 2008, 2009; Nathan 2008a, 2008b, 2009). Thus, social policy reforms are frequently rolled out based on evidence that the intervention was properly implemented, or that beneficiaries were satisfied with the intervention. This 'flexible' approach to research means that there is little consistency across studies in terms of research questions, data and design. Furthermore, the absence of research protocols means that the risk of spin is high (see chapter 7 for an empirical study of spin in policy evaluation).

The second difference is that in medical research, evidence of the cost-effectiveness of a drug has an instrumental use: it is the single most important piece of information that will be considered by regulators in their decision to authorise the drug. When such information is clear and unambiguous, the approval process can be relatively straightforward (Lavertu & Weimer, 2010). In contrast, policy evaluation results have a more conceptual use: governments are free to use results as they see fit and are by no means bound to the conclusions and recommendations of evaluators. Other, non-scientific considerations play an equally, and perhaps greater, role in shaping social policy. Those include ethics, morality, legality, policy commitments and political support. For all these matters, the 'expert' is the elected politician, not the scientist. Unlike medicines and healthcare products regulators, social policy-makers can legitimately discard evaluation results that are found unacceptable or undesirable. Importantly, such a decision implies that an evaluation was conducted in the first place. This is a strong assumption given that no government in the world is subject to a formal obligation to evaluate social interventions.

Financial stakes are a third, major difference. The development process from patent filing to product launch has been estimated to take an average of 12 years at a total cost of some £200 million (BMJ 1996). In contrast, the costs related to the development of social interventions seem to be much lower. For example, a 2008 Report from the NAO found that, between 2002 and 2006, the DWP had spent about £40 million on initiatives targeted specifically at ethnic minority employment. These included the Ethnic Minority Outreach pilot (£31 million spent between April 2002 and September 2006), the Ethnic Minority Flexible Fund (£6.8 million spent between April 2004 and March 2006) and the Specialist Employment Advisers pilot (£1.5 million spent in 2004-2006). Other initiatives were trialled for a fraction of these costs (for example, the Mental Health Court pilot and the Virtual Court Pilots were both implemented by the Ministry of Justice for an average cost of £400,000).

2.8. Conclusion

This chapter was set out to propose a theoretical framework for the study of the influence of political institutions on policy research decisions.

The first objective of this chapter was to identify the strengths and the weaknesses of the existing literature, as well as possible research gaps. The review has shown that, collectively, we know a lot about the influence of institutions on policy research. However, this knowledge is fragmented across disciplines and supported by an excessive number of concepts and theories. This chapter was unable to review the literature in a systematic way but identified the most significant contributions in research methods, sociology, public policy and public administration. It recommended carrying out more research to identify the key decision-makers and to formulate a more parsimonious theory that would be applicable to a broad range of countries and policy areas.

The second objective of this chapter was to lay the foundations an empirical strategy for future research in this area. The notion of confirmation bias, used in research methods and social psychology to qualify the tendency of individuals to favour information that supports prior beliefs and hypotheses, emerged as the most desirable option. There is an abundant literature looking at confirmation bias at the individual level and at the organisational level, particularly in medical research. Thus, the question of the effect of political institutions on policy research would be most effectively addressed by questioning the prevalence of confirmation bias in government-funded research. This entails a two-step approach. First, the scientific credibility of the research decisions made by the relevant government(s) must be systematically investigated, based on a number of common research prescriptions. Second, the effect if policy commitments on the scientific credibility of these research decisions must be assessed.

There are many questions that remained unanswered. I will mention three. First, to the extent that political institutions do influence research, this influence must be context-specific. So we need to understand the contexts in which the effect of political institutions is relatively stronger/weaker. Second, we need to know what happens at the individual level. In particular, we need to understand who the actors are and the type of incentives they are subject to. I have shown that the public policy and the public administration literatures disagree on that point; however this could also be due to the fact that they tend to investigate separate policy areas. More detailed accounts in this area would help researchers make credible assumptions regarding decision-making processes. These two questions will be addressed in chapter 3. Thirdly, we need to identify the type of research design and data needed to answer the question of the effect. This question will be addressed in chapter 4.

3. Institutional context

3.1. Introduction

In the previous chapter, I argued that the effect of political institutions on policy evaluation could best be analysed using the 'confirmation bias' theory. According to this theory, the scientific credibility of an evaluation is negatively associated with the strength of the government's commitment to the intervention being evaluated.

Importantly, these phenomena do not take place in a vacuum. Considering that different contexts are likely to strengthen or weaken any association between policy commitments and research decisions, the scoping of the data matters to a large extent. A deep understanding of the singularity of the selected case is essential to assess the credibility and the strength of these associations and to make correct inferences about other places and times.

This chapter serves three purposes. First, it describes the case selection process, with a special emphasis on how constraints and opportunities were handled. Second, it 'sets the scene' by presenting the different ministries, actors and processes which constitute the context of my empirical work. Third, it discusses the substantive implications of the case and defines a number of expectations. The focus on the Department for Work and Pensions (DWP) in some parts of the document is for convenience only, as it is the government department on which we had the most evidence.

The evidence in this chapter comes from two main sources. First, I reviewed the administrative and scientific literature to gather a maximum of background information on the organisation and the management of British ministerial departments in general and on research decision-making processes in particular. In addition to this review, I interviewed 15 policy researchers between October 2011 and February 2012 (See Annex II). Two types of interviews were conducted, all using semi-structured questionnaires. The first five interviews were meant to clarify the research process at the DWP and the role of the different actors. Each of the ten other interviews focused on one pilot in particular and addressed the effect of policy commitments and political salience. Pilots were chosen with a view of having some diversity in terms of sizes, levels of complexity and political salience. Interviewees were asked to comment primarily on that case. What follows is a thematic analysis of the evidence collected from the documentary review and interviews.

The rest of this chapter is organised as follows. Section 3.1 describes and justifies the case selection process, highlighting the numerous trade-offs between scientific rigour and efficiency. Section 3.2 provides a 'negative'

description of the case itself, contrasting the observed units of analysis with the unobserved ones (in different countries, at different times), notably to inform the upcoming discussion on external validity. Section 3.3 'sets the scene': it presents the organisations where research decisions are made, the actors involved in these decisions as well as the decision-making process. Section 3.4 sheds light on various incentives explaining why one might expect an influence of policy commitments on research decisions.

3.2. Case selection

The case selection was constrained by two factors, namely the nature of the dependent variables and the availability of data.

The first constraint was imposed by the dependent variables identified in chapter 2. The need for a proper sequencing between the research phase and the policy decision, for clear sampling mechanisms and for outcome measures unambiguously implied a focus on experimental or quasi-experimental policy-making ('piloting'). This is an important restriction, given the small amount of policy interventions evaluated in this way across the world. There is unfortunately no systematic data on the number and location of these research projects[5]. A quick scan of this data shows that experimental and quasi-experimental evaluations are concentrated in two types of countries: a few high-income countries (US, UK, Canada, France, Netherlands, Denmark to varying degrees) and some low-income countries, where experimental methods have been used to evaluate development aid programmes. The latter countries, however, do not offer a suitable context to answer my research question. Most of the interventions that have been experimented on there were not sponsored by the local government and thus were not necessarily linked to any sort of policy commitment. Rather, they were commissioned by donors or lenders such as the World Bank or grant-making foundations such the Bill and Melinda Gates Foundation. Therefore, the choice was limited to the above-mentioned high-income countries, which have the additional benefit of enjoying stable institutions and more 'traceable' bureaucratic procedures.

The second most pressing constraint was the availability of data. Two criteria had to be borne in mind during the scoping of the study. The first criterion was the quantity of the relevant research projects. It quickly appeared that the US and the UK were the only two countries that could provide the data I needed. Unlike in other countries, experimental and

[5] Some databases of experimental and quasi-experimental evaluations are available online. See for example:
- The J-PAL databse: http://www.povertyactionlab.org/evaluations
- The American Economic Association's registry for randomised controlled trials: https://www.socialscienceregistry.org/
- 3ie's Registry for International Development Impact Evaluations: http://ridie.3ieimpact.org/

quasi-experimental policy-making have been used extensively and almost routinely in those two countries in recent history. The second criterion was the need for to keep the level of institutional complexity to a minimum. Using a country with a centralised system as opposed to a federal one made the data collection much easier as the concentration of powers considerably restricted the number of potential evaluation commissioners. The UK thus appeared to be the best choice. I must also acknowledge that, being a UK resident, I knew that I would have an easier access to information and a better understanding of the phenomena described in this than in any other country. This personal consideration certainly influenced the case selection.

The decision regarding the timeframe was equally pragmatic, *i.e.* with a view to provide enough data without introducing too much heterogeneity in terms of policies and governments. In the end, I focused on research projects commissioned between May 1997 and May 2010. This period corresponds to the Labour governments led by Tony Blair and Gordon Brown. Finally, and with a view to maximise the number of observations, I included the four government departments offering the largest number of observations, namely the Department for Work and Pensions (and its predecessors, Department of Social Security and Employment Service), the Department for Education (and its predecessor the Department for Children, Schools and Families); the Home Office and the Ministry of Justice (and its predecessor (the Department for Constitutional Affairs).

Further specifications will be added in chapter 4, which deals with data and methods. At this stage, it is sufficient for the reader to know that this thesis investigates the effect of policy commitments on the research decisions made by four British ministerial departments (Department for Work and Pensions, Department for Education; Home Office and Ministry of Justice) between May 1997 and May 2010.

3.3. Generalisability

Selecting a specific case out of convenience as opposed to randomly sampling it from a whole population has a number of implications for this thesis. The first of these implications concerns the generalisability of the conclusions to other situations. Understanding the situations which the case better represents is instrumental to making meaningful inferences. The following section shows that the chosen case is more representative of (1) countries with a strong evaluation culture than countries with a weak evaluation culture; (2) post-Labour Britain than pre-Labour Britain; and (3) research in ministerial departments than research in non-ministerial departments.

3.3.1. Evaluation outside the UK

Using the UK as location calls to reflect on its specificities in terms of evaluation culture and capacity.

The operationalisation of these concepts is not without difficulty. Attempts to rank countries based on their evaluation capacity have been made in the past. For example, the authors of the *International Atlas of Evaluation*, scored 21 high-income countries based on a set of nine institutional criteria including the supply of domestic evaluators, institutional arrangements in government for conducting evaluation, and pluralism of institutions and actors conducting evaluation (Furubo, Rist, & Sandahl, 2002). Out of these 21 countries, the United States came first and the UK fifth, equally placed with the Netherlands. According to the authors, countries with a high score, such as the UK, have both internal and external incentives to evaluate their policies. This said, by the authors' own admission, the instrument is not highly scientific and thus one should refrain from trying to over-interpret these findings. More importantly, the above ranking does not say anything about the scientific rigour of the evaluations conducted in these countries.

Another way of analysing the position of the UK in terms of evaluation culture is to identify a surrogate indicator that would be both objective and comparable across countries, such as the presence of a high-level commitment to publish all evaluation findings regardless of the results. Publication is a central feature of the scientific approach; indeed it is a pre-condition to the falsification of results. In the UK, such commitment can be found on three levels. Firstly, the right to access information held by public authorities has been granted to British citizens in 2000 with the Freedom of Information Act. Although information relating to the "formation of government policy" is exempted from the Act, evaluation results are usually not considered as such. Secondly, self-regulation requires that online publication of evaluation studies be considered the default option across government departments (Government Social Research Unit, 2010b). Thirdly, the Labour Party, which had the majority in the UK Parliament between May 1997 and May 2010, expressed its commitment to transparency and access to information on several occasions when it was in government (Cabinet Office, 1997). This commitment explains why a vast majority of research outputs is published in the UK.

Unfortunately, it was not possible to collect data in a way that would allow for systematic cross-national comparisons. However, selected comparisons would indicate that the UK is one of a few countries where the government is committed to the publication of its evaluations. In the US, this decision is left to the discretion of each department. For example, the website of the US Department of Labour indicates that the Department will release results "of all evaluations that are not specifically focused on internal management, legal, or enforcement procedures or that are not otherwise prohibited from disclosure. Evaluation reports will present all results, including favorable,

unfavorable, and null findings[6]". In France, the Department of Labour publishes every year an activity report presenting all studies commissioned by the Department. However, in-house research project are not mentioned and the reports are not available online.

Against this background, I would expect the case to be more representative of high-income countries with strong evaluation capacity (such as the US, Canada, Australia, New Zealand, Denmark, Sweden, etc.) than of other countries.

3.3.2. Evaluation before and after New Labour

The second important characteristic of the case at hand pertains to the chosen timeframe, which is that of the Labour government (May 1997 – May 2010). Such a timeframe allows comparisons with this government's predecessor and successor.

The governments of Tony Blair and Gordon Brown have often been presented as a 'golden age' of policy evaluation (Furubo et al., 2002). Their strategy was set out in the 1999 White Paper 'Modernising Government', which offered a strong commitment to more evaluation, the modernisation of evaluation standards and tools, and an enhancement of the evaluative capacity of government (Cabinet Office, 1999a). One should not conclude too hastily that evaluation under New Labour was transformed 'from famine to feast'; indeed the reality was often less rosy than policy documents painted it (see Maguire (2004) for a case study). However, the New Labour era was marked by more evaluative activity directly commissioned by and for government. Whether this was the outcome of the substantial growth in public spending (between 1999 and 2007), or an attempted to fill an ideological gap within New Labour (as suggested by Furubo et al.) is open to discussion.

Labour contrasts sharply with the previous Conservative government (May 1979 to May 1997). Furubo and colleagues report that under Thatcher and Major, UK evaluation was essentially fragmentary and linked closely to resource management. There was little by way of either an established community devoted to policy evaluation or formalised procedures for initiating, conducting and utilising evaluations in the policy process. Evaluation was seen as marginal to departmental interests and overlooked by ministers other than as a tool for expenditure reduction. The idea that it might inform policy effectiveness was limited to a few enthusiasts (Furubo et al., 2002).

Whereas the victory of New Labour in the 1997 election marked a radical change in the UK government's approach to policy evaluation, the contrast

[6] http://www.dol.gov/asp/evaluation/EvaluationPolicy.htm

with the post-Labour era is much less obvious. It is still early to assess the Coalition's commitment to evidence-based policy and rigorous evaluation. The decisions made so far by the new government have sent mixed messages. On the one hand, many stakeholders have criticized the £3 million cut in spending on evaluation between 2010 and 2013 (National Audit Office, 2013). Furthermore, the NAO established in its report that there was a lack of evaluation in progress or planned for the major projects identified by each department in their business plans. On the other hand, there have been some positive developments as well (Rutter, 2012). Those include the setup of a 'What Works' network - which makes the UK one of the first countries to allocate resources to evidence synthesis on such a scale – as well as the creation of the Cabinet Office's behavioural insights team with the aim of promoting randomised controlled trials and cost-benefit analyses in policy-making.

In light of the above, I would expect the case to be more representative of the Labour and post-Labour era than of the pre-Labour era.

3.3.3. Evaluation outside ministerial departments

The third characteristic of the case at hand pertains to the fact that it focuses on ministerial departments and deliberately leaves aside non-ministerial departments, also known in the UK as non-departmental public bodies (NDPBs).

This distinction is important. Ministerial departments such as those included in my case are led politically by a government minister, normally a member of the Cabinet and cover matters that require direct political oversight, such as the formulation of new policies and their implementation. They also increasingly evaluate public policies and programmes. In contrast, NDPBs generally cover matters for which direct political oversight is judged unnecessary or inappropriate. A typical NDPB is established under statute and is accountable to Parliament rather than to the government. Research Councils are an example of NDPB in the UK. Those include *inter alia* the Arts and Humanities Research Council (AHRC), the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC). It can be argued that ministerial departments are essentially concerned with their performative reputation, whereas research councils will promote their technical-scientific reputation (Carpenter 2010).

These crucial differences justify different public service bargains (PSBs) between the government and the personnel of these organisations (Hood & Lodge, 2006). Researchers working in ministerial departments are likely to be agents and, as such, to follow the instructions given by the principal. The principal is in turn held to be responsible for the actions of the agent. Conversely, researchers working in non-ministerial departments are likely to be in a relation of *trusteeship* with the government. Trustees are subject

to fewer controls. Under a PSB of the trustee type, the tenure and rewards of public servants are not under the direct control of those for whom they act; the skills and competencies they are expected to show are not determined by the instrumental interests of elected politicians and loyalty lies to an entity that is broader than the government of the day (Hood & Lodge, 2006).

The distinction between trustees and agents is helpful to understand the different degrees of autonomy that staffs working in ministerial departments and NDPBs can enjoy; however one should not see these two concepts as mutually exclusive. Aspects of trusteeship may apply to Whitehall civil servants, even though the PSB they operate under is often described by scholars and officials as if it were purely of an agency type of responding to or anticipating ministers' decisions (Hood & Lodge, 2006).

Against this background, I would expect the case to be more representative of research carried out in ministerial departments than in NDPBs.

3.4. Research decision-makers in UK ministerial departments

Understanding what shapes policy research decisions first requires identifying the decision-makers. The following section shows that, contrary to popular wisdom and some previous accounts (notably the public policy literature, see section 2.2.3), policy research decisions are not made by elected or senior policy-makers. Evaluation is the responsibility of middle managers (Wilson, 1989) or "first floor bureaucrats" (Page & Jenkins, 2005). Among them, two types of officials have an extensive responsibility on research decisions: the 'analyst' and the 'policy-maker'.

The following section is supported by Exhibit 8 (courtesy of Boa et al., 2010), which describes the annual research cycle at the DWP's Work Welfare and Employment Group (WWEG). It shows that although analysts are present at all steps, the process is actually dominated by individuals having a vested interest in the success of the reform.

3.4.1. Ministers

Whilst the idea that ministers can use their position to influence research decisions cannot be ruled out given the high levels of political legitimacy and political acumen that most of them have, it comes with too many assumptions to be really credible.

Exhibit 8 – Programming research at the DWP

Time	Action	Who's involved
September to October	Discussions with policy colleagues about research priorities	Analysts and policy makers – occasionally external stakeholders
November	Presentation to EASG of short research priorities papers. EASG identify synergies and take an initial view on the prioritisation of research priorities for WWEG next year.	Analysts
November to December	Detailed Project Initiation Documents (PIDs) are drafted for all proposed projects. These are peer reviewed by other WWEG analysts.	Analysts and policy colleagues
December	Shortened PIDs are prepared for projects previously approved for current year but not yet started.	Analysts
January	Detailed PIDs are considered by EASG. EASG focus on expensive projects or those with reservations expressed at peer review. Previously approved projects are reconsidered at same time rather than being automatically approved.	Representatives from policy and finance, along with analysts.
February	Submission to ministers outlining the proposed programme – including major evaluations funded from programme budgets.	Ministers
After April	Research projects are given funding approval.	

Source: DWP; Boa, Johnson and King (2010).

Firstly, the assumption that ministers have a direct interest in research decisions is a far-fetched one. The wide range of roles that ministers have to perform is rarely understood outside Whitehall. The diverse constitutional and political constraints they are subject to means that they are dependent for their standing on the need to satisfy a wide range of people and groups, and, above all, the prime minister (Riddell, Gruhn, & Carolan, 2011). Comparisons, often erroneous, are made with the heads of private sector organisations. In fact, as noted by Rhodes in its observation, the diversity of issues and audiences that ministers face means that there is no obvious reason to prioritize economic rationality over political rationality, rather the converse (Rhodes, 2013). So, much government is not about strategy and priorities but the appearance of rule: "Keeping things going, preventing anarchy, stopping society falling to bits. Still being here tomorrow" (Lynn & Jay, 1984).

Secondly, the assumption that ministers know how to make a research decision bears little credibility. Some have underlined how ill-suited and

under-prepared most ministers are for their posts (Riddell et al., 2011). Most come to the role without adequate training and experience, often with little expertise in the subject matter of their department, knowing that the insights required performing the job effectively may only be gained through experience. It is hard to think of another profession or career where an individual could rise to the very top, and assume a position of heavy responsibility, having had no previous acquaintance with that line of work.

Even assuming that some ministers had an interest in policy research decisions, it is unlikely that they would have a 'political base' within their department to pull strings. The high ministerial turnover in the UK – as shown in Exhibit 9 – has often been pointed as impeding the effectiveness of ministers (Riddell et al., 2011). The unusual nature of the rapid turnover in some posts in the UK is vividly illustrated by a comparison with Germany. Riddell *et al.* showed that, since 1949, Germany (including the former West Germany) has had just 15 ministers for the economy (excluding finance), while the UK has had 35 ministers in the equivalent position (in the Department of Business, Innovation and Skills and its predecessors).

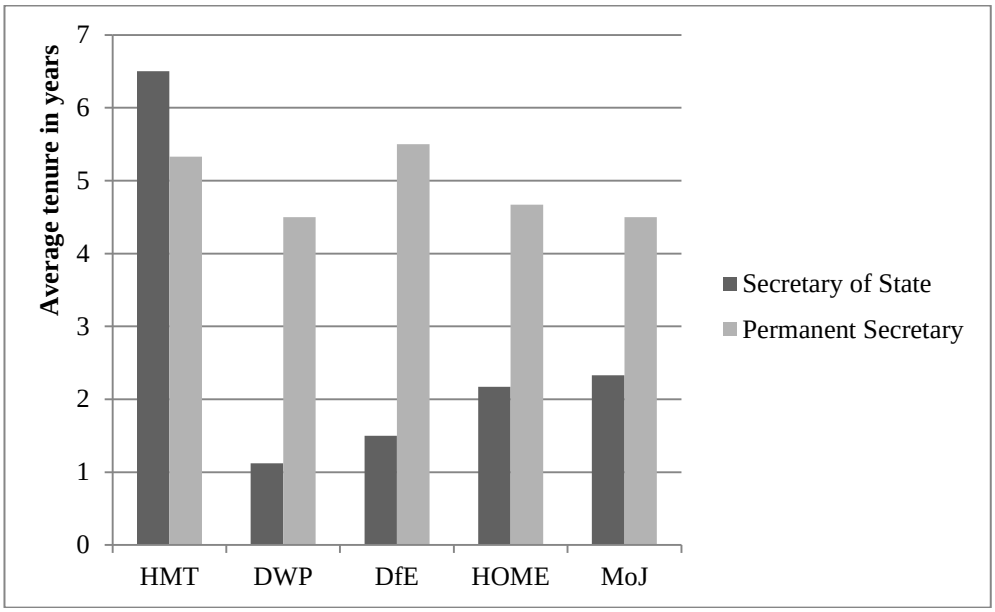
Descriptions of the research process in ministerial departments have mentioned that ministers usually step in quite late to formally approve the evaluation programme and authorise funding (Boa et al., 2010). This suggests that, to the extent that ministers influence research decisions directly, it is more as veto players than decision makers.

3.4.2. Permanent secretaries

The second actor who might be thought to play a role in research decision-making is the department's permanent secretary. The permanent secretary is the most senior civil servant of a British government department, charged with running the organisation on a day-to-day basis. His role includes: policy advice, securing policy implementation, the management of the 'day-to-day business, financial management and a role as 'guardians of propriety' and of the rules and conventions of how government should operate (Paun & Harris, 2013). As noted by Rhodes, the roles and responsibilities of permanent secretaries overlap to a large extent with those of ministers, so much so that it would be more accurate to talk about a "class of political-administrators" to describe the politicians and administrators at the top of the Civil Service (Rhodes, 2013). Against this background, expecting permanent secretaries to play a direct role in the making of research decisions is unrealistic, for the same reasons as ministers. This claim is in line with the findings of Page and Jenkins, according to whom senior civil servants have a wide range of responsibilities and cannot be closely involved in the work of them all. Also, they often do not possess the technical expertise needed to understand the work middle-ranking officials

do. Their contribution to the work of middle-ranging policy officials is mostly indirect and informal (Page & Jenkins, 2005).

Exhibit 9 – Turnover of UK secretaries of state and permanent secretaries between May 1997 and May 2010



Notes:

- *Secretary of State for Work and Pensions: Average tenure between June 2001 and May 2010;*
- *Secretary of State for Education: Average tenure between June 2001 and June 2007;*
- *Minister of Justice: Average tenure between June 2003 and May 2010;*
- *Permanent Secretary to the DWP: Average tenure between May 2002 and May 2011;*
- *Permanent Secretary to the Department of Education: Average tenure between 2001 and 2012 (includes Department for Education and Skills);*
- *Permanent Secretary to the Ministry of Justice: Average tenure from 2003 to 2012 (includes the Department for Constitutional Affairs).*

3.4.3. Policy teams

The first group of officials shaping policy research decisions includes the respective 'policy teams' of each department. As noted by Page and Jenkins, policy-making is not only a political activity – involving the manoeuvring of different politicians, groups and individuals to shape policy – but also a bureaucratic one. Policy teams shape policies into a form that can be put to ministers and a wider audience and turned into a set of policy instruments in the form of a law, plan, budget, consultation document, etc. Politicians need bureaucrats to develop and maintain policy, not simply for 'advice' on how to do it (Page & Jenkins, 2005).

The role of policy teams in policy research decisions is evidence in several documents. First, the description of the research process at DWP shows that, formally, policy teams are involved in all research decisions that might affect the government's agenda (Boa et al., 2010). This includes the identification of 'research priorities' as well as the drafting of Project Initiation Documents, which the blueprint of the study to be conducted (scope, research question, timing, budget, division of tasks, etc.). The reader should assume that this description is an adequate representation of other departments as well.

Second, the role of policy-makers in the British Civil Service is outlined in the *Policy Skills and Knowledge Framework*, which was revised in March 2013. As shown in Exhibit 10, three core competences are mentioned, namely (1) the expertise to produce and use evidence for policy purposes; (2) an understanding of political constraints; and (3) the skills required for the implementation of policy. These competences reflect to a large extent the concerns of middle-managers, in particular the duty to deal with constraints and to comply with the organisations' priorities (Wilson, 1989).

Third, there is empirical evidence of the role of policy teams in the research process (The LSE GV314 Group, 2014).

Exhibit 10 – Role of policy professionals as per the Policy Skills Framework

Competence	**Role/expectation**
Evidence	– "Compile, assimilate, distil, interpret and present a strong evidence base from a wide range of types of evidence and opinions".
Politics	– "Translate ministerial vision into a clear outcome, and develop a clear and shared understanding of what the problem is and what success looks like; test mutual understanding of the problem and goal". – "Support ministers' engagement with parliament and enable public accountability in their area".
Delivery	– "Systematically identify issues that could affect implementation and addressing them/steps to mitigate gaps or weaknesses throughout the life of the policy". – "Maintain political legitimacy, and mandate, throughout the life of the policy, working across government to co-ordinate progress towards shared objectives".

Source: https://civilservicelearning.civilservice.gov.uk/sites/default/files/link_3_-_policy_skills_knowledge_-_curriculum_map_with_cpd.pdf

3.4.4. Analysts

The second group of officials shaping policy research decisions are the department researchers. The British civil service employs a number of researchers including economists, statisticians, biologists, psychologists, sociologists, etc. Together, they form the 'Analyst' profession, whose main responsibility is to provide evidence for policy-makers (Government Office for Science, 2013).

The formal role of analysts in the UK civil service is defined in the Government Social Research Competency Framework (see Exhibit 11). The framework provides behavioural indicators for the five levels of the profession, from Research Officer to Chief Research Officer. The expected skill set of an analyst, according to the Framework, includes (1) intellectual capacity; (2) delivery skills; (3) interpersonal skills; and (4) leadership and management skills.

The first skill of analysts is to be capable of designing, managing and reviewing policy research projects. Exhibit 11 shows the government's expectations for each role. The GSR framework confirms that analysts are responsible for making methodological decisions. Interestingly, the framework makes no mention of the criteria that should guide these methodological decisions. There are many references to "innovative methods"; however there is no definition of what makes a methodology 'innovative' and no justification for the desirability of 'innovative' methods. It is also worth noting that no explicit reference is made to scientific norms such as the production of adequate, valid and reliable empirical evidence; and the application of logical consistency (Merton, 1942; Zuckerman, 1988).

Empirically, the role of analysts in the research process has been best described in the LSE GV314 study (2013). According to research contractors, analysts are those, within government, who seem to be the most concerned with the scientific quality of research outputs.

Exhibit 11 – Role of Government Social Researchers, as per the GSR Competency framework

Grade	Examples of 'intellectual skills'	Examples of 'delivery skills'
Research Officer	– "Designs small scale and less complex research projects for either in-house work or commissioned projects"	– "Identifies who the customer and key stakeholders are for each project; works with others to identify customer needs".
Senior Research Officer	– "Draws upon a track record of designing medium sized or more complex projects to translate a policy question into a viable research specification or in-house project".	– "Engages actively with customers to clarify and determine their needs; ensures those needs are addressed".
Principal Research Officer	– "Takes the lead on a number of 'technical' matters within the wider GSR/ analytical community, for example, this could be methodological". – "Supports SROs/ROs on selection of methods and can deal with more complex problems without detailed knowledge of project".	– "Influences and negotiates effectively with a range of stakeholders/contractors and in different situations, even when the audience is sceptical or hostile".
Senior Principal Research Officer	– "Encourages staff to consider new and innovative methods in social research and evaluation". – "Actively encourages the use of innovative research methods and analytical techniques among team members".	– "Generates workable solutions to complex problems while taking into account the full range of stakeholder perspective and risks". – "Sensitive to customers' wider political and organisational priorities".
Chief Research Officer	– "Keeps abreast of critical methodological developments within social research and identifies the value to the department, and across Whitehall, of new research techniques and approaches". – "Mediates effectively when there is a professional dispute, for example, on issues of methodology".	– "Anticipates changing priorities and manages this through strategic contingency planning". – "Consults with customers and partners rather than imposing solutions; involves stakeholders in deciding what has to be done and what can be done better".

Source: http://www.civilservice.gov.uk/wp-content/uploads/2011/12/gsr_competencies_framework.pdf

3.4.5. Contract researchers

Most of the policy evaluation research in the field of employment and welfare is conducted by external organisations (Boa et al., 2010). The DWP commissions research from a framework of approved expert suppliers. A framework is an agreement with a group of suppliers, which sets out the terms and conditions under which specific purchases can be made. The tasks involved in the 2013 Framework include *inter alia*:

- Qualitative and quantitative research and evaluation;
- Data collection through fieldwork and/or interrogation of administrative data supplied by the department;
- Survey design/methodology;
- Sampling;
- Pilot studies and experiments;
- Literature reviews;
- Evaluation of policy measures.

The involvement of contractors in research decisions varies from one project to another. Sometimes they are called on very early on to conduct feasibility studies, give advice on sampling or identify the type of data that could be used in the evaluation of a particular programme. More frequently though, their job is limited to research planning, data collection, analysis and reporting, with more 'fundamental' decisions taken by the civil servants in charge of the project. It is safe to say that, overall, the role of contractors increases as the project goes along; however not in a linear way. Analysts and policy-makers occasionally step up when key decisions need to be made, such as the design of survey questionnaires.

3.5. Expected effect of policy commitment on the research process

The above shows that all the actors involved in the research process face the same dilemma. On the one hand, they all share an interest in getting the best possible evidence. On the other hand, they work in an organisation, which is tasked with the implementation of a policy agenda. And this policy agenda is largely beyond their control. In what follows, I show that, in the context of UK ministerial departments, the two objectives are not equal. There is strong evidence that scientific considerations are secondary only to performative considerations.

3.5.1. The business of ministerial departments

It is only a slight exaggeration to say that policy research and evaluation is a drop in the ocean of government business. Comparable data is difficult to get across departments however some figures are telling.

First, policy evaluation is a minor expenditure for most departments. For instance, the DWP spent on average about £20 million per year on 'external

research' (Government Office for Science, 2012). This amount must be compared with the DWP's *departmental expenditure limit* – i.e. the budget allocated for the running of the services that it oversees and the everyday cost of resources such as staff – which was £8.3 billion in 2012-2013. It can also be compared with the department's *annually managed expenditure (AME)*, i.e. the amount it spends on programmes which are demand-led – such as welfare, tax credits or public sector pensions. In 2012-2013, the DWP's AME was £166 billion.

Second, policy research occupies few people in government. Looking again at the DWP, we can see that in 2011, 679 people were working on policy research (Government Office for Science, 2012). As a comparison, in January 2014, the department employed nearly 100,000 staff (including Jobcentre Plus), which made it the biggest government department in the UK.

Third, policy evaluation is not a very scrutinised activity. This will probably not come as the surprise to the reader given the two above-mentioned points. Out of the 1,486 reports published on the NAO website between January 1999 and March 2014, only six of them focused on the practice of evaluation, including four on regulatory impact assessments. Since 2010, the Parliament's Public Administration Select Committee (PASC), which controls the matters relating to the quality and standards of administration within the civil service, has launched 49 inquiries on subject as diverse as crime statistics, public engagement in policy-making, the Civil Service Reform or on public procurement. Not a single inquiry dealt with the practice of policy research. Thematic committees such as the Home Affairs Committee, which examines the activities of the Home Office and associated public bodies, have not shown a greater interest in the issue. It would seem that the only parliamentary report dealing with the use of scientific advice in government is a House of Commons' Science and Technology Committee report of 2007 (House of Commons, 2007).

3.5.2. Effect of policy commitments on ministers and permanent secretaries

More in-depth analyses of the utilization of research by British policy-makers have corroborated the idea that the 'scientific' mission of ministerial departments was secondary to their 'implementation' mission.

This is the case among ministers and senior civil servants. In his very detailed account of the policy-making process in UK ministerial departments (mentioned earlier), Rhodes argues that permanent secretaries are anxious to ensure the implementation of the decisions made by the government. According to him, "Both Conservative and Labour governments want departments to implement their policies effectively. The permanent secretary must get on with the job of ensuring the departments

'deliver'" (Rhodes, 2013). The incentive is even stronger when policy implementation depends on third parties (private contractors, local authorities, etc.). Ministers and permanent secretaries must compensate for the fact that "they have a hands-off, not hands on, link to policy implementation" (Rhodes, 2013).

This focus on performance has significantly increased under Labour, with the set-up of the Prime Minister's Delivery Unit and of the Public Service Agreements (PSAs) (Barber, 2008). PSAs were first introduced in the 1998 Comprehensive Spending Review which set around 600 performance targets for around 35 areas of Government (Cabinet Office, 1998). These were refined on several occasions until 2010, when the Coalition government scrapped them.

The UK government's performative logic can be seen in the way senior civil servants approach research findings. Evidence is more likely to be used 'symbolically', i.e. to justify or legitimate a policy or decision, than instrumentally, to inform a decision. Policy-makers construct the story line by asking "what happened and why?" They also ask whether a story is defensible (to both internal and external audiences); accurate (in that it is consistent with known and agreed 'facts'), believable (in that it is consistent with the departmental philosophy). Crucially, as practiced, rational analysis is retrospective not prospective. It is used to justify decisions already taken by other means and for other reasons. And the other reasons are usually political ones.

3.5.3. Effect of policy commitments on policy teams

Policy teams are also subject to strong incentives to implement government policies. Some of these incentives are formal, as evidenced by the Policy Skills Framework (see section 3.4.3). However, most of ministers' authority on policy-makers is exerted informally (Page & Jenkins, 2005). Ministers rarely issue direct and clear instructions to policy officials that define what they should do with any precision, and senior officials tend to offer advice and support rather than commands and injunctions. Thus, middle-ranking policy officials often need to exercise discretion. They know that any significant policy initiatives, or even any significant features of policy initiatives, either need to be sanctioned by ministers or have to be treated as if they were subject to being sanctioned by ministers. "Discretion is exercised within this context of ministerial sanctioning – actual, deemed or anticipated – and this context shapes the way policy officials think about their roles. The difficulty with bending over backwards is that ministers often have few clear ideas about what they want" (Page & Jenkins, 2005).

Policy teams impose their performative logic on the research process in two ways. First, by using their relative authority. Being generally more senior than analysts, they can formally impose their views on all research decisions

affecting the government's agenda. Furthermore, the description of the research process at the DWP indicates that the degree of involvement of policy teams is positively correlated with the cost of the intervention. Thus, evaluations of larger and more expensive programmes are likely to be more skewed towards implementation than other evaluations.

Second, policy-makers hold the purse-strings. They control programme budgets, which are substantially larger than social research budgets and are often used to fund in-depth evaluations (Boa et al., 2010; Government Office for Science, 2012). Yet, budgetary constraints can also serve to support political ammunition objectives in commissioned research. Salisbury et al. (Salisbury et al., 2011) show how the constraints set by research design features specified by commissioning departments – the budgets, the timelines, as well as the specification of the methods to be used – can prevent the generation of clear judgements of how well or badly a policy is working (The LSE GV314 Group, 2014).

3.5.4. Effect of policy commitments on analysts

The effect of policy commitments on analysts' decisions is both direct and indirect.

The direct effect is through the GSR Competency Framework presented in Section 3.4.4. It shows that research skills are not the only skills required from analysts. Analysts are also expected to show political sensitivity. Some 'delivery skills' mentioned in the framework are presented in Exhibit 11. There, the word 'customer' refers to the 'audiences' defined earlier in this thesis; and include programme beneficiaries (jobseekers, pupils and parents, victims and criminals, etc.), policy and implementation teams within ministries and, to some extent, the members of the Cabinet.

Some ministerial departments provide an additional incentive to guarantee the implementation of ministerial decisions. For example, most of the DWP's analysts are not part of a separate 'research unit' within the department, which would guarantee some autonomy, but are embedded in the policy teams they serve. Analysts are therefore made co-responsible for the implementation of ministerial policy decisions.

The indirect effect of policy commitment has to do with the organisational structure of ministerial departments, which put generalists on top. Analysts can only get promoted within their profession to a limited extent. Indeed, there are few (very) senior researcher positions within a given department. Thus, career-maximising analysts will need to give up their specialisation at a certain point and become generalists (policy-makers) themselves. This could influence their decisions in anticipation.

3.5.5. Effect of policy commitments on contractors

It is difficult to evaluate the influence of contractors on the scientific credibility of research decisions.

On the one hand, contractors are hired based on their research credentials, reputation and expertise. The technical specifications DWP Research Framework indicates that bidders will be evaluated against the following criteria, with a minimum score set for each one:

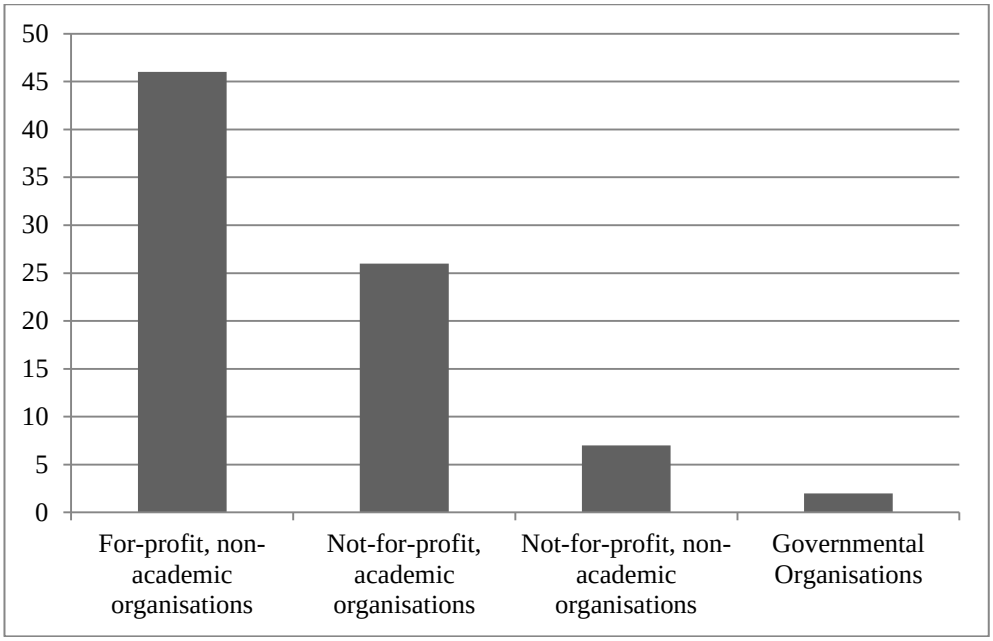
- Methodological expertise;
- Quality of outputs;
- Research ethics;
- Research strengths;
- Examples of relevant research.

The above might suggest that the decisions made for these research projects will be in the interest of robustness and scientific quality.

On the other hand, the contractual nature of the relationship between the department and the consultant limits the autonomy of the latter and encourages reciprocation. The relative dependence of contractors is due to the competitive nature of framework contracts. These agreements are only an umbrella agreement setting out the basis and the terms and conditions on which subsequent call-off contracts are established, but which places no obligations, in itself, on the department to purchase any services. Potential suppliers who are successfully awarded a place on a framework agreement may be invited to compete in "mini competitions" where they are capable of providing the services to be called off. These mini competitions generally involve between two and ten contractors. Besides, it should make no doubt that these contracts are of a commercial nature. The previous DWP framework (2009-2013) included 88 organisations from both the profit and not-for-profit sectors, although the latter organisations essentially use these contracts as money-spinners (see Exhibit 12).

Empirically, the evidence is mixed. The LSE GV314 Study (2013) shows that when asked to make some changes to the final report, academics tend either to oblige or meet their sponsors half-way (three-quarters of respondents). Many respondents mentioned the contentiousness of a given reform to explain policy-makers' hands-on approach to the evaluation. However, the survey provides little evidence that this helps produce supportive reports.

Exhibit 12 – Number of DWP Research Framework contractors per sector (2009-2013 Framework)



3.6. Expected variation across research decisions

In the previous section, I showed that policy teams had extensive influence on all aspects of 'their' reforms, including evaluation. However, the evaluation process is long and technical. Policy-makers, whose responsibility is to lead the reform process, are unlikely to be involved in all decisions. Some will matter more to them than others. The purpose of the following section is to identify these research decisions. It first starts with a general description of the research cycle in a typical government department and then considers the three research decisions analysed in this thesis, namely the timeframe, the selection of pilot sites and the reporting of outcomes.

3.6.1. The research cycle in a typical government department

The research cycle starts once the work programme has been established and approved by ministers (see Exhibit 13). Three main phases can be distinguished: the design phase (steps 1 to 4); the data collection and analysis phase (step 5); and the reporting phase (step 6).

The design phase is concerned with the decisions that have policy and managerial implications, namely the definition of timeframe of the pilot, its

scope, the definition of its objectives, and on some occasions the selection of pilot sites. These decisions are made centrally by policy teams and analysts and must be formally approved by the relevant minister (steps 1 and 2). A research design is drawn up by analysts before going out to competitive tender (step 3). The design may be revisited by tendering organisations in their tender proposals and revised again before the research is commissioned and executed (step 4). However, the most fundamental decisions will have already been made. Feasibility studies are occasionally commissioned for most sophisticated studies.

The data collection and analysis phase starts soon after an evaluator has been appointed (step 5). These tasks are typically performed by the contractor under the supervision of analysts. Within the WWEG the day-to-day management of projects varies quite considerably depending on their size and status. Individuals interviewed by Boa et al. (2010) stated that project management tended to be more proactive on a day-to-day level at the design and reporting stage than in other government departments. This proactive input tends to focus mainly on quality assurance of the work, and the presentation of research results. For large evaluation projects, the DWP project manager can be involved full time.

The research cycle ends with the reporting of the evaluation results. In principle, this task is performed by the sole contractor but, as mentioned earlier, guidance or pressure from both analysts and policy-makers cannot be excluded.

Against this background, one could expect decisions made during the research design phase to be strongly influenced by policy commitments (with some variation depending on the complexity of the decision). Conversely, decisions made during the data collection and analysis phase are expected to be weakly influenced by policy commitments. Reporting is expected to fall somewhere in the middle.

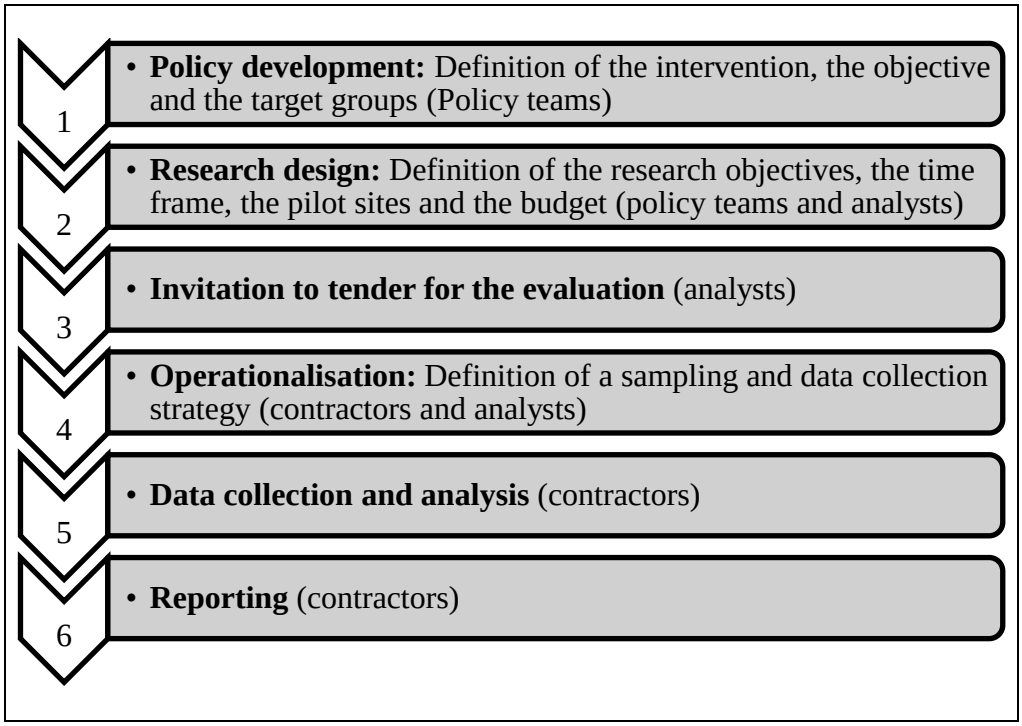
3.6.2. Timeframe

The timeframe of a pilot is decided by the relevant policy teams, as ‘managers’ of a given reform. From an organisational viewpoint, this makes sense as the piloting phase has important repercussions. These repercussions are political (results will need to be available before the next election or the next Spending Review) as well as financial (longer pilots are more expensive than shorter pilots) and managerial (local agencies must be prepared for a possible national rollout). Good coordination among services is thus essential, which is why this competence is given to a generalist rather than to a specialist. Once approved, timeframes are usually conscientiously monitored by policy-makers. The timely implementation and evaluation of a pilot is often a key requirement for the relevant teams. There is clear and consistent evidence that timeliness is a key issue (Jowell 2003; Magenta

Book). Analysts are consulted but have no decision-making power. Research contractors are usually not consulted on the timeframe of research.

Against this background, I would expect policy commitments to have a strong effect on the duration of pilots.

Exhibit 13 – The research process in a typical ministerial department



3.6.3. Site selection

Like the timeframe of a pilot, the selection of pilot sites is usually made by policy teams. Again, this is justified by the organisational implications of this decision, such as the need to negotiate the implementation of the pilot with local policy-makers and front-line agents and the cost of running a pilot across multiple locations.

However, the selection of pilot sites is one of the decisions where analysts have the greatest influence. Indeed, the quality and the size of samples are of paramount importance in research. There is anecdotal evidence of the influence of analysts in this matter. Boa et al. (2010) cite the evaluation of the Pathways to Work pilot, in which DWP analysts were successful at getting the pilot redesigned so that the evaluations provide more meaningful data. They indicate that, having made a convincing case, the size of the pilot doubled from three to seven areas. One of my interviewees confirmed that, although analysts can advocate more or different pilot sites, ultimately the

decision belongs to policy-makers. Research contractors have typically no say.

Against this background, I would expect a weak effect of policy commitment on the selection of pilot sites.

3.6.4. Reporting

The write-up of evaluation reports is the sole responsibility of the research contractor or, in the rare instances where the research has been conducted in-house, of the department analysts.

In theory, one would expect the reporting of findings to be a highly contested research decision. First, it matters to policy-makers. On an instrumental level, findings will help policy-makers make decisions regarding the rollout of the programme and possible adjustments. On a more symbolic level, results will help policy-makers 'legitimise' the intervention among stakeholders. Second, findings matter to analysts, who will 'fight their corner' and try to preserve the scientific integrity of the project and, thereby, their reputation. Finally, it matters to research contractors, who will seek to build up their reputation as experts and reliable business partners.

The LSE GV314 study shows that the reality is more nuanced. Requests to change or scale down critical content are actually far from systematic. The authors report that 52% of respondents were asked to make changes affecting the interpretation of findings or the weight given to them (against 46% of respondents who were not). However, the free responses in the survey as well as the interview material gives ample evidence of those sponsoring the research seeking to shape the way the results are reported. Moreover, survey data suggest that when asked to make some changes to the final report the academics tend either to oblige or meet their sponsors half-way (The LSE GV314 Group, 2014). Here again, these results must be taken with a pinch of salt, as they might be biased by selective memory and social desirability.

Against this background, I would expect policy commitments to have a weak effect on the reporting of evaluation outcomes.

3.7. Expected variation across departments

It has been said repeatedly in this thesis that organisations matter. To the extent that they can freely set their research priorities, hire their staff and allocate resources, different ministerial departments can have different approaches to policy research and evaluation. The following section compares the approaches of the four departments included in this case study

and considers, for each of them (1) the commitments to research and leadership; (2) research procedures; and (3) their use of evidence in policy-making.

3.7.1. Department for Work and Pensions

The ministerial departments which make the policy research decisions that I analyse as part of this project are like any other organisation: they face multiple audiences and pluralistic interests. This makes the arbitration amongst them all the more difficult and interesting.

The Department for Work and Pensions (DWP) was created in 2001 as the result of a merger of the Department of Social Security and the Employment Service. It reports to the Secretary of State for Work and Pensions who defines its agenda and priorities. It is the biggest public service delivery department in the UK serving over 20 million customers.

The DWP has been regularly praised both within the government and outside, for the quality of its research. Between 2008 and 2011, the Government's Office for Science (GO Science) conducted two science and analysis capability reviews of DWP. In 2008, the Department was assessed to be 'strong' in its ability to base choices on evidence, the highest rating. In their second review in 2011, the reviewers made again a positive assessment of the Department. In particular, they identified a "strong commitment across the Department to using analytical and scientific evidence to inform the development and the delivery of policy". They found that the focus on analytical and scientific evidence was supported by the presence of analysts and scientists in several senior policy delivery roles. The review also found "consistently high levels of enthusiasm, commitment and retention among analytical staff which reflects and helps to perpetuate the focus on use of science and analysis".

Differences may also appear in the way departments conduct research. In a recent report, the National Audit Office noted that DWP "did not properly evaluate pilots before launching Pathways to Work. The flawed evaluation gave too positive a view of expected performance". However, the NAO also noted some high-quality evaluations in the area of active labour markets: eight of ten labour market evaluations were of a sufficient standard to have confidence in the impacts attributed to policy (National Audit Office, 2013).

Last but not least, the commitment of a department to research can be assessed through the use of research findings. In its 2013 Report on *Evaluation in Government*, the NAO looked at the percentage of regulatory impact assessments in 2009-2010 referring to evaluation findings. They found out that 80% of DWP's impact assessments were based on evaluation findings – one of the highest scores among government departments.

In light of the above, one would expect the research commissioned by the DWP to be strongly influenced by scientific norms.

3.7.2. Department for Education

The Department for Education (DfE) was formed in May 2010 by the incoming coalition government, taking on the responsibilities and resources of the Department for Children, Schools and Families (DCSF). It reports to the Secretary of State for Education. The DfE is responsible for issues affecting people in England up to the age of 19, including child protection and education.

In 2009-2010, GO Science made also a strong positive assessment of the DfE's use of science and analysis. The investigators noted a clear focus on the use of analytical evidence to inform and guide the development and delivery of policy. The Department's senior leadership was found to play a key role in driving this analytical, evidence-based approach. GO Science also noted that the Department also had many strong links with the academic and wider research community and with delivery partners who are often involved in research and data collection. In its 2013 Report (already mentioned), the NAO report also noted high-quality evaluations in the area of education: six of nine education reports were of a sufficient standard to have confidence in the impacts attributed to policy. However, the DfE was found to perform poorly in terms of research utilisation. None of the impact assessments produced by the Department in 2009-2010 were based on evaluation findings (80% of DWP's impact assessments).

In light of the above, one would expect the research commissioned by the DfE to be moderately influenced by scientific norms,

3.7.3. Home Office and Ministry of Justice

The Home Office (HO) was formed in 1782. It reports to the Home Secretary. It is also responsible for immigration, security, and law and order. It is also in charge of government policy on security-related issues such as drugs and counter-terrorism. In May 2007, some functions of the Home Office were combined with the Department for Constitutional Affairs to form the Ministry of Justice (MOJ). The MoJ reports to the Secretary of State for Justice. Its stated priorities are to reduce re-offending and protect the public, to provide access to justice, to increase confidence in the justice system, and uphold people's civil liberties. In the remainder of this thesis, the HO and the MoJ will be considered as one department.

The two departments featured quite prominently in the Report of the Parliamentary Science and Technology Committee. The members of the Committee also regretted the insufficient scientific leadership within these

departments. The Committee noted that the "Home Office DCSA seemed to have had little input to the transformation of the Forensic Science Service (FSS), a key scientific resource for the Government, describing 'the low visibility of the Home Office Chief Scientific Adviser' as 'a source of concern, particularly in view of the history of weak scientific culture in the department'" (p.23). This view was apparently shared by the research staff of the Home Office, who felt a consistent "lack of appreciation of the value and importance of scientific evidence among (especially senior) officials". Officials noted that it had an adverse impact in many respects: lack of strategic planning or horizon scanning, commissioning hurried and poor quality 'fire-fighting' research, a reluctance to make use of evidence when it is available, poor communication of issues to the outside world, etc.

The House of Commons Committee of Science and Technology gives also an account of the Home Office's research practice. Committee members were concerned to hear allegations from certain academics that departments have been commissioning and publishing research selectively in order to 'prop up' policies. Professor Tim Hope, a criminologist from the University of Keele who has worked with the Home Office, indicated that of two case studies looking at burglary reduction commissioned by the Home Office, the department decided to only write up one: "Presumably [...] because the area-wide reduction was greater here than elsewhere". Professor Hope also accused the Home Office of manipulating the data so as "to capitalise on chance, producing much more favourable findings overall", despite the fact that "for individual projects, the [Home Office] method produces considerable distortion". Other academics have voiced similar concerns. For example, Reece Walters of Stirling University claimed of the Home Office's treatment of research results: "It is clear the Home Office is interested only in rubber-stamping the political priorities of the Government of the day [...] To participate in Home Office research is to endorse a biased agenda".

Looking at research utilisation, the 2013 NAO Report observed that 6% of the impact assessments conducted by the HO in 2009-2010 and 10% of the impact assessments conducted by the MoJ referred to evaluation findings (DWP: 80%, DfE: 0%).

In light of the above, one would expect the research commissioned by the HO and the MoJ to be weakly influenced by scientific norms,

3.8. Conclusion

This chapter was set out to identify a suitable context for this study and understand the substantive implications of this context on the expected prevalence and severity of confirmation bias in policy research. It should be

borne in mind that a different context could have been chosen, and that this different context could lead to different conclusions.

The first objective of this chapter was to identify an appropriate location and time to address the research question as well as reflect on the 'external validity' of the conclusions to be drawn from this case study. This chapter made clear that the choice of the time and place was dictated by convenience rather than probabilistic methods. Thus, throughout this thesis, generalisations beyond the case will be limited. Furthermore, the UK has been until recently a fairly isolated case in terms of policy evaluation. Beyond the UK, evaluation is a quasi-systematic exercise only in the US, Canada, Australia and New Zealand. Other European countries might also evaluate their programmes, however more punctually.

The second objective of this chapter was to 'set the scene' and identify the key actors as well as their motivations. The review highlighted a number of 'stylised facts'. First, it showed that the policy research decisions considered in this thesis were made by middle managers. Ministers and permanent secretaries may occasionally influence these decisions, as veto players, however systematic intervention must be ruled out. Second, these decisions are typically shared between three main groups: policy-makers, analysts and research contractors. Policy-makers have the highest level of seniority and the greatest capacity to influence research decisions. Third, ministerial departments in the UK have strong incentives to implement the government's agenda. They have stronger incentives to perform than to demonstrate scientific expertise.

The third objective was to gather qualitative information with a view to improve the conclusion validity of this thesis. The review revealed that, given the administrative architecture and culture of UK ministerial departments, one would expect the effect of policy commitments on research decisions to be positively correlated with the degree of involvement of policy teams in these decisions. Thus, one would expect policy commitments to have a greater effect at the *beginning* of the research cycle, and for decisions like the duration of pilots. In addition, I would expect the effect of policy commitments to be negatively correlated with the research culture of each department. Of the four departments included in this study, the DWP is expected to be the least subject to confirmation bias. It is followed by the DfE. The HO and the MoJ are on the third step of the podium.

Having defined the theoretical framework underpinning this study and described the context on my analysis, I now need to specify the research design and the data that will be used to answer the research question. This is the purpose of chapter 4 overleaf.

4. Research design

4.1. Introduction

In chapter 2, I discussed the reasons why a government-sponsored policy evaluation might be subject to confirmation bias and presented some of the research decisions that might reflect such a bias. In chapter 3, I argued that the effect of institutions on research decisions was context-dependent and suggested a specific case to take my analysis forward. Britain's Labour government (1997-2010) quickly emerged as the most desirable option. It is now time to operationalise the research question, i.e. to decide what types of data and research design are most likely to bring an answer to the question at hand, given the contingencies imposed by the context.

The goal of this chapter is to identify the most suitable research design to assess the extent of confirmation bias in a policy context. Two more specific objectives have been assigned to it. First, this chapter reviews the methodologies used so far to study the effect of institutions on research decisions, in both a democratic and market context. Second, it introduces the PILOT dataset that I developed for this purpose. PILOT includes observational data on over 230 policy evaluations conducted by the British government between 1997 and 2010 in three policy areas: employment and welfare; crime and justice and education and parenting. PILOT will be subsequently used to test the general hypothesis that the strength of policy commitment leads to different research decisions. This will be done either through regression analyses or through qualitative research; the dataset providing the structure for the rigorous selection of cases.

The rest of this chapter is organised as follows. Section 4.2 presents and critically appraises the different methodologies used so far to analyse the effect of institutions on research decisions and discusses the relevance of a 'meta-research' design, whereby individual studies are used as unit of analysis. Section 4.3 introduces the PILOT dataset, defines the population of interest and describes the characteristics of the sample. Section 4.4 presents the data sources used to populate the dataset as well as the procedures followed to limit selection bias. Section 4.5 presents the main variables included in the dataset gives some details on their operationalization. Section 4.6 discusses PILOT's strengths and limitations as well as the scalability of the method. Section 4.7 concludes.

4.2. Review of methods

Studies on the effect of political institutions on research decisions have so far relied on a number of methods including: participant and nonparticipant observations, interviews, surveys and meta-research. The following section illustrates the trade-offs entailed by each method and shows why the latter design is the most appropriate to the research question.

4.2.1. Participant and nonparticipant observation

Organisational ethnography is not a common research tool in political science given the difficulty to get access to senior decision-makers, however there have been a few interesting contributions (see Rhodes 2013 for a review). In his own account of the British Civil Service, Rhodes draws on three sources of information, which he describes as the pattern of 'practice', 'talk', and 'considered writing' (Oakeshott, 1996; Rhodes, 2013). On *practice*, Rhodes observed the office of four ministers and six permanent secretaries for between two and five days each. On *talk*, he conducted repeat interviews with permanent secretaries, cabinet members and other officials. On *considered writing*, he consulted newspaper reports, copies of speeches and public lectures, and committee and other papers relevant to the meetings he had observed.

Participant and nonparticipant observation are particularly valuable when very little is known about a group and to capture the motivations of social actors as well as the meaning of their everyday activities. It generates descriptive accounts which are valuable in their own right (Hammersley & Atkinson, 1995). Thanks to Rhodes's account, we now have a better understanding of the rhetorical power of the notion of 'evidence-based policy', which can be seen as a form of 'storytelling' (Rhodes 2013). According to Rhodes, civil servants identify and construct their story line by asking "what happened and why?" So, they test 'facts' in committee meetings and rehearse story lines or explanations to see what they sound like and whether there is agreement. In this way, they can anticipate the reaction of external audiences. Other interesting accounts includes Metcalf's on how policy-makers occasionally lean on contract researchers to provide congenial results (Metcalf, 2008) and Allen's, who narrates his experience of contract research in the area of housing and urban policy (Allen, 2005).

However, this method presents number of disadvantages. Firstly, it requires the consent of the observed, which is a difficult thing to achieve. Given the secrecy of policy-making – documents concerning the formulation of government policy are not subject to the Freedom of Information Act – this method requires a high level of trust from policy-makers, which compromise the independence of the researcher and in any case, limits it to the most seasoned researchers. Secondly, the method requires patience,

endurance and the collection of a huge amount of data (participation to meetings, access to documents, etc.), which makes it hardly replicable and cost-effective. Thirdly, it does not allow the analysis of variations across policy areas, across time and research decisions, as desired.

4.2.2. Interviews

Interviews of researchers in single or comparative case studies have been used in the past. For example, this methodology was used to describe how a major US nuclear weapons laboratory (the Lawrence Livermore Laboratory, LLL) controlled the process of research within its boundaries in the face of conflicting norms imposed by the scientific profession and its patron agencies (Sutton, 1984). More recently, it was used to investigate the role of cooperative research centres (CRCs) in Australia as a medium for facilitating R&D collaboration between academic and government researchers (Garrett-Jones, Turpin, Burns, & et al., 2005).

Using interviews in the context of single case studies can be helpful for theory-building and when the number of cases is insufficient for a quantitative analysis. Thus, this method makes sense in the case of the LLL. As noted by Sutton, the LLL was an anomaly in terms of functionalist theory because it conformed neither to the academic ideal of disinterested inquiry nor to the image of applied science as a parasitic and derivative activity. Using rich descriptions gathered from interviews with resident scientists, Sutton concluded that research norms were situationally defined. Conversely, the CRCs analysed by Garrett-Jones *et al.* are "hybrid organisations" drawing upon the practices and cultures of all their participants. Although these organisations are probably more numerous than organisations such as the LLL, there was, at the time of the study, very little understanding of their added-value, which justified the use of interviews. The study of Garrett-Jones *et al.* shows that ultimately, CRCs may be in competition with their participant organisations for human and financial resources in relation to activities such as the commercialisation of research results (Garrett-Jones et al., 2005).

Despite its strengths, this methodology would be inappropriate to the research question this study seeks to answer. Indeed, the research question posed in this thesis is broad and shallow, and thus requires a *nomothetic* approach. Conversely, interviews based on a few cases would lead me to take an *idiographic* approach, namely rich descriptions of narrowly defined situations.

A more satisfactory option would be to interview scientists in relation to a larger number of research decisions or cases. The motivations and uses of this method are diverse. For example, it was used to describe the role of pilot schemes in policy-making in the UK (Jowell 2003). Later, a group of researchers used interviews to provide information on the frequency and

reasons for outcome reporting bias in clinical trials (Smyth, Kirkham, Jacoby, & et al., 2011). Interviews spanning a large number of cases can be helpful when the researcher is eager to identify a pattern in his observations but a quantitative design is not possible due to a lack of affordable and reliable quantitative data. For example, Jowell (2003) collected information about 123 pilot schemes across nine UK government departments. However, his research question, as well as its search for 'mini case studies' meant that the design had to be qualitative. Thus, around 30 face-to-face interviews with policy-makers and policy researchers were conducted to discover their special perspective on the pilots for which they had been responsible.

Although they are better suited to answer questions of prevalence and variation than the previous method, interviews related to multiple cases would still be a weak design in relation to this thesis. Firstly, they give little guarantee of yielding information that is strictly comparable across cases. Secondly, they are prone to social desirability bias. Last but not least, it is a costly method with a limited chance of success, as shown by the Smyth, Kirkham and Williamson study (Smyth, Kirkham, Jacoby, & et al., 2011).

4.2.3. Surveys

Against this background, discussions regarding the most appropriate research design to analyse the drivers of research decisions will naturally lead us to consider a quantitative methodology. Questionnaires are often the first method that comes to mind. Questionnaires can be a cost-effective research method, especially when it comes to survey professions as well organised as researchers. For this reason, surveys have been often used to analyse the influence of institutions on research directions (LSE GV314; Jowell 2003; Amara et al.).

However, the method has important limitations. As with any unethical or socially stigmatised behaviour, self-reported survey data are likely to underestimate the true extent of the phenomenon. Respondents have little incentive, apart from good will, to provide honest answers (Fanelli, 2009).

Different strategies have been devised to overcome this social desirability bias and generate more reliable estimates. For example, a recent survey of psychologists incorporated explicit response-contingent incentives for truth telling and supplemented self-reports with impersonal judgments about the prevalence of practices and about respondents' honesty (John, Loewenstein, & Prelec, 2012). Such incentives led to higher and – according to the authors – likely more valid, prevalence estimates of questionable behaviours. Other surveys have asked questions on colleagues' behaviours rather than the respondent's, assuming that a different formulation would yield more reliable results (Greenberg & Goldberg, 1994; Tavare, 2012). Here again, important variations have been observed.

Results from one method to another can vary significantly. A recent meta-analysis compared the results of 18 such surveys (Fanelli, 2009). The average self-report admission rate was 2.3%. Interestingly, the average report on colleagues' misconduct was 14.5%. However, the interpretation of such a gap is subject to speculation. On the one hand, the effect of social expectations in surveys asking about colleagues could depend on the particular interests of respondents. In general, scientists might tend to protect the reputation of their field, by minimising their knowledge of misconduct. On the other hand, some respondents might have particular experience with misconduct and might be very motivated to report it. In addition, surveys on colleagues' behaviour might also lead to inflated estimates of misconduct because the same event might be reported by several respondents. Finally, the wording of questionnaires was found to matter and when interpreting survey results, one needs to bear in mind that people have different perceptions of what does and does not constitute research misconduct. Scientists were less likely to reply affirmatively to questions using the words 'fabrication' and 'falsification' rather than 'alteration' or 'modification' (Fanelli 2009).

More importantly for the study we are concerned about, the possible causal link between incentive and questionable research decision is hard to establish in this kind of survey. A natural conclusion would be that the incentive pre-dated the research decision but the reverse cannot be excluded for certain. Moreover, the investigator often has little control on the type of cases that are covered by survey respondents. It is not inconceivable that a large number of respondents would be involved in a few research projects, whereas other interesting cases would not be commented.

4.2.4. Meta-research

Meta-research consists in systematically coding and analysing research decisions *as they appear to the meta-researcher* rather than *accounts* of these decisions reported by stakeholders.

Meta-research is typically conducted to evaluate the mean effect of a medical treatment across multiple studies. However, it is equally applicable to any other kind of effect. One could, for example, compare similar studies conducted by different teams and analyse the extent to which the type of institution has an effect on findings. This is quite frequent in medical research, where similar prescription drugs have been trialled by industry-funded and government-funded teams of researchers. One such study for example, looked into the benefit of statin, a high-selling class of drugs used to lower cholesterol. This study found 192 trials in total, either comparing one statin against another, or comparing a statin against a different type of treatment. Controlling for other factors, they found that industry-funded trials were 20 times more likely to give results favouring the test drug (Bero, Oostvogel, Bacchetti, et al., 2007). There are many more examples of such

problem in the medical literature (see for example (Bekelman, Li, & Gross, 2003; Kelly et al., 2006; Lexchin, Bero, Djulbegovic, & et al., 2003; Sismondo, 2008).

Meta-research has numerous advantages. Unlike ethnographic methods, interviews and surveys, it is unobtrusive, i.e. it does not require the researcher to be physically present. This is an important characteristic given the problems of access, subjectivity and social desirability bias mentioned earlier. In addition, meta-research is better suited to collect large amounts of information in a comparable and consistent way. Indeed, using interviews or surveys to collect information such as the timing of pilots would entail a high risk of knowledge/memory bias, especially considering the broad scope of the study (230 studies spanning 13 years and four government departments). Using administrative data (i.e. evaluation studies) is a more reliable option.

However, the method is not without its weaknesses. First, unobtrusive measures reduce the researcher's control over the type of data collected. The method assumes that the data needed by the researcher is (1) largely available; (2) consistently reported across documents; and (3) reliable. These are very demanding assumptions, especially when the documents to be reviewed do not follow any reporting standards or guidelines (a point which will be made many times throughout this thesis). The number of missing values for each variable can be used as indicator of how effective the method is.

A second challenge to overcome in meta-research is sampling bias. Sampling bias is a systematic error due to a non-probability sample of a population, causing some units of the population to be less likely to be included than others, and leading to biased inferences. Systematic reviews were devised in the 1980s to address this specific issue. It is defined as the attempt "to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question. Researchers conducting systematic reviews use explicit methods aimed at minimizing bias, in order to produce more reliable findings that can be used to inform decision making" (Higgins, Green et al., 2011).

Although collecting studies in a systematic fashion drastically reduced the risk of sampling bias, it cannot address the fact that some studies are deliberately not published. Publication bias is the tendency of researchers, editors, and pharmaceutical companies to handle the reporting of experimental results that are positive (*i.e.* showing a significant finding) differently from results that are negative (*i.e.* supporting the null hypothesis) or inconclusive, leading to a misleading bias in the overall published literature (Song, Parekh, Hooper et al., 2010). This issue was first formally analysed in the mid-20th century (Sterling, 1959) and since then has become very well documented, especially in the medical literature (see Kirby Lee, Bacchetti, & Sim (2008) for a review of this literature). So far, the most

effective strategy to overcome publication bias has been the resort to Freedom of Information requests, however those are only possible in a few countries such as in the US and the UK (Fowler, Agha, Camm, & Littlejohns, 2013; Joober, Schmitz, Annable, & Boksa, 2012).

4.3. The PILOT dataset

Having shown why a design based on the observation of naturally occurring policy research decisions was the most adequate, I now turn to the definition of the unit of analysis, the population of interest and the sample underpinning my empirical work. The objective here is to make correct inferences between *what I can observe* and *what I want to know*. The following section contends that the PILOT dataset can be considered in two different ways: as a 'self-contained' case, or as a sample drawn from a hypothetical population.

4.3.1. Research scope

It is useful to briefly remind the reader of the institutional context chosen to carry out this study (see section 3.2 for a more detailed account). Given the type of information needed to answer the research question, it was decided to focus on policy research conducted in the UK. The Labour government was chosen as time frame (May 1997 – May 2010). The selected policy areas include employment and welfare, education and parenting and crime and justice.

4.3.2. Unit of analysis

Meta-research uses datasets in which the unit of analysis is a discrete intervention or treatment. In the area of biomedical research, where the method was first used, these treatments typically include drugs and other therapies. In the area of social research, which is the focus of this thesis, an intervention is a policy aiming to address a type of social disorder. In both the medical and social areas, an intervention is most of the time 'simple', i.e. a single molecule/policy instrument – for example, a type of statin to treat cholesterol or the provision of free school meals to improve the educational attainment of poorer pupils. However, it can also be 'complex', i.e. a specific *combination* of molecules/policy instruments. The reason why complex treatments or policies are considered as one intervention has to do with the underlying theory that the whole is more than the sum of its parts. For example, tri-therapies used to treat people infected with HIV can be considered as a treatment in its own right, given the interaction occurring between the different molecules. Likewise, programmes like the New Deal for Disabled People (implemented by the DWP) is based on the assumption

that, provided together, case management, financial incentives to work and training are more effective than separately. Complex interventions are thus defined as such by the drug manufacturer or the government and taken at face value by the meta-researcher.

I have mentioned earlier in this document that not all policy interventions were equally fit for the purpose of this research (see section 2.5.1 for a justification). Indeed the research decisions most likely to reflect confirmation bias are specific to interventions conducted in an 'experimental' spirit, albeit not necessarily with experimental methods. These 'pilot interventions' are the units of the PILOT dataset.

Pilot interventions have a number of specificities. Firstly, a pilot tests a *national policy intervention*: only pilots initiated by the central government have been included. Conversely, pilots initiated by local authorities or non-governmental organisations have been excluded. I have also excluded pilots initiated by the three devolved administrations of the UK (Northern Ireland, Scotland, Wales), insofar as the scope of their competences differs according to the region and the policy area. Secondly, a pilot has a *known duration*; in other words, its end date is known when the pilot starts (however, pilots can subsequently be extended or shortened). Thirdly, its implementation is restricted to a *fraction of the territory* where it is meant to be rolled out. This last criterion was probably the most difficult to apply. Indeed, whilst most programmes in this dataset were clearly labelled as 'pilots' or 'trials', some other, often small-scale, projects were more ambiguous in terms of the government's intentions. This definition is in line with the British legislation[7].

As noted by Ettelt, there has been a gradual interest in policy piloting under the New Labour governments, specifically between the publication of the 1999 White Paper *Modernising Government* and the 2007/8 fiscal crisis (Ettelt & Mays, 2013). Due to numerous legal, ethical and practical constraints, piloting has been essentially limited to the making of distributive policies, such as employment programmes, schooling programmes, childcare services, rehabilitation programmes for criminals, etc. Distributive policy is moderately prone to conflict and rarely involves primary legislation, as opposed to regulatory and redistributive policies (Lowi, 1972). However its correct implementation is contingent on a network of local agencies. According to Jowell (2003), pilots in the area of distributive policy became so popular after the publication of the 1999 White Paper that they became a norm:

> *For many respondents, however, the decision on whether or not to conduct a policy trial was a matter of opportunity. If it was possible to conduct and evaluate a trial before national roll-out, then it was generally commissioned nowadays more or less as a matter of course. The exceptions were when, say, an indelible manifesto commitment*

[7] http://www.legislation.gov.uk/ukpga/1998/14/section/77/enacted

> *existed in favour of a particular approach, or when insurmountable technical difficulties were likely to arise (Sanderson, 2002; Martin and Sanderson, 1999). In the absence of such obstacles, however, a presumption in favour of piloting new policies seems to be becoming normative in most departments (Jowell 2003).*

4.3.3. Data sources

Interventions are presented and analysed in evaluation reports or studies. These studies are the 'interface' between the intervention and the meta-researcher. Indeed, they often are the only source of information needed for a meta-research. Conveniently, each study usually reports the effect of a single intervention, which means that in many cases, there is a perfect correspondence between the 'intervention' and the 'study'. In the area of biomedical research, these studies are known as 'clinical studies'. In social research, these studies are referred to as 'evaluations'.

Whereas a study usually focuses on one intervention, an intervention can inform multiple studies. For example, the efficacy of an intervention can be evaluated at different points of time and each measurement phase can be reported separately. Furthermore, an intervention can be evaluated from different angles (efficacy, cost-effectiveness, implementation, user satisfaction, etc.) or different teams, which is possible when the data is publicly available. When this happens, the meta-researcher must respect the assumption of statistical independence and make sure the intervention is included only once.

4.3.4. Population

The systematic approach to data collection as well as the limited number of observations (both of which are discussed in section 4.5 below) mean that the sample drawn for this study includes virtually the entire population of pilot interventions conducted in the UK between 1997 and 2010 in the relevant policy areas.

This being said, test statistics and inferences are still useful for two reasons. On one level, test statistics help us assess the plausibility of a partial association (or the lack thereof) in the sample. Low P-values suggest that an effect of the size observed in the sample is substantially plausible. On another level, it can be assumed that the studies in the PILOT dataset are a sample of a 'hypothetical population' comprising other types of policy evaluations, carried out in other policy areas and at different times (e.g. post 2010).

4.4. Search strategy

The selection process is shown in Exhibit 14.

4.4.1. Published evaluation reports

The research started with the identification of all studies commissioned by the relevant government departments (DWP, DfE, HO and MoJ) during the period of reference (May 1997 to May 2010). For this purpose, I searched (1) the DWP's Research website[8]; (2) the DfE's Research & Statistics Gateway[9]; (3) the HO's Research Development and Statistics website[10]; and (4) the MoJ's Research and Evaluation website[11]. These websites were systematically searched, without restriction in terms of publication 'series' (for example, the HO has nine different types of research publications). From this sample, I selected all evaluations and excluded other types of studies (customer satisfaction surveys, scoping studies, evidence reviews, etc.). From this sample, I selected all evaluations of pilot interventions and excluded other types of evaluations. The definition of 'pilot' used in this exercise was presented in section 4.3.2. This decision was made based on the abstracts and introductions of these studies. When several evaluations were conducted on the same policy intervention, this intervention was recorded once and for all to ensure the statistical independence of each unit in the dataset. However, all relevant evaluation studies were kept to provide background information on the pilot intervention.

4.4.2. Unpublished evaluation reports

Despite the government's commitment to publish all publicly funded research, not all reports were found online. There are many reasons why evaluation reports are sometimes withheld. Firstly, the format and content of the published research output remains at the discretion of the commissioning department and releases may be paper-based. This mostly applies to pre-2000 evaluations however. Secondly, departments are not expected to publish research on those rare occasions when publication would "threaten national security, destabilise the economy, or not be in the public interest". Thirdly, the quality of the report might be judged insufficient for publication (Government Social Research Unit, 2010a). The UK Government's Social Research (GSR) Service regularly publishes guidelines for assessing the credibility, rigour and relevance of individual research studies[12]. Fourthly, the study might have been commissioned by a

[8] http://goo.gl/yVdNhJ
[9] http://goo.gl/BjSmR1
[10] http://goo.gl/cpvDYH
[11] http://goo.gl/Syh9B2
[12] http://www.civilservice.gov.uk/networks/gsr/publications

department and evaluated by another (this can be the case when the intervention has implications for several departments). Finally, it could be that a study was published by the research team in a peer-reviewed journal, although this seems to be exceptional.

To limit the risk of publication bias to a minimum, I cross-checked the list of pilots obtained through my own research with two other sources of information. First, I used a number of official documents to identify pilots that would have been planned or conducted but not evaluated or not published. Those included (for the period of interest): (1) the annual reports of all relevant departments; (2) all Budget and Pre-Budget Reports; (3) all Green Papers published by the relevant departments; (4) all parliamentary research briefings published in the areas of interest; and (5) written questions from Members of the Parliament. The research was conducted through automatic searches for the following keywords: "pilot", "pathfinder", "trailblazer", "experiment" and "evaluation". Four documents in particular proved very helpful in gathering information about pilot schemes: three parliamentary research briefings on employment programmes (House of Commons, 2000, 2003, 2005a, 2005b) and the answer of the Secretary of State for Justice to a parliamentary question asking for the list of all external research projects commissioned by the MoJ since its inception in 2006 (House of Commons, 2012).

In addition, I also sent a total of 15 Freedom of Information (FoI) requests to the four relevant government departments as recommended by the literature (Fowler et al., 2013). Departments were asked to provide information about unpublished evaluation reports. When the content of the missing reports was needed to populate the dataset, subsequent requests were made to get access to these reports.

Freedom of Information requests did help me identify a few studies which were not available online. However, the procedure has its flaws. Firstly, it is more appropriate for 'confirmatory' enquiries – that is, when the researcher wants to get hold of a specific study – than for 'exploratory' enquiries – i.e. when the researcher wants to find out the number and nature of unpublished studies. The definition of a 'reasonable' request is too strict to allow broad questions that concern a whole department and its agencies. Thus the researcher must limit his request to a specific unit or bureau. It is possible that some evaluations commissioned outside the 'research' or 'evaluation' units have been missed; however the number is probably low.

Secondly, government departments consider that they are only required to publish final reports. Oftentimes consultants will produce one or several interim reports (especially for the evaluation of large programmes), however it is unclear why some of these reports are accessible online and other not. Interim reports are relevant to my research because they give a first indication of the effectiveness of the programme. Also, they help to see if the intentions of the government are consistently reported (e.g. with regards

to the rollout of the programme). Unfortunately, it was not possible to know how many interim reports were not published or to get hold of the missing reports.

4.4.3. Unevaluated pilots

Whereas there are ways of spotting a study that was carried out but not published, identifying an unevaluated pilot proved somewhat trickier. Two main scenarios need to be considered here. First, the pilot was implemented but not formally evaluated. For example, it could be that results were simply monitored. The second scenario is that of a pilot that never was implemented but that went close enough to implementation to be visible in policy and administrative documents and generate relevant data for the researcher. This was typically the case of the last few pilots planned by the Labour government but terminated by the new Coalition government after the May 2010 coalition. These pilots are far from random and thus needed to be included as well.

Given the heaviness of the procedure, this extra research was carried out for DWP pilots only. Command papers published between May 1997 and May 2010 by the DWP as well as all Budget and Pre-Budget Reports were reviewed to get a list of announced policy changes. This list was matched with the list of research publications available on the web, to estimate the number and nature of (1) evaluations that were not publicly announced; and (2) evaluations that were announced but either not conducted or not published. Four extra DWP pilots were thus added to the list. This list can be found in Annex I.

Exhibit 14 – Selection process

	DWP	DfE	HOME	MoJ	TOTAL
Studies	886	926	557	200	2,569
Including evaluations	331	288	134	83	836
Including pilot studies	218	143	48	54	463
Including single interventions	58	114	26	25	233

4.4.4. Example

To illustrate how the data was collected for this study, I use the Pathways to Work (PtW) programme, which is one of the interventions piloted by the DWP between 1997 and 2010, and briefly discuss the process which led to

its inclusion to my dataset. PtW was an employment programme for claimants of incapacity benefits which was first announced in 2002 and introduced in October 2003 on a pilot basis.

As virtually all DWP 'programmes' (*i.e.* policies involving the provision of a service or financial assistance), PtW was evaluated and thus the first sign of PtW was found in the research database of the DWP. One of New Labour's flagship reforms, PtW was a large and costly programme targeting one of DWP's core target groups so it is not surprising that it was extensively evaluated: no less than 31 separate evaluation studies were first identified, including: 16 implementation studies, nine outcome studies, two cost-benefit studies and four 'lesson-learning' reports. PtW was piloted from October 2003 and in April 2004, and then started being rolled out to the rest of the country.

Being such a prominent programme, it was easy, in the end, to identify PtW as a single pilot scheme and thus as a unit, even if the occasional reference to 'Incapacity Benefit Reform' very slightly complicated the identification. On other occasions, the identification proved much easier (e.g. when a pilot was subject to a single study) or much trickier (e.g. when two studies contained contradictory information on the pilot status of a reform).

Exhibit 15 – Snapshot of the PILOT dataset

	Pilot	START	END1	END2
1	Empowering Young People Pilots	2008/01	2009/03	2009/03
2	Achievement for All	2010/01	2011/06	2011/06
3	Extended Flexible Entitlement for 3 and 4 YO pathfinder	2007/04	2008/09	2008/09
4	Pathfinder UK Online Centres	1999/10	2000/02	2000/02
5	Time to Talk	2007/01	2008/04	2008/04
6	Schools Plus Teams Pilot	2001/09	2002/07	2002/07
7	Aiming High: African Caribbean Achievement Project	2004/03	2005/07	2005/03
8	Transition Information Sessions	2006/06	2008/05	
9	Support Childminder Pathfinder	2003/09	2004/04	2004/04
10	Second Great Parenting Experiment	2006/07	2007/06	2007/06
11	Schools Linking Network			
12	Drug and Alcohol Courts Pilot	2007/09	2010/09	2010/09
13	Pilot Beacon Schools	1998/09	1999/09	1999/09
14	Helping Families Programme	2009/07	2010/03	2010/03
15	Single Level Test Pilot			
16	Early Professional Development	2001/09	2004/07	2004/07
17	Leadership and Management Development Programme (LMPD)	2004/08	2005/07	2005/07
18	Short breaks Pathfinder	2009/		
19	Excellence Fellowship Awards	2002/02	2004/06	2004/06
20	Family Nurse Partnership	2007/04	2008/03	2008/03
21	Trust School Pathfinder	2006/09	2008/09	2008/09
22	14-19 Pathfinder Initiative	2003/01	2005/07	
23	Progress File	1997/09	1998/07	1998/07
24	Re-Ach Project	2006/06	2008/05	2008/05
25	Making Good Progress	2007/09	2009/08	2009/08

Most of the information needed for the present research project was found in these evaluation reports. However, at times, extra research had to be conducted because I could not find what I was looking for. This information

was found in the *Pathways to Work* Green Paper (Department for Work and Pensions, 2002).

4.5. Variables

The series of 17 interviews conducted with policy researchers (already mentioned in section 3.1) was instrumental to identify the key variables influencing the duration of pilots and the selection of pilot sites. In addition, these interviews helped me assess the availability of the data.

For each unit in the dataset, primary information was coded manually based on a variety of administrative sources. Two types of variables were identified. Information pertaining to the research design (*e.g.* research dates, type of research carried out, etc.) and the intervention (*e.g.* target group of the intervention, objectives and policy instrument), was coded primarily based on the evaluation reports when they were available, on other documents when they were not (including technical specifications, secondary legislation, and administrative documents).

Control variables (*e.g.* time remaining before the next election, performance of the government in that policy area) – were built using the above-mentioned sources as well as any other document published by the government, the Parliament and other quasi-governmental organisations. Those included, *inter alia*, press releases, reports of the National Audit Office, Hansard, etc. As many pilots have their own 'brand' (e.g. *New Deal for Lone Parents*, *Beacon Schools*, *Pathways to Work*, etc.), this was, in most cases, possible. More information on these variables, as well as descriptive statistics, can be found in Annex II as well as in the relevant empirical chapters (5 and 6).

4.5.1. Dependent variables

Out of the three dependent variables considered in this project, two only have been subject to statistical analysis: the duration of a pilot and the selection of pilot sites. The third variable (completeness of reporting) has been analysed qualitatively and thus is not discussed in the following section. Descriptive statistics are provided in the relevant empirical chapters.

4.5.1.1. Pilot duration

As recommended in Section 2.6.1, the first dependent variable is the duration of a pilot. The duration of a pilot is defined as the number of months between start and finish. The reported start date for each pilot is that

of the launch of the programme in the first pilot site, as indicated in relevant evaluation reports. The end date proved trickier to establish. Several indicators are proposed in chapter 5, including the actual end date, the planned end date and the 'supposed' end date. Thus different end dates have produced different duration variables for each pilot, all of them measured in months. Distinguishing 'observed', 'planned' and 'supposed' duration was meant to improve the construct validity of the variable and make the notion of duration less 'thick'. The assumptions made regarding the measurement of the 'supposed duration' of a pilot mean that this variable should be considered in an exploratory way. The reliability of these measurements is warranted by the factual nature of the variable.

4.5.1.2. Site selection

In line with the prescriptions of section 2.5.2, the second dependent variable indicates the regions in which the reform was piloted. The exact operationalisation of the variable, as well as the context in which it has been used, are presented in section 6.4. What matters for this chapter is that the variable has two components. The first component is the type of 'area' defining the boundaries of the pilot. This area can be geographical or 'cultural' (e.g. North London, Mercia, Teeside), but most often they will be administrative (e.g. local authority, county, Jobcentre Plus district, probation area). The second component is the list of areas chosen as pilot sites. We will see in chapter 6 that, as a result of the variety in the type of areas, my analysis could only be performed for one policy (employment and welfare). The areas mentioned in the dataset are those reported in the evaluations studies. Extra research had to be carried out, given the incompleteness of many studies. The reliability of the measurement was warranted by the factual nature of the variable.

4.5.2. Independent variables

As already mentioned, there is no single, objective way of measuring the government's commitment to a policy. For that reason, the concept of 'commitment' was measured using three different variables: (1) the stated aim of the pilot; (2) whether the reform was derived from a pre-election pledge; and (3) the seniority of the reform 'champion'.

4.5.2.1. Pathfinders

In line with the prescription of section 2.4.3, the first independent variable used in this study indicates the purpose of the pilot, building on the different typologies found in the literature (Ettelt & Mays, 2013; Jowell, 2003; S Martin & Sanderson, 1999; Sanderson, 2002). To my knowledge, this is the first time that such a variable is used for an empirical purpose.

A dichotomous variable was created to distinguish the pilots, which the government was committed to roll out regardless of the outcome ('pathfinder' pilots) from other, more 'formal' pilots. This question exemplifies the 'black box' problem for three reasons. Firstly, a government might be committed to roll out a pilot, but if the intervention is significantly altered between the pilot and the rollout phase, should we regard it as the same policy? Secondly, can we confidently assume that, when no evidence of commitment can be found, the government is indeed ready to consider all options, including a termination of the policy? Thirdly, is the assumption that the government always 'sticks to its initial plan' a reasonable one? Although contrasting pathfinder and experimental pilots does not solve all the afore-mentioned issues, I think it is the least imperfect measurement of the government's intention, given that the government coined the term specifically to send a message to stakeholders regarding its intentions. The reliability of the measurement is ensured by the fact that virtually all pathfinder pilots are labelled and 'marketed' as such.

4.5.2.2. Manifesto

In line with the prescription of section 2.4.4, the second independent variable used in this study indicates whether the proposed reform was derived from a pre-election pledge.

The operationalisation of this dichotomous variable is similar to that of previous studies (Klingemann et al., 1994; Rallings, 1987; Rose, 1980). A pledge has two components, the first of which is the statement of a specific objective, such as reducing unemployment or increasing school performance. However, these objectives are not enough to identify a pledge, as they can be found in all party manifestos and are very consensual. The second and most important component is the policy intervention chosen to meet the said objective. Policy interventions include taxes, expenditures, regulations, deregulations, etc. They are the value-ridden part of the policy and governments from different end of the political spectrum will be expected to resort to different interventions. Importantly, both the objective and the intervention are needed to identify a pledge. The question of whether a reform was derived from a pre-election pledge was addressed by comparing two documents: the evaluation study, which contained the intervention that was piloted and its objective, and the party manifesto from the previous election.

4.5.2.3. Seniority of the announcer

In line with the prescription of section 2.4.5, the third and last independent variable used in this study was a measure of the seniority of the official announcing the reform. This decision builds on the idea that senior officials

operate for most choices a trade-off between credit-claiming and blame-shifting (Hood 2011). To my knowledge, this is the first time such variable is used in an empirical project.

The variable was created in two steps. Firstly, a taxonomy of 'announcers' was built from the data itself. The sources used to get to this information included, in that order: (1) the evaluation study; (2) policy documents related to the reform, including command papers and the Hansard; and (3) local newspapers, as some pilots were announced during ministerial visits. Although not recorded, the number of cases where evidence was unclear or contradictory was minimal and resolved by an expert judgement of the strength of the evidence (number of concordant sources, credibility of the source). Secondly, an ordinal variable with six categories was created, from the lowest to the highest level of seniority: (1) No apparent announcement; (2) Civil servant; (3) Junior minister; (4) Secretary of State; (5) Chancellor; and (6) Prime minister. The reliability of the measurement is warranted by the factual nature of the variable, which limited the risk of misinterpretation.

4.5.3. Main control variables

The PILOT dataset includes additional variables which have been used as controls. This section presents a few of them. I refer the reader to the empirical chapters for a more detailed presentation.

A first set of variables includes the various research decisions made for each pilot in addition to those used as dependent variables. One of them is the research design elaborated for each pilot using a slightly modified version of the Maryland Scale of Scientific Methods (Sherman et al., 1998). The scale used in the dataset goes from 0 to 5, where 0 indicates studies based on solely qualitative information. In the following chapters, the categories of this variable are collapsed into a dummy isolating the qualitative design from the other designs.

A second set of variables concerns the intervention which was piloted. A policy intervention is defined as (1) a policy instrument (2) targeting a specific population or sub-group (3) with a clearly stated objective. In line with this definition, the following variables have been included:

- A dummy to distinguish interventions that are mandatory from interventions that are voluntary.
- Whether or not the pilot was rolled out.

A third set of variables pertains to the organisational context in which pilots took place. For example, I indicated which of the four government departments commissioned a given pilot: DWP; DfE; HO and MoJ.

4.5.4. Data analysis

Empirically, the PILOT dataset has been used in two ways. First, it provided the data needed for the regression analyses performed in chapter 5 and 6. The type of regression was dictated by the measurement scale of the dependent variable. Thus, event-history analysis was used to model pilot duration and binary logistic regression was used to model the selection of pilot sites. Qualitatively, PILOT was used for the systematic selection of the cases analysed in chapter 7.

4.6. Conclusion

This chapter was set out to propose a research design to assess the extent of confirmation bias in policy research.

The first objective of this chapter was to critically appraise the different methodologies used so far to analyse the effect of institutions on research decisions and to identify the most promising one. The review identified four types of methods including participant and nonparticipant observation, interviews, surveys and meta-research. There are clear disciplinary differences in terms of methods between studies interested in the effect of democratic institutions and those focusing on market institutions. In particular, the latter studies tend to rely more on quantitative methods. The review has concluded that meta-research was the most appropriate design to answer the research question posed in this thesis.

The second objective was to introduce the PILOT dataset. PILOT includes observational data on over 230 policy evaluations conducted by the British government between 1997 and 2010 in three policy areas: employment and welfare; crime and justice and education and parenting. PILOT adds value to existing methods to study the effect of institutions on research in three important ways. First, it analyses this effect in a systematic way, whereas previous research have mainly relied on single or small-N case studies. Second, it relies on structured and factual methodological choices as opposed to verbal accounts of decision-making processes. It also includes data pertaining to the type of intervention that was piloted and to the political context. Third, it looks at three policy areas, whereas previous studies have mainly focused on one.

Having presented and justified the theory underpinning this thesis, the context in which this theory will be tested as well as the methods that will be used to answer the research question, I now turn to the empirical part of this project. The remainder of this document uses PILOT to analyse the extent of confirmation bias in three research decisions: the time allocated to pilots, the selection of pilot sites and the reporting of pilot outcomes. These questions are addressed in chapters 5, 6 and 7 respectively.

5. Effect of policy commitments on pilot duration

5.1. Introduction

The pilot studies conducted by the UK government between 1997 and 2010 to test the effectiveness of its social policy reforms before a possible nationwide introduction present an interesting variety in terms of methodology and approach. For example, whereas an average intervention was given 20 months to demonstrate its worth, some had as few as two months to do the same job and some others up to four years. The complexity of some of these interventions – be it in terms of implementation or in terms of evaluation – are one, perfectly sensible, explanation for this outcome. But the political institutions which commission these studies expose civil servants and researchers to higher-level constraints and incentives which cannot be ignored. If the resources which fuel such studies are provided by organisations which are not policy-neutral, the influence that these resources create on the research process must be examined.

Time is a critical factor in research. On one level, time is a resource given to an agent for the execution of a task. As other resources (budgets, people, expertise), time is scarce and subject to equilibrium effects (the time allocated to a project is taken away from another project). Thus, the amount of time allocated to a pilot can be seen as an indicator of its relative value in the eyes of government officials. On another level, time acts as a moderator variable, i.e. as an independent variable strengthening or weakening the effect of an intervention on its beneficiaries. Thus, an organisation or an individual with a vested interest in the success of such an intervention might find it convenient to interrupt its evaluation earlier than initially planned owing to favourable interim results. This phenomenon, known as truncation (Bassler, Briel, Montori, Lane, et al., 2010), has been studied in medical research. Some have argued that pharmaceutical companies have a financial incentive to truncate their clinical trials and market new drugs early (Trotta, Apolone, Garattini, & Tafuri, 2008).

The goal of this chapter is to answer the question of whether similar mechanisms can be observed in the policy sphere. It builds on the notion of ‘confirmation bias’ (defined in section 2.3) and its possible effect on the duration of research projects (presented in section 2.6.1). In what follows, I address two more specific questions. First, I analyse the extent to which the pilot studies conducted by the British government between 1997 and 2010 abide by two scientific prescriptions regarding the duration of research projects. These prescriptions are (1) the proportionality of this duration with the complexity of the intervention to be evaluated and the complexity of the

evaluation itself; and (2) the strict observance of the pre-defined research timescales. Second, I analyse the effect of policy commitments on these two prescriptions.

The rest of this chapter is structured as follows. In section 5.2, I discuss theoretically – using professional research guidelines as well as the political science literature and previous empirical evidence – why and how one might expect policy commitments to affect pilot timeframes. Section 5.3 introduces the data and methods for my analysis. Section 5.4 lists the hypotheses to be tested. Section 5.5 shows some descriptive statistics. Section 5.6 presents the results of my analysis, first looking at pilot duration, then at early pilot interruptions. Section 5.7 discusses these results in relation to the literature. Section 5.8 concludes.

5.2. Expected effect of policy commitments

As discussed in section 3.6.2, the timeframe of a pilot is determined by the relevant policy teams who would typically consult analysts on research requirements (Boa, Johnson & King, 2010). Given the costs associated with pilots that are too long or too short, policy-makers are faced with an *optimal stopping problem* (Carpenter 2002).

5.2.1. Commitment to scientific norms

The timeframe of a research project is subject to two scientific prescriptions. The first prescription is that the time given to a study be determined solely on the basis of the research question. More specifically, two factors need to be taken into account. The first factor is the type of intervention. Some interventions – such as changes in the school curriculum or health campaigns – may take years or even decades to produce a measurable effect. Other policies are designed to have an almost immediate impact. In any event, most policies take time to bed in and the timetable for their policy trial needs to be adjusted accordingly. Unless the period of the trial is long enough to detect certain impacts, it can create a false impression of policy failure, which would have been contradicted by a later reading (Jowell 2003). The second factor impacting the duration of a pilot is the type of effect to be estimated. Evaluating its effect in terms of access to/take-up of a given service is expected to be quicker than in terms of satisfaction/opinion or behaviour. For example, the *Magenta Book* (2003), which outlines the UK Cabinet's methodological standards for policy evaluation, reckons that RCTs should be given about "two to three years". It can be extrapolated from the above that the more time is given to a study, the more researchers will learn about the effect of the intervention. Longer research projects will allow researchers to estimate the short-term and long-term effects of an intervention whereas shorter research projects will be limited to the former.

The second prescription of science is that, regardless of the duration of a study, the timeline must be carefully planned and executed. In clinical trials, early stopping (or truncation) must be limited to pre-defined cases of extreme *benefit* (the experimental intervention apparently has superior efficacy to the control intervention); *safety* (the experimental intervention apparently has unacceptable adverse effects); or *futility* (there is apparently no prospect of this study showing superior efficacy to the experimental intervention) (Trotta, Apolone, Garattini, et al., 2008). Research guidelines recommend that timing and frequency of interim analyses, as well as early stopping rules, be specified in research protocols (e.g. Consort statement). However, I would argue that early stopping rules do not concern policy evaluation. First, social interventions do not entail a health risk, so early stopping for safety is irrelevant. Second, interrupting a policy evaluation for benefit or futility would be unwise given the complexity of the social sphere and that the effect of social interventions is rarely stable over time. Time variation has been observed in the effect of many policy interventions including employment (Card, Kluve, & Weber, 2010; National Audit Office, 2010) and crime/justice (Martin, Butzin, Saum, & Inciardi, 1999). Thus, I would normally not expect early interruptions of policy pilots. At the very least, procedures should be in place to ensure that pilots cannot be cut at a 'propitious time'.

5.2.2. Commitment to the intervention

A commitment to the intervention is expected to lead to shorter evaluations as research generate important waiting costs for agencies (Carpenter, 2002). These costs are above all political. Patients want a drug for their disease, and firms that profit from drug sales want entry into potentially lucrative markets. To delay the authorisation to market the drug is to impose a cost upon these interests, and when these interests are well organised and influential, they can make it costly for the agency to delay (Carpenter 2002).

Waiting costs depend on several factors. One of the most obvious is the expected benefit of the intervention. The higher the expected benefit is, the higher the pressure will be to make it available to all. This benefit can be assessed from two different perspectives. The first perspective is that of the beneficiary. In the area of clinical trials, this benefit is known as the therapeutic novelty of the intervention. In their study of new drug approved in the US between 1950 and 1986, Dranove and Meltzer (1994) show that more innovative drugs are developed and approved more rapidly than less innovative drugs. The second perspective is that of the drug manufacturer. The Dranove and Meltzer study shows also that drugs with a greater market potential are developed and approved more rapidly than other drugs.

Waiting costs are also affected by the degree of organisation of the beneficiaries of the intervention. Agencies might find it harder to wait when the people affected by the problem are better organised and better funded.

Several scholars have documented the influence of organised patient groups over FDA behaviour (Carpenter, 2002; S. Epstein, 1996; Vogel, 1990).

The newsworthiness of a particular problem also makes waiting politically costlier. Social disorders with greater severity and entrenchment are usually more newsworthy. When media allocate substantial coverage to a problem, then potential solutions receive more attention from the public, stakeholders and politicians (Baumgartner & Jones, 1993). As a result, any delay in the implementation of the new intervention are amplified and the costs of waiting the publish evaluation results rise. Carpenter (2002) showed that the amount of media coverage given to a disease was significantly and negatively correlated with drug approval time.

Finally, waiting costs are influenced by the reputational strategies of the different actors. In the area of clinical trials, this applies first to drug manufacturers. Olson showed that regulators respond to firm-specific characteristics when evaluating new drug applications. For example, firms that are less diversified and more R&D intensive are subject to shorter reviews for their applications than more diversified and less research-intensive firms (Olson, 1997). The reputational strategy of agencies is also an important factor. Carpenter (2002) showed that the FDA will optimise the waiting cost related to the review of an application by weighing the danger of adverse drug reactions from approved drugs against the political cost of delaying the approval of the new drug.

5.2.3. Social policy

The above shows that we have accumulated a significant amount of information on how executive agencies in highly regulated policy areas allocate time to research. In contrast, the view that there might be a 'double standard' in evaluation procedures depending on the political salience of the social policy reform is widely shared but mainly based on hearsay. The considerable variation in the duration of evaluation projects as well as the absence of research protocols or any transparent rule to allocate resources has led some to induce that important non-scientific criteria were at play (Fay, 1996; Jowell, 2003; Sanderson, 2002; Walker, 2001).

The public policy literature has provided ample evidence that was an important constraint for policy-makers. In-depth interviews with senior civil servants and ministers in the UK have shown that the decision to pilot was based largely on pragmatic considerations – the most salient of which was the timeframe available (Boa et al., 2010; Jowell, 2003). The roll-out of many new policies was widely acknowledged to be governed by timetables quite unable to accommodate lengthy policy trials. Once a major new policy had been announced to the public at large, the political and practical momentum in favour of rolling it out nationally – both without delay and without modification – was sometimes impossible to resist (Boa et al., 2010;

Jowell 2003). This problem is just slightly attenuated for the interventions which are granted pilot status. While appreciating the important contribution that early evaluation can make to the development and delivery of new policies, the ministers and policy civil servants interviewed by Jowell also complained that researchers were too seldom willing to recognise how short the optimal time period was in which to roll them out. They were predictably opposed to "the evaluation tail wagging the policy dog", especially as, as one minister put it, "pilots are often seen to give unequal access to benefits for often very deprived people or areas" – a perception that was politically unsustainable for long periods (Jowell 2003).

As predictably perhaps, many researchers interviewed by Jowell put the opposite case, referring to time scales for some pilots that were patently too short to achieve their aims. They argued that, if the very purpose of such pilots was to help refine new policies or practices prior to their national roll-out, there was no point in working to a timetable that was incapable of accurately answering the primary questions being addressed (Jowell 2003). One or two evaluations (not mentioned to protect the identity of respondents) were singled out as examples of unrealistic timetables that had proved to be an embarrassment. By not allowing a sufficient period for the policy to bed in before measuring its impact, these pilots had wrongly presaged a failure of the policy when – as it later turned out – this was not the case (Jowell 2003). These findings confirm a hypothesis made earlier by other researchers (Coleman, 1979; Nathan, 2008).

The aim of my study is to find out whether the above-mentioned results apply to policy research. Specifically, it tests the hypothesis that reforms of greater political 'importance' might be subject to shorter pilots.

5.3. Measuring the duration of a pilot

Contrary to new drug applications (NDAs), which are subject to regulation[13], there is no procedure and no protocol specifying how long a pilot will be. Thus, the 'real' duration of a pilot is something that can be estimated rather than measured.

The duration of a pilot can be estimated in different ways depending on how much credit one gives to formal institutions, decision-makers and processes. The approach taken in this study is to look at researchers' intentions.

[13] In the US, the 1962 Drug Amendment to the Food, Drug, and Cosmetic Act of 1938 define the regulatory standard used by the FDA to evaluate NDAs. Pharmaceutical firms must show, first, that a new drug is safe, and second, that it is effective for its intended use. Furthermore, the 1962 Amendments outlined a multistage process for firms and the FDA to follow and obtain approval for a new drug. The process begins with laboratory and animal studies, and continues with three phases of clinical studies. When the firm completes all of these studies (which can take 8 to 10 years), it compiles all of its evidence and then submits it in the form of an NDA to the FDA for review (Olson 1997).

Accordingly, the duration is the number of months between the start of the pilot and its *planned* termination. The variable makes sense substantively, as one can assume that the duration of a pilot is determined by the political circumstances at the time of – or shortly before – the time of its launch. The planned duration is the dependent variable modelled hereafter.

An alternative measurement would have been to report the date when the differentiated treatment of pilot and non-pilot sites terminated. This information is the one that is the most consistently given in evaluation reports. However, the *observed duration* can be misleading as it does not necessarily reflect the initial intentions of policy-makers. By the time of their termination, some pilots will have been extended; others will have been shortened for reasons that were unknown at the time when the decision was made.

The problem with the above-mentioned indicators is that they take the duration of a pilot at face value and assume that policy-makers will wait until the end of the pilot to consider a possible roll-out. This might not always be what happens. Walker (2000) reported about the New Deal for Disabled People – a pilot conducted in the late 1990s by the Department for Work and Pensions – that the intention of policy-makers had always been to make a decision regarding a possible roll-out half-way through the two-year pilot period and before the results of impact analyses were available. Some authors have also made similar points about different reforms (Chitty, 2000; White & Dunleavy, 2010). Against this background, the *supposed duration* of a pilot could be defined as the number of months between its launch and the publication of the first evaluation report. This makes a difference only for the longer pilots, for which interim evaluation reports are often commissioned at an early stage.

However, this indicator is not without problem either, as it rests on the assumption that these reports are made publicly available shortly after their presentation to policy-makers. However, I have some evidence that it might not always be the case. Firstly, exploratory research showed that the delay between the implementation of the pilot and the publication of first results can vary significantly from one department to the next. Secondly, I know that some reports can be subject to occasional publication embargos from the relevant departments (Metcalf, 2008). In fact, it has been argued that the control over the acceptance of contract deliverables and the timing of their release was the client's main weapon in influencing the content of research reports (Metcalf, 2008). Whether they are due to embargos or genuine discussions between research commissioner and researcher, such delays suggest that the report's content might be 'bended' in ways that threatens objectivity.

5.4. Hypotheses

My first hypotheses build on the knowledge that more salient interventions are subject to shorter research procedures than less salient interventions (see section 5.2.2). An intervention can be considered politically salient if the government has publicly expressed its intention to roll out the reform. This type of pilots, known in the UK as 'pathfinders', is considered by many closer to prototyping than to experimenting (Sanderson 2002; Jowell 2003; Walker 2000; Boa, Johnson, King 2010; Martin and Sanderson 1999). Therefore, I would expect that such reforms were subject to shorter pilots. To test this idea, I formulate the following hypothesis:

H1: Other things being equal, pathfinder pilots are shorter than formal pilots.

Another way of assessing a government's commitment to a specific reform is to check whether it was announced in the ruling party's manifesto for the previous election (Rose, 1980). When policy initiatives arise from election manifestos, policy-makers are impatient to receive results that will provide evidential support for decisions to proceed with full implementation. Such a political interest potentially conflicts with the interests of evaluators, the interests of which are served long-term, in-depth analysis of the effects of pilots (Jowell 2003, Sanderson 2002, Walker 2002). Against this background, it is unlikely that an office-seeking government would commit large resources to pilot a measure that contributed to its electoral success. I will test this idea based on the following hypothesis:

H2: Other things being equal, pilots that are directly related to a pre-election manifesto pledge are shorter than other pilots.

The relationship between citizens and elected representatives is the core concern of democratic theory and elections are typically assigned the principal role in structuring this relationship. They are a means by which the public can make governments accountable and influence policy directions. In institutions with strong political control of the bureaucracy and adequate incentives, this pressure for accountability trickles down to civil servants, who are encouraged to take a specific course of actions. In such circumstances, the political pressure exerted by an office-seeking government on the civil service is expected to be much greater towards Parliament's end of term. A study in Brazil found that the approval of environmental licenses varied according to the electoral cycle and distributive politics motivations. In years of gubernatorial elections, more environmental licenses were approved, especially in municipalities with a large presence of loyal voters to the governor. In years of mayoral elections, the approval rate is larger where the mayor belonged to the same party as the governor (Ferraz, 2007). This responsiveness of bureaucrats to elected institutions has also been demonstrated in other settings (Coate, 2002; Frye

& Mansfield, 2004). The effect of the electoral cycle will thus be tested through the following hypothesis:

H3: Other things being equal, pilot duration is positively correlated with the amount of time remaining before the next general election.

It should be said that hypothesis H4 needs to be taken with a pinch of salt when it comes to Britain's New Labour. Arguably, the outcomes of the 2001 and 2005 elections were never really in doubt. So the 2010 election was really the first one since 1997 that the Conservative Party had a good chance to win. This insight might have had an impact on the duration of pilots in Labour's third term. To test that idea, I formulate the following hypothesis:

H4: Other things being equal, pilots launched during Labour's third term are shorter than those launched in the two previous terms.

In particular, I would expect the time before the next election to be a particularly salient issue in Labour's third and last mandate. Therefore, I have added an interaction term between these two variables.

H5: Other things being equal, the effect of the time remaining before the next general election on pilot duration is stronger during Labour's third term in government.

Hypotheses 6 and 7 concern the DWP only as data could not be collected for all departments.

Hypothesis 6 builds on the idea that the time afforded to research varies according to the target group (Carpenter 2002). More 'important' target groups are expected to be associated with shorter pilots. This is relevant for the DWP, which ranks its 'customer groups' using a point-based system. Against this background, I formulate the following hypothesis:

H6: Other things being equal, pilot duration is negatively correlated with the symbolic importance of the target group.

Hypothesis 7 builds on the idea that the time afforded to research varies according to its expected benefit (Dranove & Meltzer 1994; Olson 1997). More 'beneficial' interventions are expected to be associated with shorter pilots. This is relevant for the DWP, which administers both voluntary and mandatory employment programmes. Although evidence of the economic effectiveness of mandatory work programmes is mixed (Department for Work and Pensions, 2012), there is wide cross-party support in favour of mandatory work programmes and benefit sanctions in the UK (Grice, 2012). Against this background, I formulate the following hypothesis:

H7: Other things being equal, mandatory employment programmes are subject to shorter pilots than voluntary programmes.

My final hypothesis concerns pilot truncations. Bearing in mind that there is no strong scientific justification for the interruption of a policy pilot, but that there might be political benefits in it, I hypothesise that pilots to which the government is strongly committed are more likely to be truncated than pilots to which it is weakly committed. Given the limited amount of data, this hypothesis will be tested in a qualitative, non-inferential way. Hypothesis 8 reads as follows:

H8: A pilot is more likely to be stopped early when the government is committed to the policy and when interim results show an apparent benefit.

This study includes two control variables. First, the optimal duration of a pilot depends on the level at which one wants to evaluate the intervention. Evaluations of pilot processes (including inputs/outputs, and client/implementer's *experiences)* can be undertaken as the programme is being rolled out. Thus, they can be relatively short. Evaluations of 'soft outcomes' (employability, attitude in school, attitude to crime, etc.) would typically require the programme and its participants to 'mature', so they are expected to be longer than process evaluations. Evaluations of 'hard outcomes' (effect of the intervention on employment status, school performance, recidivism, etc.) will be the longest to conduct, as they involve substantial data collection. To capture this idea, I included a categorical variable for the type of effect with three values.

Second, the duration of a research project is contingent on financial and human resources, as well as the expertise available within an organisation. I use the department commissioning the evaluation as a proxy for this notion.

5.5. Data and methods

The following analysis is based on the PILOT dataset presented in chapter 4, which includes 233 pilots spanning 13 years (1997-2010) and three ministerial departments (DWP, DfE and HO and MoJ).

The duration variable indicates the lapse of time, in months, between the reported start of a given pilot and its planned end date. I will show later that, although one would intuitively conceive duration as a continuous, interval-level variable, in this specific case, it was treated as a discrete variable recorded in months. This was justified by the limited number of duration values.

Chapter 4 also describes the key independent variables that will be used throughout this thesis. As a reminder, I considered *pathfinder* pilots as a proxy for the government's commitment to a reform and constructed the corresponding dummy variable.

A reform was considered a manifesto commitment if both the objective and the intervention were mentioned in the Labour Party manifesto of the previous general election.

The Election variable captures the number of months between the start of a pilot and the next general election. I also created a dummy variable for each of the three terms that Labour spent in government (1997-2001; 2001-2005; 2005-2010).

The type of evaluation conducted as part the pilot was measured using an ordinal variable with three categories: (1) process evaluation, (2) outcome evaluation; and (3) impact evaluation. Process evaluations address the effect of the intervention on daily operations, including the take-up of the intervention by target group, the burden on staff, etc. Outcome evaluations are concerned with changes in attitudes and dispositions at one point in time (e.g. the employability of jobseekers). Impact evaluations are those which measure the effect of the intervention using a counterfactual and/or several measurement phases. When several studies were available for the same pilot, I recorded the highest-order effect.

The commissioning department was coded as mentioned on the report.

I have used the scale used by DWP to prioritise its customer groups. According to this scale, Jobcentre Plus earns 12 points for the placement of a lone parent or a person with a disability, 8 points for the placement of a long-term unemployed and 1 point for the placement of a person already in employment.

Finally, I created a dummy to distinguish interventions that are mandatory to target groups from interventions that are not. This information was collected from evaluation reports.

The duration of a pilot given the government's commitment to the intervention was modelled using event history analysis. The technique models the 'hazard' of an event, that is to say that an event occurs at particular time given that it has not occurred before that time. Event history models, sometimes called duration models or survival models, originate from biomedicine where they are used to model how observed variables (such as smoking) are associated with the amount of time from a starting point such as a treatment to an event such as death (see for example (Box-Steffensmeier & Jones, 2004) for an overview). These models are now frequently used in the social sciences. For example Dranove and Meltzer

(1994) used event history analysis to model the timing of a drug's approval given its scientific and commercial importance.

Event history analysis is appropriate when the data are subject to censoring, that it to say when a subject leaves the study before an event occurs, or the study ends before the event has occurred. To be clear, there is no censoring in the data used for this study. However, duration models are still appropriate and make fewer assumptions than linear regression models regarding the distribution of the dependent variable. Normality, in particular, is not required.

For the purpose of this study, the 'survival time' of a pilot is considered discrete as opposed to continuous. Each month of pilot implementation is coded 0 until the month when the pilot terminates (it is then coded 1). Thus, the equations presented in the results section are of the logistic kind and model the odds of a pilot terminating after a duration of a specific number of months, given that it has not terminated before and controlling for a number of other variables.

5.6. Descriptive statistics

The collection and description of data on pilot duration reveals a number of findings that are noteworthy. The first is that this information is sometimes not available. In more than 10% of cases, I failed to find the precise start date and/or the end date of the pilot, despite the extra research carried out in policy and administrative documents and the media. In fact, if I had limited my research solely to evaluation reports, the number of missing values would have been significantly higher. This shows that important information is often not reported in government-sponsored evaluations. This observation concerns all government departments and is not limited to the question of pilot duration. Chapters 6 and 7 will confirm that government evaluation reports rarely allow replication or research synthesis.

The second finding concerns the duration of pilots. Exhibit 16 shows some descriptive statistics for the *planned* duration of pilots. I find an average duration of 20 months with a standard deviation of 11 months and a range going from two months to four years. The frequency distribution shows a clear pattern: over a fifth of pilots (51) has a planned duration of exactly 24 months. The 2nd most frequent value is 12 months (41 pilots) and the 3rd is 36 months (18 pilots).

The PILOT dataset gives 50 occurrences of pilots for which the *observed* duration is not equal to the *planned* duration (21%). Out of those occurrences, 14 were due to a missing value (six missing planned durations; eight missing observed durations) and 36 were due to an early or late stop of the pilot. Out of those, 25 were extended and 11 were shortened.

Exhibit 16 – Frequency distribution of pilots per planned duration (N=207)

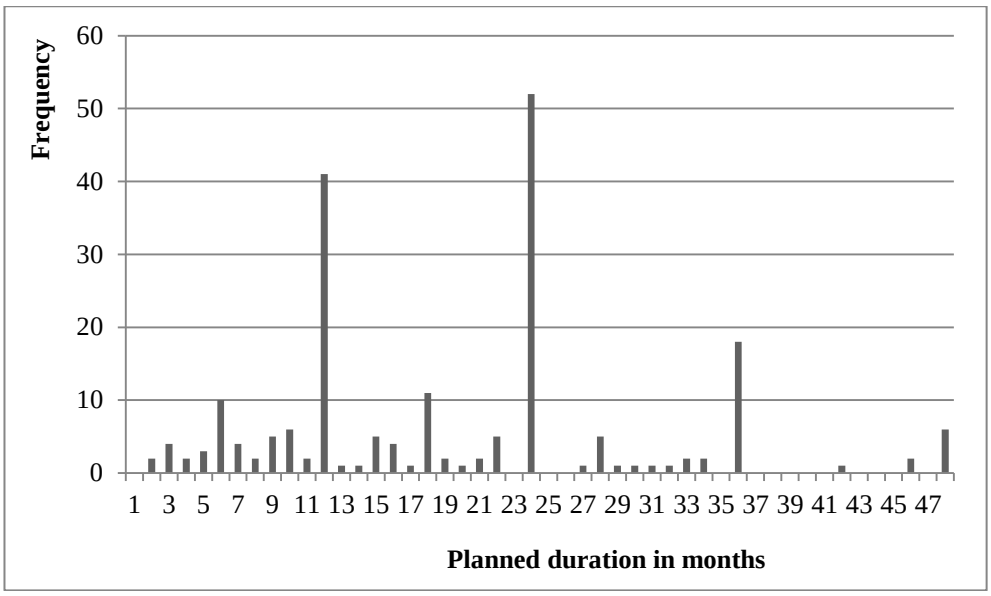
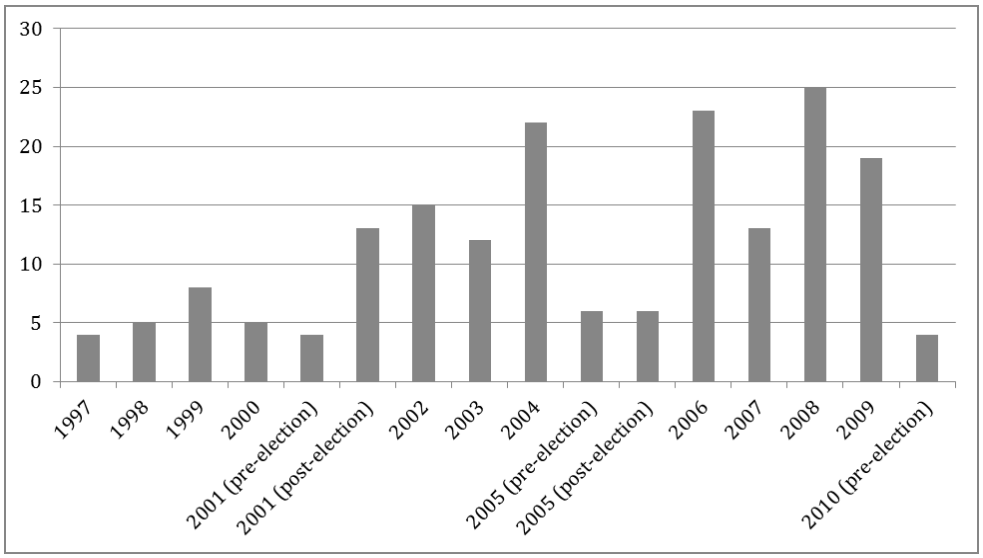


Exhibit 17 – Number of new pilots launched per year



Looking at the distribution of the number of pilots per year (see Exhibit 17), we can see an upwards trend between 1997 and 2010. This trend is most obvious when one considers the number of new pilots launched between two elections: 26 during Labour's first term (May 1997 to June 2001); 68 during its second term (June 2001 to May 2005); and 90 during its third term (May 2005 to May 2010). The distribution also indicates that the number of new pilots launched does not seem to go down as the next election gets closer: indeed, out of the four years which saw the launch of the highest number of pilots, two were pre-election years (2004 and 2009).

Descriptive statistics for all variables can be found in Exhibit 18 below.

Exhibit 18 – Descriptive statistics (pilot duration models)

Variable	N	Min	Max	Mean	SD	Freq
Duration	217	2	48	19.8	10.6	--
Pathfinder	233	0	1	--	--	56
Manifesto	233	0	1	--	--	112
Election	221	0	58	25.4	14.8	--
Term 1	229	0	1	--	--	38
Term 2	229	0	1	--	--	78
Term 3	229	0	1	--	--	113
Process	215	0	1	--	--	112
Outcome	215	0	1	--	--	46
Impact	215	0	1	--	--	57
Department – DfE	233	0	1	--	--	114
Department – DWP	233	0	1	--	--	58
Department – HOME	233	0	1	--	--	26
Department – MoJ	233	0	1	--	--	35
Mandatory intervention*	54	0	1	--	--	20
Target group*	50	1	12	8.2	3.1	--

** Available for the DWP only*

5.7. Results

5.7.1. Planned duration

The results of the statistical analysis for the four departments are presented in Exhibit 19. The event modelled here is the 'hazard' of termination, i.e. the probability that a pilot terminates at month *m*, given that it has not ended before and given the independent variables introduced in section 5.4. As a reminder, the models are of the binary logistic type and a positive regression coefficient for an independent variable means that the hazard of termination is higher, and thus the pilot shorter. Three different specifications are proposed. Model A focuses on the association between my various

indicators of policy commitment and the planned duration of pilots. Model B adds the variables related to the electoral cycle. Model C is a more parsimonious proposition.

The first six variables (in grey) in the output are part of the *intercept* of the model, i.e. the value of the dependent variable if all parameter coefficients were equal to 0. This information is commonly reported but rarely interpreted, given the lack of 'substantive' meaning. Thus I will not comment these lines. Besides, each cell of the output contains two lines: the first one represents the coefficient as a log odds ratio and the second line the same coefficient as an odds ratio (i.e. the result has been exponentiated). In this section and in the rest of the document, I will rather use the latter coefficient. The interpretation of the coefficients is as follows.

Hypothesis 1 states that pathfinder pilots are shorter than formal pilots. There is strong evidence that it is true: all three models indicate a positive and statistically significant partial association between the two variables. The hazard of a pilot terminating at any time is between 3.03 times and 3.52 times higher for pathfinders than for formal pilots. In other words, this hazard is between 203% and 252% higher for pathfinders than for pilots. This result is strongly significant across models. Hypothesis 1 cannot be rejected.

Hypothesis 2 states that reforms directly related to manifesto pledges are associated with shorter pilots. The data shows that the opposite is true: controlling for other variables, being a pre-election pledge multiplies the odds of the hazard of a pilot terminating at any time by between 0.76 and 0.84. In other words, it decreases the hazard of a termination by between 24% and 16%. However, the effect is not statistically significant. The duration of pilots is evidently not influenced by election pledges and hypothesis 2 can be rejected.

Hypothesis 3 states that the duration of a pilot is positively correlated with the time remaining before the next general election. Model 2 shows that it is not the case. Indeed, each additional month to the next general election multiplies the odds of the hazard of a pilot terminating at any time by 1. Unsurprisingly, this effect is not statistically significant. Hypothesis 3 can be rejected.

According to hypothesis 4, pilots launched during Labour's third term in government (2005-2010) are expected to be shorter than those launched in the two previous terms. The baseline in Exhibit 19 is Labour's first term. The results show that, controlling for other variables, being launched in Labour's second term multiplies the odds of the hazard of a pilot terminating at any time by 0.58 compared with a pilot launched during the first term (they are 42% lower). This result is not significant. Likewise, being launched in Labour's third term multiplies the odds of the hazard of a pilot terminating at any time by 0.55 compared with a pilot launched during

the first term (they are 45% lower). This result is not significant either. Hypothesis 5 can be rejected.

Hypothesis 5 states that the effect of the time remaining before the next general election on pilot duration was stronger during Labour's third term in government. Model 2 shows that, each extra month before the next general election multiplies the hazard of a termination by 1 when this pilot was launched in Labour's first or second term and by 0.99 in the third term. Unsurprisingly given the results above, none of these results are statistically significant. Hypothesis 6 can be rejected.

Exhibit 19 contains two control variables. First, the research question addressed in the evaluation report was included, based on the hypothesis that longer-term questions would require longer pilots. Models 1 to 3 analysed the partial effect of process evaluations, outcome evaluations and impact evaluations separately. I found that, on average, an outcome evaluation would multiply the odds of a pilot terminating at a certain time by 0.44 (i.e. reduce them by 56%) compared with a process evaluation. Likewise, an impact evaluation multiplies the odds of a pilot terminating at a certain time by an average of 0.66 (i.e. reduce them by 34%) compared with a process evaluation. In other words, both outcome and impact evaluations are longer than process evaluations; however the difference is only statistically significant for outcome evaluations. Model 3 offers a more parsimonious model, whereby outcome and impact evaluations are merged into a single category. The model concludes that, controlling for other variables, conducting an impact or outcome evaluation multiplies the odds of a pilot terminating at a certain time by 0.59 (i.e. reduces them by 41%) compared with a process evaluation. This result is significant at the 5% level.

Second, I analysed the effect of organisations on pilot duration, based on the hypothesis that pilots commissioned by the Home Office or the Ministry of Justice would be shorter than those of the DWP or the Department of Education. I found that, controlling for other variables, being commissioned by the DWP multiplied the odds of a pilot terminating at a certain time by between 0.80 and 1 compared with the DfE. Conversely, controlling for other variables, being commissioned by the Home Office or MoJ multiplied the odds of a pilot terminating at a certain time by between 1.02 and 1.33 compared with the DfE. Although the direction of these effects is in line with my initial hypotheses, none of them are statistically significant.

Exhibit 19 – Pilot duration models (all departments)

Method: Duration model for discrete data
NB: In each cell of the output, the first line represents the coefficient as a log odds ratio and the second line the same coefficient as an odds ratio.

	(A)	(B)	(C)
Time	.06** *(1.06)*	.06** *(1.06)*	.06** *(1.06)*
Time6	1.22** *(3.38)*	1.27** *(3.56)*	1.23* *(3.42)*
Time12	2.87** *(17.63)*	2.98** *(19.68)*	2.86** *(17.63)*
Time18	1.56** *(4.75)*	1.61** *(5.00)*	1.56** *(4.75)*
Time24	3.97** *(52.98)*	3.93** *(50.90)*	3.95** *(51.93)*
Time36	4.04** *(56.82)*	4.07** *(58.55)*	4.05** *(56.82)*
Pathfinder	1.26** *(3.52)*	1.11** *(3.03)*	1.15** *(3.15)*
Manifesto	-.27 *(0.76)*	-.26 *(0.77)*	--
Election	--	.00 *(1)*	--
-- Term 1 (b)	--	--	--
-- Term 2	--	-.54 *(0.58)*	--
-- Term 3	--	-.59 *(0.55)*	--
Election x Term 3	--	3.21e-06 *(0.99)*	--
-- Process (b)	--	--	--
-- Outcome	-.82** *(0.44)*	-.80** *(0.44)*	--
-- Impact	-.36 *(0.69)*	-.44 *(0.64)*	--
Outcome/Impact	--	--	-.52* *(0.59)*
-- DFE (b)	--	--	--
-- DWP	-.19 *(0.82)*	-.22 *(0.80)*	-.00 *(1)*
-- HOME_MoJ	.03 *(1.33)*	-.02 *(1.02)*	.09 *(1.09)*
N	2532	2458	2532

** p<0.05* *** p<0.01* *(b) Baseline*

5.7.2. Factors specific to DWP

Exhibit 20 provides two extra models based on DWP data only. Model D is identical to Model C (but with a focus on the DWP). Model E adds the variables related to the target group and the degree of constraint of the intervention. The interpretation of coefficient remains unchanged.

Across the models, the effect of pathfinders remains strong and significant. However, the effect of the type of study on pilot duration weakens when only DWP data is considered and ceases to be significant.

Exhibit 20 – Pilot duration models (DWP only)

Method: Duration model for discrete data
NB: In each cell of the output, the first line represents the coefficient as a log odds ratio and the second line the same coefficient as an odds ratio.

	(D)	**(E)**
Time	.07** *(1.07)*	.09** *(1.09)*
Time6	1.47 *(4.35)*	2.03** *(7.61)*
Time12	3.08** *(21.76)*	3.48** *(32.46)*
Time18	1.76** *(5.81)*	1.98** *(7.24)*
Time24	4.60** *(99.50)*	4.64** *(103.54)*
Time36	4.39** *(80.64)*	4.15** *(63.43)*
Pathfinder	3.65** *(38.47)*	3.63** *(37.71)*
Target group	--	-.01 *(.099)*
Mandatory intervention	--	.23 *(1.26)*
Outcome/Impact	-.37 *(0.69)*	-.47 *(0.62)*
N	822	818

** p<0.05* *** p<0.01* *(b) Baseline*

Hypothesis 6 states that target groups with a higher level of priority are subject to shorter pilots. The evidence shows that the effect of target group is very small and insignificant. Hypothesis 6 can be rejected.

Hypothesis 7 states that mandatory interventions will be subject to shorter pilots than non-mandatory interventions. The evidence shows that the 'direction' of the effect is as anticipated. Controlling for other variables, piloting a mandatory intervention multiplies the hazard of a termination by 1.26 compared with a voluntary intervention, i.e. increases them by 26%. However, this effect is not statistically significant. Hypothesis 7 can be rejected.

5.7.3. Early stopping

Hypothesis 8 states that a pilot is more likely to be stopped early when the government is committed to the policy and when interim results show an apparent benefit.

The observed duration of a pilot was shorter than the planned duration in 11 instances. All of them were interrupted further to the May 2010 election, which saw a change of majority in Parliament. Thus, the evidence suggests that political commitments can lead to early pilot termination, but not in a way that could have been predicted by the literature on clinical trials. It was the change of policy strategy that triggered the early termination of these pilots rather than the will to see them rolled out as soon as possible.

It is difficult to know precisely what role evidence played in these decisions. However, there are some indications that, in this particular circumstance, it was marginal. Out of the 11 pilots interrupted because of the election, four were cancelled by the incoming government before they had even started. In five other cases, the change of government resulted in the rescoping and the scaling back of the evaluations, which suggests that the decision to interrupt the pilot had little to do with results. For example, the evaluators of the Child Development Grant pilot reported that:

> "The methodology changed significantly following the announcement to bring forward the close of the CDG pilot. Original plans for longitudinal interviews with parents were not carried forward" (Child Development Grant Evaluation, p.ii).

In one case (*Right2BCared4 pilots*), it was not possible to identify whether, on the whole, the intervention had had a positive or negative effect and whether this effect was significant.

In only one case (*Find Your Talent*), the pilot seems to have had a positive effect on the target group. As indicated in the evaluation report:

> "The findings from Year 1 and 2 demonstrate early evidence of programme additionality in several areas, as observed and reported by partners and stakeholders". (Find Your Talent Evaluation, p.27),

However; the policy was not rolled out, which shows that the decision to interrupt the pilot was not related to the findings. Against this background, Hypothesis 8 can be rejected: in policy research, early stopping is unlikely to be used to favour the intervention over the counterfactual.

5.8. Discussion

The findings of this study agree with the rest of the literature in three ways. Across the models, the most robust finding is that reforms which the government is committed to roll out (pathfinders) are significantly shorter than 'formal' pilots. This result is noteworthy: it confirms that reforms which the government intends to roll out are less 'researched' than other reforms, as previously written (Chitty, 2000; Sanderson, 2002; Walker, 2000; White & Dunleavy 2010).

The second finding counterbalances the first one. It shows that the scientific prescription according to which the duration of a research project should be dictated by the research question is respected: I found that, overall, process evaluations are significantly shorter than outcome/impact evaluations (when considered as a single category). The slightly surprising finding of the lack of significant difference between outcome and impact evaluations can be explained. A number of evaluation reports indicated that the government had initially planned to evaluate the impact of the new intervention. However formal and informal studies later on concluded to the infeasibility of an impact evaluation owing to insufficient sample size, inappropriate research design or lack of data. Less 'ambitious' research designs have sometimes been adopted without affecting the duration of the pilot. This is exemplified in the evaluation of the *Time to Talk* programme, which indicates that:

> "The original research design included a focus on outcomes as measured by questionnaires to the participating parents before and after their involvement. However, because of delays and challenges in recruiting staff, implementing the programme locally and engaging parents, the research design was adapted, with the agreement of the Teenage Pregnancy Unit, to one based on qualitative interviews, supplemented by basic quantitative data" (Davis, Cullen, Davis, & Lindsay, 2010, p.6).

Third, I found out that pilots were sometimes interrupted early after instructions given by ministers, as suggested in the medical literature. However – and this is a key difference with the medical literature – those interruptions did not aim to favour the intervention. All interruptions happened further to a change of government and the evidence suggests that little attention was given to the evaluation results. This finding is interesting, as it gives ground to the 'symbolic' role of research in policy

(Weiss, 1979). According to this theory, evaluation is less used to inform policy decisions than to justify decisions that have already been made on the basis of ideas, professional experience, self-interest, organizational interest, a search for prestige, or any of the multiplicity of reasons that go into decisions about policy and practice. Besides, this finding triggers the interesting question of the effect of changes of government on the type of evidence used in policy-making.

This chapter also offered some more surprising results. First, I found no significance difference in the duration of pilots based on whether or not they were related to a pre-election pledge. This could be explained by the fact that some 'formal' pilots were announced in election manifestos. Presenting an intervention as a pilot (and thus, somehow, as a 'policy experiment') does not commit the government to a specific decision. In fact, the government might even take credit for its pragmatism – at least as long as it concerns a limited number of uncontentious reforms. Conversely, the financial crisis that hit Britain in 2008 triggered a number of pilots and reforms – many of them fairly salient – that could not be foreseen at the time of the previous general election in May 2005.

Second, I showed that the duration of a pilot was not affected by the electoral cycle. This result is somewhat troubling as it contradicts the strongly held view that the timetables of research conflicts with the timetables of politics. It could be that, for a few specific reforms, timing could indeed have been an issue, giving the impression to some that this was a widespread problem. In other words, timing could be a punctual problem but in the grand scheme of things, it is not.

5.9. Conclusion

This chapter was designed to examine whether politically salient reforms were subject to shorter pilots than reforms with lower salience.

The first objective was to analyse the extent to which pilot studies fulfil the two scientific prescriptions that concern timeframes. Looking at the proportionality of timeframes, I found that whilst government officials are generally subject to greater time pressure than researchers working independently or in academia (which can be explained by the political cycle, scarce resources or both), they still allocate resources such as time based on research considerations. Whereas the mean duration (20 months) would seem credible in relation to the professional guidance given by government's scientists (Magenta Book), about a third of pilots lasted for 12 months or less, which raises some questions regarding the type of research that can be done in so little time. Generally, I found that there was a great amount of variation in the duration of pilots (from 2 to 48 months). Having said that, the data shows that, overall, the duration of pilots was proportional

to the type of research undertaken, with impact evaluations given more time than process evaluations. Unfortunately, it was not possible to assess whether more complex interventions were subject to longer pilots, as recommended by the research methods literature. Furthermore, I found that very few pilots had been truncated (about 5%). All of the truncated pilots were interrupted following the change of government in May 2010, which thus excludes the possibility that these interruptions were decided for convenience.

The second objective of this chapter was to investigate whether pilots to which the government was committed were associated with shorter pilots. I found limited evidence of an association. The data shows that, controlling for other variables, pathfinders are significantly shorter than formal pilots. However, there is no significant difference in terms of duration between interventions related to a pre-election pledge and interventions which were not. In addition, I did not find that the duration of pilots was influenced by the electoral cycle, as expected. This is an important finding, which contradicts the rest of the literature and most of the anecdotes heard on the subject. I also did not find that the duration of pilots was influenced by the political 'salience' of the intervention as measured by the target group or the fact that the intervention may be sanctioned. Finally, the fact that all the truncated pilots were interrupted following the change of government in May 2010 can be interpreted in opposite ways. On the one hand, it can be seen as supporting the idea that policy U-turns (or changes in policy commitments) have an effect on the duration of pilots. The results of an existing pilot might be embarrassment for a new government with different policy priorities. On the other hand, this can be seen as a specificity of democratic systems, whereby the most desirable policies are those which are supported by a majority, not necessarily those that are the most effective.

Unfortunately, I have some reasons to believe that this study did not fully answer the question asked in the introduction. First, this study did not control for the complexity of the intervention. Yet it is one of the main reasons why a pilot might be longer. It is easy to understand why pilots requiring deep institutional or organisational changes – such as the merger of the Employment Service and the Department of Social Security into Jobcentre Plus – need more time to 'bed in' than simple adjustments to training programmes. None of the variables tested for that purpose proved satisfying. Second, it was not possible to control for the 'size' of the pilot. Here again, the underlying hypothesis was that 'larger' pilots would require more 'piloting time' than smaller pilots. The number of pilot participants was too inconsistently reported across evaluation reports to be included in my models. The same remark applies to the number of areas or units of delivery (schools, jobcentre plus offices, courts, etc.). Third, the reader needs to bear in mind that the duration of a pilot can only be *estimated*. The real duration remains a 'black box' issue unknown of the public.

[This page was intentionally left blank]

6. Effect of policy commitments on pilot site selection

6.1. Introduction

Just as the amount of time afforded to a pilot can tell us about the respective influences of professional and political logics on research decisions, the distribution of pilot sites across a country can provide information about the goals of policy research. In the previous chapter, I remarked that the variation in pilot duration could not only be explained by the type of research. In particular, this decision was found to be associated with the government's commitment to the reform. Armed with this knowledge, I can now re-examine the question of the selection of pilot sites. For that purpose, I will look more specifically at the welfare-to-work pilots run by the DWP through Jobcentre Plus, which is organised in 40 districts across England. I will try to answer this question: how can it be that, whilst an average Jobcentre Plus district hosted about 10 pilots between 1997 and 2010, some of them were selected only three times and some others up to 23 times? As in the previous chapters, both scientific and 'real-world' considerations will be controlled for, so I can best estimate the partial effect of the political logic.

The objective of the following chapter is not simply to describe the process leading to the selection of pilot sites but to explain *why* I see what I see. Our central hypothesis – presented in detail in section 3 of chapter 2 – is that, using information on the political salience of some of these pilots, researchers will select districts in a way that minimises the uncertainty of the pilot outcome. In other words, politically salient pilots are expected to be tested in districts chosen for their exemplarity, rather than for their representativeness. As in the previous chapter, I aim to build the statistical model that best explains this decision.

The rest of this article is structured as follows. The first section discusses theoretically – using professional research guidelines as well as the political science literature and previous empirical evidence – why one might expect a pilot site to be *representative* in the scientific logic, as opposed to *exemplary*, in the political logic. Section 2 briefly presents the findings from interviews conducted with policy researchers in the 'feasibility' stage of this study. Section 3 introduces the data and the variables for the logistic regression. Section 4 discusses the hypotheses. Sections 5 and 6 present descriptive and inferential statistics respectively. Finally, section 7 discusses the findings.

6.2. Expected effect of policy commitment

I briefly discussed in chapter 2 why sampling could be an issue in applied research projects and came to the conclusion that the degree of representativeness of a research unit (both at individual and regional level) could be a meaningful indicator of their responsiveness to market and political institutions. The following section looks more specifically at policy pilots and summarises what I know on the subject.

6.2.1. Commitment to the scientific method

According to the *Magenta Book* (*The Magenta Book: Guidance for evaluation*, 2011), there are two main threats to the external validity of a pilot. The first case is when those affected by a pilot are not representative of the wider population. For example, if a policy is only piloted in parts of London, it would be unwise to assume that the observed effects would be the same in other parts of the country. A well-designed pilot study would address this by including a variety of different types of area. Even so, it is unlikely to be an exact representation of the whole population. Where it is possible to quantify how the pilot areas differ from the country as a whole, it may be possible to correct for this bias. This can be particularly valuable if the choice of pilot areas (or participants) is constrained, for example, if there is a greater than average representation of urban areas in the pilot. A second and more difficult case to deal with is where the pilot areas (or people, or units) are self-selecting, for example, if local authorities were asked to volunteer to participate. In such cases, the generalizability of the pilot findings to areas that are compelled to participate in a later implementation stage cannot be assumed.

The only way to be *certain* that the results of an evaluation represent the behaviour of a particular population is to ensure that the units are randomly selected from that population. For example, out of the 433 local authorities in England, each one should have a probability of 1/433 to be selected for a given pilot if indeed the intervention is provided at the level of local authorities. Importantly, the method does not need to be a simple random sample, however it should include some kind of randomness (multistage sample, cluster sample, etc.). Regardless of the method used, it is considered essential that the roll-out schedule be not correlated with the outcome. For example, the performance of local agencies should not be used for sampling purposes.

There are implications for our research. If the purpose of the study is to use the results from the pilot sites to derive conclusions that will apply for the whole country, then the set of pilot sites should be reasonably representative of the country. In statistical terms, this means that a probability sample should be used, *i.e*. a sample where each unit has a known, non-zero chance

of being selected. Conversely, any kind of sample that does not rely on probabilistic properties is prone to a certain type of bias. Just like normal conditions will lead to normal results, exceptional conditions lead to exceptional results.

6.2.2. Commitment to the intervention

From an institutional level, pilots have an important property: they introduce a temporary geographical variation in the administration of a given jurisdiction. The UK for example had to change its legislation in 1996 to allow different levels of benefits to be tested across the country. From a political point of view, this differentiated treatment creates opportunities both for the 'central' and the 'local' policymaker.

6.2.2.1. Pilot sites as exemplars

Although I have already made this point several times in this study, it is worth repeating this crucial fact: public policies in the UK – and probably in many other countries – are evaluated by the organisations that conceive and implement them. This institutional set-up makes the evaluation of these policies unlikely to be value-free. Studies finding that a specific reform has had an insignificant effect – or a negative one – on the group it was targeting will put the government in a rather uncomfortable situation (Bovens et al., 2008). If the government has a vested interest in showing that its policies have a positive effect, then decisions such as the choice of pilot sites do matter. I would indeed expect policy-makers to use their knowledge about these sites to select those which are most likely to generate a positive outcome.

Pilots with a relatively high level of political salience are virtually indistinguishable from gradual reforms. Gradual reforms are deep and far-reaching policy changes which are implemented in sequences, one group at a time, and in the economic literature they are opposed to 'big-bang' reforms. The literature on gradual reforms – which was first developed to explain the successes and failures of reforms in transition countries – has shown the importance of building constituencies for reform through appropriate sequencing. This usually means implementing a reform on the most compliant groups first and taking advantage of favourable exogenous events to roll out the reform to other, more resistant groups. Sequencing gives governments the opportunity to maximise the probability of moving forward to the next stage of the reform process (Dewatripont & Roland, 1996).

Pilots can be seen as strategies to build constituencies for reform. If the objective of the pilot is not to yield generalizable results but as a way of building internal support for change and overcoming resistance, it is better

not to look for representative sites, but rather to look for *exemplars*. According to this line of thinking, pilot sites should not be randomly drawn, but selected with a purpose. Hasluck suggests that such reasoning might not be absent from policy-makers' decisions and that "the resources devoted to the pilot may exceed that available at national roll-out" (Hasluck, 2000). Some authors have seconded that point of view, alluding that the use of pilots could be more akin to prototyping than to experimentation (Brodkin & Kaufman 1997; Walker 2001). Others have warned against the "structural danger of unrepresentativeness" of pilots in a context where there is a strong political commitment to a policy, and the pilot receives generous resourcing (Sanderson, 2002).

There are some empirical results supporting this claim. In his study of an environmental programme implemented in Madagascar, the *Opération Menabé* Pilot, Billé (2010) observed that the pilot region had been chosen amongst other reasons because of the existence of a Regional Development Committee, unique in Madagascar and without any official existence in the national politico-administrative system. This committee was considered reliable, energetic and with strong leadership under the authority of a motivated local dignitary (Billé, 2010). Billé concluded that the first step in setting up a pilot experiment is usually to identify a space (territory, community, sub-basin, administrative unit, etc.) in which the conditions before the intervention seem favourable enough to offer the best promise of success (Billé, 2010).

1.2.2. Clientelism

The second opportunity offered by pilots is clientelism, *i.e.* the exchange of goods and services for political support. Clientelism may arise when two conditions are met: (1) the piloted intervention brings an obvious immediate benefit to its target groups, such as cash payments and other financial incentives; and (2) the piloting phase is long.

There is anecdotal evidence of clientelism in the selection of pilot sites. Jowell (2003) reported that the pilot of the Education Maintenance Allowance, which tested different models and levels of monetary reward for young people to stay on at school, had created intense political pressure, especially from MPs in constituencies bordering the pilot sites. Similar anecdotes were reported by one of our interviewees (see chapter 4). Rogers-Dillon came to a similar conclusion in her analysis of the Welfare-to-Work reform launched in 1996 by US President Bill Clinton. The Personal Responsibility and Work Opportunity Act (PRWORA) gave US states the power to pilot time-limited benefits. Rogers-Dillon observed that the Act created a win-win situation, whereby governors were afforded political power in exchange for their support of the reform (Rogers-Dillon, 2004).

This type of selection, though morally questionable, does not necessarily affect the findings of the evaluation, as long as the sampled sites are not correlated with the outcome of interest. It is not addressed in the rest of this chapter.

6.3. An account of sampling decisions at the DWP

The following section reports on a series of 17 interviews conducted with policy researchers between October 2011 and February 2012. The aim of these interviews was to test the feasibility of a quantitative study of sampling decisions, as well as to collect a maximum of information on how this decision was made. Two types of sampling decisions were discussed: (1) the sampling of regions; and (2) the assignment to conditions in the case of studies with control groups.

6.3.1. Decision makers

There was a consensus among respondents that sampling decisions did not follow a formal, invariable procedure. Having said that, all agreed that it was primarily the competence of policy-makers. Some knew that the decision was made after consultation of key groups including analysts and Jobcentre Plus District (JCPD) Managers. There was also an agreement on the fact that external evaluators were never involved. According to one interviewee, "the set-up of the pilot, including the selection of sites, is brought on a plate to the contractor". With respect to a possible involvement of elected policy-makers, the majority of respondents could not rule it out, but in the meantime no one had evidence to support one claim or another (see below).

6.3.2. Opinion about the use of probability mechanisms for the choice of sites/individuals

Questioned about their perception of the meaning and relevance of probability sampling among stakeholders, all interviewees agreed to say that the notion was highly divisive.

Politicians were perceived to have ambivalent positions on the subject. One the one hand, the need to pilot a new intervention in a set of regions that is representative of the country as a whole was felt to be understood. Occasionally, MPs would even hold the government to account on this issue, as shown by the following statement, found in a Select Committee Report:

> "We understand the reasons why the present pilot areas were chosen, but the Government will need to bear in mind during the evaluation the fact that the pilot areas are not fully representative of the country as a whole. We recommend that, even at this late stage, the Government should give consideration to adding a pilot area which covers a predominantly London Area or Northern City geographical type." (House of Commons, 1999a).

To which the government responded:

> "The Government notes the Committees' concerns. We are confident that the pilot areas are sufficiently representative of the country as a whole for us to make sound estimates of the national impact of ONE. The selection of the pilot areas was determined primarily by the need to ensure that the pilots covered a range of labour markets and demographic characteristics, and the areas selected (such as Lea Roding and Leeds) include characteristics of concern to the Committees such as deprivation and representation of ethnic minorities (…). Adding another pilot area at this stage would increase substantially the cost of the pilots, and would be impractical at this stage, without significantly increasing the depth or robustness of the evaluation." (House of Commons, 1999b).

On the other hand, at a more micro level, the use of probabilistic formulas to assign individuals to conditions was perceived by interviewees as problematic for ministers and interest groups, who tend to find the method unethical. Likewise, mid-level and frontline bureaucrats tend to find methods like randomisation unfair and occasionally imposing extra burden on public services without prior discussion. Conversely, some interviewees reported that probability mechanisms had strong proponents among the economists working for the Treasury. One of them said that Treasury officials were often ready to trial any kind of intervention, provided it was evaluated with experimental methods. For example, the Employment Retentions and Advancement demonstration was above all a research project testing the feasibility of large-scale RCTs in the UK. Treasury and DWP officials had to identify a low-profile intervention to make it happen.

Among the community of researchers, feelings about probability mechanisms were thought to depend on the department to which researchers were affiliated as well as their academic background. For example, one interviewee reported that RCTs were highly regarded at the Department for Health, which employs a high number of individuals with training in medical research. In fact, this interviewee reported that the use of randomisation had been the condition of the Department's participation to the Job Retention and Rehabilitation Pilot initiated by the DWP. Conversely, researchers at the Department for Education (which employs a large number of analysts with qualitative research skills) were found more

sceptical about the usefulness of experiments and quasi-experiments. DWP analysts were found to be in the middle.

6.3.3. Selection criteria

There appears to be no single sampling method used and probably no 'pure' method used either. The interviewees indicated that three factors were considered when selection pilot sites, with different levels of influence depending on circumstances.

The first consideration spontaneously mentioned is normative: site selection has a direct impact on the validity of results. However, there was a disagreement among interviewees as to what criterion was taking precedence. Some said that sites had been selected with a view to maximise internal validity. In other words, regions with the highest number of people belonging to the target groups were more likely to be selected than others. This view was mostly held by researchers having evaluated pilots for well-defined groups such as lone parents. For others, the selection of pilot sites was motivated by the idea of generalizability or external validity. It had to cover a range of different characteristics including geography (North/South…), urban/rural, economic structure of the region, etc. One interviewee mentioned the importance of not having any other pilot running for the same target group. Some respondents pointed the presence of quotas. For example, London, which is in many respects different from the rest of the country (due to its cost of living, its high proportion of migrants and the dynamism of its economy), will almost always be included in a national pilot.

The second consideration mentioned was more practical or managerial. According to one respondent, "some Jobcentre Plus District Managers are more compliant to pilot and cooperate with DWP than others" and this was view as an important factor in the implementation of pilots. Another one seconded that statement, citing the example of Manchester, which is a recurrent pilot site for that very reason. Overall, practical considerations were found important by a majority of respondents. One of them concluded that the sum of all criteria to fulfil made the process of site selection "probably not very scientific".

Last but not least, two researchers drew our attention to the fact that the selection of sites had gone through a call for expression of interest to private or public service providers. This was the case of the Job Retention and Rehabilitation Pilot and the Work-Focused Services in Children's Centre pilot.

6.3.4. Evidence of political influence

Interviewees disagreed on whether political considerations had influenced the selection of sites for the pilot they were involved in. For about half of respondents, a political bias of any kind in the selection of pilot sites was unlikely. Respondents felt that the pilot they had evaluated had too low a profile to trigger ministerial interest.

6.4. Data and methods

6.4.1. Principles

As already mentioned in the introduction to this chapter, our analysis is based on a subset of the PILOT dataset presented in chapter 4, namely the pilots conducted by the DWP. Although a cross-policy analysis would have desirable, such study would have very complicated to carry out. The problem is that the 'administrative geography' of the four government departments included in the full PILOT dataset is not comparable. For example, the DfE works mainly with the 152 local authorities (LAs) of England, however these LAs are free to pilot new interventions jointly or at a higher administrative level. The Home Office coordinates the work of the 43 police forces in England and Wales. On its end, the MoJ 'manages' 42 probation areas (which are coterminous with police force area boundaries but are served by 35 Probation Trusts), 650 HM Courts and Tribunals as well as 130 HM prisons in England and Wales. Lastly, the DWP relies on its network of 40 Jobcentre Plus districts (JCPD) in England for the administration of unemployment benefits and on local authorities for the administration of the Housing Benefit.

6.4.2. Jobcentre Plus districts as units

Given the objective of this study as well as the above-mentioned constraints, it was decided that I would focus on the pilots run by the DWP. Fifty of them were identified in chapter 4.

In terms of the geographical units, I used Jobcentre Plus districts to define the regions where pilots were conducted. This decision was motivated by the fact that pilots are often implemented at this level. The number of such districts changed over time. I use the 40 distinct districts of England (version prior to April 2011) to be able to match data items consistently across datasets. Other classifications were also considered such as the 93 NUTS 3 regions of England, the 152 local authorities, the 354 districts or the 850 or so JCP offices, however JCPDs were considered to be the best option.

For each of the 50 pilot schemes run by the DWP, pilot sites were selected from identical pools of 40 districts (N=2000 pilot-districts).

The problem with Jobcentre Plus districts is that few indicators are aggregated at this level. So the data collection work involved the construction of a pool-up table matching each local authority (counties and districts) with NUTS 3 regions and JCPDs.

6.4.3. Operationalisation

An exemplar pilot site is a region where a pilot is most likely to produce a positive effect, be it in terms of process (*i.e.* new rules are applied 'by the book'), outcome (*i.e.* customers get new qualifications) or impact (*i.e.* customers find a job), thereby acting as 'role-model' for other regions. Thus, the exemplarity of a region may refer to two distinct traits: behaviour and performance.

6.4.3.1. District managers as local policy entrepreneurs

From a mid- and street-level bureaucrat point of view, pilot implementation is not different from policy implementation. It does not matter much for local civil servants that the policy be delivered for a limited time (given the perpetually changing policy framework) and in small number of districts. Therefore, pilot implementation is likely to be affected by the same shirking behaviours that characterise policy implementation (Lipsky, 1980). Against this background, some JCPDs might be better 'test beds' for new interventions because local management and procedures have a reputation for being supportive of the policies made centrally in Whitehall. JCPD managers play a key role in this respect, as they are the links between the government and the street-level bureaucracy and thus can act as policy entrepreneurs.

A policy entrepreneur has been defined as an individual "who exploited an opportunity to influence policy outcomes in order to maximize his/her self-interests, without having the resources required for achieving this goal alone" (Cohen, 2011). This influence is usually exerted through networking, a contribution to policy-making and the building of coalitions (Mintrom, 1997). JCPD managers are in a key position to play this role, which is why their support is so important to the success of a pilot. Some authors have highlighted the greater level of commitment and the 'pioneering spirit' amongst staff involved in pilots (Hasluck, 2000). As noted by Billé:

> "One of the fundamental parameters often taken into account is the presence of key individuals, talented and charismatic leaders thriving towards innovation (Saunders, 2003). Later on, the anticipated up-scaling of the experience is hindered by personalities less driven by

> innovation, less motivated and less prone to change, be it out of lack of conviction, for reasons of personal agenda (such as career opportunities), because of decisions on the allocation of available resources, or others." (Billé, 2010).

Unfortunately for this study, I have very little information about JCPD managers. A Freedom of Information request was sent to the DWP in September 2012 to get the list of all district managers since the creation of Jobcentre Plus as well as the permission to contact them. However our request was rejected on the ground that the requested information was not readily available. Furthermore, district managers were not authorised by DWP to answer my questions (DWP, personal communication).

6.4.3.2. District performance

Another way of looking at the notion of exemplarity is to consider the relative performance of a given JCPD. In fact, this performance is closely monitored by the DWP through a series of six indicators including the 'Job Outcome'. Job Outcome is a point system measuring the number of JCP customers who move into work, whether through a referral by an adviser or one of JCP contracted providers or via self-service channels. When there is a match, the job outcome is converted into points depending on the customer group. The higher the priorities of the customer, the more points are achieved. For example, helping an unemployed lone parent into work earns a JCPD or office 12 points, whereas helping an employed person change job will give it only one point. Every year, new targets are established centrally by the DWP for each district and office based on previous performance and labour market circumstances. At the end of the year, a job outcome performance is measured in terms of percentage against target. However, a discussion with a DWP official revealed that this type of indicator is very volatile. Furthermore, rules and definitions seem to have changed several times since the introduction of JO targets in 2006.

6.4.3.3. Favourable labour market conditions

The above-mentioned performance monitoring system used by the DWP is based on the proportion of benefit claimants moving into work. As the targets set by the government to Jobcentre Plus are unknown for most of the period under consideration, I use the absolute value as proxy for Jobcentre Plus district exemplarity. In other words, a new intervention is more likely to produce quick and positive results in a district with a fluid labour market than in a district where conditions are not as favourable. More specifically, I use the Jobseeker Allowance (JSA) exit rate to jobs as indicator (Nunn & Jassi, 2010; Riley, Bewley, Kirby, Rincon-Aznar, & George, 2011). To reduce noise, the value included in the dataset is the annual average JSA exit rate to job of a given district the year before the start of the pilot. To the

extent that exemplarity matters, implementing a new intervention in district where the labour market is fluid is probably the government's best way of making the new intervention 'look' successful. The other advantage of using off-flow rates is that it is an indicator commonly used by JCP, as indicated by a DWP analyst in a non-recorded discussion.

6.5. Hypotheses

Piloted interventions tend to attract the attention of politicians, the media and stakeholders. If the pilot seems to work well, the government is unlikely to get credit for it; however if it goes wrong, the government is likely to be blamed (Hood, 2011; Weaver, 1986). Therefore, it is reasonable to think that the probability of a successful implementation is a criterion among others in the selection of pilot sites. This probability is positively correlated with the degree of exemplarity of a given district. To test this idea, I formulate the following hypothesis:

H1: Other things being equal, there is a positive association between the favourability of labour market conditions in a Jobcentre Plus district and its probability of being selected as pilot site.

I have shown in chapter 4 that the government's commitment to a reform could influence some research decisions such as the time afforded to a pilot. Against this background, we would expect the government to pay even greater attention to the level of Jobcentre Plus exemplarity when the pilot is a pathfinder, which the government is committed to roll out (Sanderson 2002). To test this idea, I formulate the following hypothesis:

H2: Other things being equal, the effect of JCPD exemplarity on the probability of being selected as pilot site is greater when the pilot is a pathfinder

Likewise, pre-election pledges are expected to make the government more anxious to deliver.

H3: Other things being equal, the effect of JCPD exemplarity on the probability of being selected as pilot site is greater when the reform originates from an election manifesto.

Bearing in mind the discussion regarding the politics of welfare-to-work programmes, I would expect the government to pay greater attention to JCPD exemplarity when the piloted programme is mandatory, and thus politically contentious. Likewise, I would expect greater care in the selection of pilot sites when the intervention targets high-priority DWP customer groups. To test these ideas, I formulate the two following hypotheses:

H4: Other things being equal, the effect of JCPD exemplarity on the probability of being selected as pilot site is positively correlated with the degree of priority of the target group.

H5: Other things being equal, the effect of JCPD exemplarity on the probability of being selected as pilot site is higher when the programme is mandatory.

Some welfare-to-work interventions are delivered by private-sector organisations, local authorities, charities or consortia in which Jobcentre Plus may or may not be involved. It can be argued that it is easier for government to shift the blame in case of failure when the implementation of a programme is not led by a government agency such as Jobcentre Plus.

H6: Other things being equal, the effect of JCPD exemplarity on the probability of being selected as pilot site is higher when Jobcentre Plus is lead implementer.

This study includes a number of controls. Firstly, I know from evaluation reports and interviews that the proportion of benefit claimants is always an important factor in the selection of pilot sites. However, the use of this indicator varies from one pilot to another. Sometimes policy-makers would look for variation/contrast (priority given to external validity). On other occasions, they would rather select districts with a high proportion of claimants (priority given to internal validity). Against this background, I control for the proportion of Income Support (IS), Jobseeker Allowance (JSA) and Incapacity Benefit (IB) claimants in the active population.

Secondly, geography matters. As indicated in evaluation reports and interviews, policy-makers tend to select pilot sites from different parts of the country. For example, UK-wide pilots will very often include at least one region from each country (England, Northern Ireland, Scotland and Wales depending on the intervention and the corresponding competence of the national government in that policy area). This study focuses on England so I have aggregated the nine English regions into four larger regions: North, Midlands, South and London.

Thirdly, demographic variables have been included. Those include the population of the Jobcentre Plus, its population density and the proportion of ethnic white people in the adult population.

Fourthly, I would expect a negative association between the number of pilots already running in the district and the probability of seeing this district chosen for a new pilot. Therefore, I control for the capacity of the JCPD at the start of the pilot, *i.e.* the number of pilots already running in the district.

6.6. Descriptive statistics

Considering our limited knowledge of the processes leading to the selection of pilot sites, I found it useful to present some descriptive statistics before a more in-depth analysis.

My first observation is that the location of pilot schemes is not always reported and the selection criteria seldom indicated (see Exhibit 21). Out of the 50 pilots included in this study, I reviewed the evaluation reports of the 45 that were implemented (Π=0.9). For 35 of those, I got a complete list of the districts in which the intervention was piloted (Π=0.83). Missing values were obtained from other documents (legislation, policy briefs, media). Out of these 35 pilots, 21 listed the selection criteria (Π =0.5). Still, among those, it was sometimes difficult to understand how the selection criterion was applied (looking for variation or large number). There is no more systematic answer to this question, which makes this paper even more relevant.

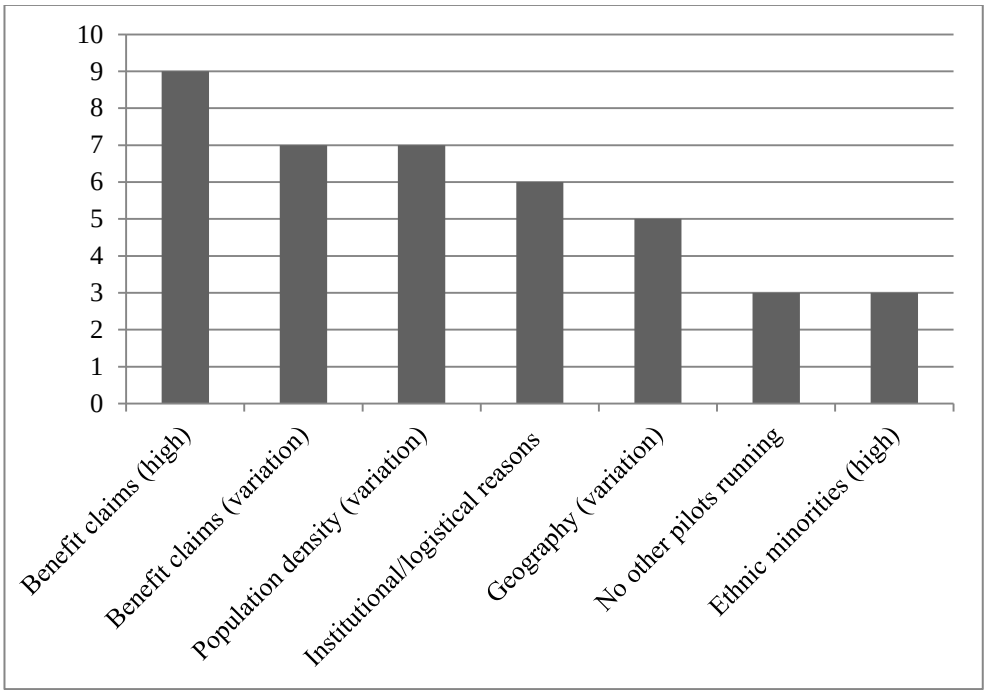
Exhibit 21 – Published information

	Numerator	Denominator	Proportion
Number of pilots	50	--	--
Number of implemented pilots	45	50	0.9
Number of published evaluations	42	45	0.93
Number of studies listing pilot districts	35	42	0.83
Number of studies listing selection criteria	21	42	0.5

My second observation is that there is no set protocol for the selection of pilot sites. A small number of criteria seem to be used to make the decision but the exact formula changes from one pilot to another. Importantly, the selection of pilot sites is never random. Exhibit 22 below shows the type and frequency of selection criteria used in the 21 pilot studies for which I found this information. It shows that the number of benefit claims in a given district is the most recurrent selection criterion used by the DWP. However this indicator is not always used in the same way, depending on whether researchers sought internal validity or external validity. For example, whilst in nine pilots, researchers looked for districts with a high number of benefit

claimants; in seven others they selected districts with different levels of benefit claimants. Variation in the population density of pilot sites is another important criterion, followed by logistical and institutional constraints and geographical considerations. Some studies mentioned the absence of other pilots running in the district for the same target group as an important element. Lastly, in three pilots, researchers were interested in testing a new intervention in districts with a high proportion of ethnic minorities. These findings are broadly in line with the comments made by the policy researchers I interviewed.

Exhibit 22 – Criteria used for the selection of pilot sites (frequency distribution, based on 21 studies)



In terms of the distribution of pilots across the country, the data shows that, out of 2000 *possible* pilot-districts, I have 411 *effective* pilot districts. This means that, on average, an intervention was piloted in eight districts and that a district hosted an average of 10 pilots (as shown in Exhibit 23).

Exhibit 23 – Frequency distribution of pilot-districts (N=411)

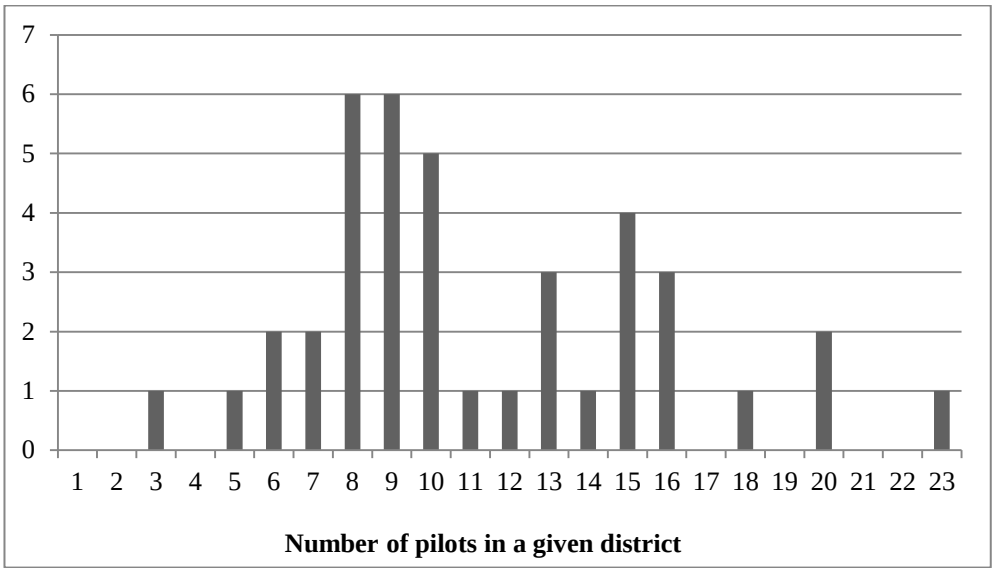


Exhibit 24 – Descriptive statistics

Variable	N	Min	Max	Mean	SD	Freq
Pilot	2000	0	1	--	--	411
Region – North (b)	2000	0	1	--	--	550
Region – Midlands	2000	0	1	--	--	400
Region – London	2000	0	1	--	--	450
Region – South	2000	0	1	--	--	600
Pathfinder	2000	0	1	--	--	360
Manifesto	2000	0	1	--	--	680
Mandatory	1880	0	1	--	--	800
JCP lead	1800	0	1	--	--	307
JSA exit rate (%)	1960	3.98	31.11	16.34	3.11	--
Benefit claimants (%)	2000	1.3	7.3	3.08	1.43	--
Working age population (in 100,000)	2000	3.98	15.37	7.72	2.75	--
Population per ha (in 10)	2000	0.1	26.8	2.92	5.24	--
Ethnic white (%)	2000	57	96	85	10.45	--
Capacity	2000	0	6	0.94	1.16	--
Target_DWP	2000	1	12	8.56	2.92	--

6.7. Results

The results of the binary linear regressions are displayed in Exhibit 25. Eight different specifications have been tested. Model 1 focuses on the association between our measure of fluidity and the odds of being selected as pilot site. Model 2 considers only the selection criteria mentioned in evaluation reports and in the interviews. Models 3 to 7 all include the measure of fluidity and the control variables but each of them tests a different interaction, as per the hypotheses presented in section 6.4. In model 3, the interaction is the government's commitment to the policy and the measure of fluidity. The interaction in model 4 is that of performance and fluidity. In models 5 and 6, I focus on target groups and mandatory programmes respectively. Model 7 tests the hypothesis that pilots implemented by Jobcentre Plus will increase the effect of fluidity on a district's likelihood of being selected as pilot site. Model 8 is a more parsimonious proposition.

Hypothesis 1 states that high-performing JCPDs are more likely to be selected as pilot sites. Five models out of six show that the opposite is true: indeed the odds of a district being selected as pilot site decrease by between 1% and 10%) for each additional percentage point of performance depending on the specifications. This result is statistically significant in models 1 (at the 1% level) and 7 (at the 5% level) only. Thus, hypothesis 1 is rejected. To the extent that there is an effect in the population, this effect is more likely to be negative.

Hypothesis 2 states that the effect of a district's performance on its odds of being selected as pilot site will be greater when the pilot is a pathfinder. Model 3 shows that, when there is no commitment to the reform, the odds of being selected as pilot site decrease by 1% for each additional percentage point in performance when the pilot is not a pathfinder and decrease by about 5% when the pilot is a pathfinder. The interaction is not statistically significant. Hypothesis 2 can be rejected.

Hypothesis 3 states that the partial effect of a district's performance on the odds of being selected as pilot site is higher when the pilot originates from a pre-election pledge. Model 4 tests that hypothesis and shows that the odds of being selected as pilot site decrease by 1% for each additional percentage point in performance when the pilot does not originate from a manifesto and increase by 3% when it does. Given that these results are not significant, I reject this hypothesis.

According to hypothesis 4, the partial effect of a district's performance on the odds of being selected as pilot site is higher for high-priority JCP customers, using the DWP Job Outcome Points table. This hypothesis is tested in model 5, which shows that the odds of being selected as pilot site decrease by 9% when the pilot tests an intervention for non-priority target

groups (baseline) and increase by 1% for each additional point in the JO point grid used by the DWP. This result is not significant.

Hypothesis 5 states that the partial effect of a district's performance on the odds of being selected as pilot site is higher when the pilot tests a mandatory programme. Model 6 shows that the odds of being selected as pilot site increase by 2% for each additional percentage point in performance when the pilot does not test a mandatory labour market intervention and decrease by 3% when it does. These results are not significant; the hypothesis is thus rejected.

Finally, hypothesis 6 states that the partial effect of a district's performance on the odds of being selected as pilot site is higher when Jobcentre Plus is lead implementer. To test that effect, model 7 adds the appropriate interaction term. It shows that, controlling for the other variables in the model, the odds of a district being selected as pilot site decrease by 10% for each additional percentage point in performance when Jobcentre Plus does not implement the pilot. This result is significant. Conversely, when Jobcentre Plus is in charge of the pilot, each additional percentage point in performance increases the odds of a district to be selected by 4%. However, this result is not significant. This last hypothesis is rejected as well.

Models 2 to 8 include control variables, which are the known criteria used by the DWP to select its pilot sites. Results across models are highly consistent.

I will first consider the proportion of benefit claimants in the working age population of a JCPD. My review of evaluation reports and the interviews carried out as a preliminary stage of this study had given a blurry picture of how this criterion was used in the selection of sites. In some circumstances, DWP would look for diversity; in others it would mainly select districts with a high proportion of claimants, supposedly to increase the internal validity of the evaluation. In models 2 to 7 the partial effect of each additional percentage point in the proportion of benefit claims on the odds of being selected as pilot site varies between -1% and +11% but with 5 models out of six showing a small, positive effect. None of these results are statistically significant.

In terms of geography, and holding everything else constant, districts in the North of England (the baseline) are more likely to be selected as pilot sites than districts in any other part of the country. The difference is particularly large and statistically significant for the London and southern districts. For example, the odds of a southern district to be selected as pilot site are, depending on the specifications, between 49% and 62% lower than for a northern district, controlling for other variables. This cannot only be explained by the respective size of each region. To understand this result, it is important to remind the reader that the North of England comprises 11 districts, the Midlands 8 districts, London 9 districts and the South 12

districts. If one district from each region was selected as pilot – as implied by some evaluation reports and interviews, a given southern district would have an 8% chance of being selected and a northern district a 9% chance. If our assumption was true, the odds of a southern district to be selected as pilot site would be expected to be 0.92 those of a northern district, *i.e.* only 8% lower. More strikingly, the odds of a London district would be expected to be 1.22 those of a northern district, *i.e.* 22% *higher*. Thus, there seems to be an additional reason for piloting in the North of England, even when other known selection criteria are controlled for.

The demographic variables used as controls include the working age population, the population density as well as the population of ethnic white people in the working population. Here again, the results are pretty consistent across models. First, I notice a positive association between the working age population and the odds of being selected as pilot site. Indeed, for each additional 100,000 people in the working population, the odds of a region being selected increase by 6% and 8% depending on the specifications. The effect is significant at the 1% level in six out of seven models. Likewise, districts with higher population density have a better chance of being used as pilot site. On average, an increment of 10 people per hectare in a given district will increase the odds of this district being selected by 5%. This result is significant at the 1% level in all but one specifications. The likelihood of being selected as pilot site decreases when the proportion of ethnic white people in the working population increases. The effect (3% to 5%) is consistent and significant across all models.

Finally, I find that capacity matters. Each additional pilot run in a district at a given time decreases the odds of the next pilot being implemented in that district by between 10% and 12%. This result is statistically significant in all but one models.

Exhibit 25 – Probability of being selected as pilot district

- Binary logistic regression
- Y = PILOT
- Coefficients are odds ratios

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
JSA exit rate (%)	0.94**	--	0.99	0.99	0.91	1.02	0.90*	--
Benefit claimants (%)	--	1.08	1.08	1.09	1.11	1.07	0.99	--
Midlands	--	0.96	0.94	0.95	0.95	0.91	0.93	0.89
London	--	0.31**	0.31**	0.32**	0.31**	0.27**	0.18*	0.23**
South	--	0.51**	0.49**	0.51**	0.51**	0.51**	0.38**	0.44**
Working age population (in 100,000)	--	1.08**	1.08**	1.08**	1.08**	1.07**	1.07	1.06**
Population per ha (in 10)	--	1.05**	1.05**	1.05**	1.05**	1.05**	1.02	1.05**
Ethnic white (%)	--	0.97*	0.97*	0.97*	0.97*	0.96*	0.95*	0.96**
Capacity	--	0.90*	0.89*	0.90*	0.90*	0.89*	0.88	0.90*
Pathfinder	--	--	1.14	--	--	--	--	--
Commitment x JSA exit rate	--	--	0.95	--	--	--	--	--
Manifesto	--	--	--	0.83	--	--	--	--
Manifesto x exit rate	--	--	--	1.03	--	--	--	--
Target_DWP	--	--	--	--	0.77*	--	--	--
Target_DWP x exit rate	--	--	--	--	1.01	--	--	--
Mandatory	--	--	--	--	--	1.77	--	--
Mandatory x exit rate	--	--	--	--	--	0.97	--	--
JCP lead	--	--	--	--	--	--	57.40**	--
JCP lead x JSA exit rate	--	--	--	--	--	--	1.04	--
N	2000	1960	1960	1960	1960	1840	1760	2000

** p<0.05*
*** p<0.01*

6.8. Discussion

In the case of the UK pilots, none of the hypotheses predicting the selection of pilot sites in connection with the salience of the intervention has been confirmed. I do not claim that purposive selection never occurs; I note only that I find no evidence in this case. Given the relative weakness of the evidence base, different interpretations can be given.

The idea that the selection of pilot sites might be influenced by policy commitments can be rejected, as shown by the lack of significant effect of both manifesto pledges and pathfinders. This latter result is more surprising given that the government's commitment to a reform was found to be a strong predictor of pilot duration.

The salience of the reform – in terms of target group and the 'severity' of the intervention – does not affect the selection of pilot sites either. This finding can be interpreted in two different ways. The first interpretation is that the choice of pilot sites has too little political salience to be visible through a 'thick' quantitative design. Group dynamics might be at play, I just cannot see them. The second hypothesis is that the political logic plays a significant role in the selection of pilot sites, albeit in a different way. I refer the reader to the earlier discussion on the benefit of implementing the pilot in constituencies where the government has strong or opportunistic allies.

The selection of pilot sites reflects some 'managerial' considerations. On the one hand, I observed that the DWP did not pay more attention to the fluidity of local labour markets whether the pilot was implemented by JCP or by another organisation. On the other hand, the consistently negative and significant effect of capacity suggests that the managerial logic does play a role in the selection of pilot sites. The idea that 'busier' JCPDs are less likely to be selected for the piloting of a new intervention than districts with more capacity not only makes sense from an organisational viewpoint, it also concurs with the interviews realised in the 'feasibility' stage of this project.

One of the most surprising findings is the geographical distribution of pilots across England. Even after controlling for population density, fluidity of the labour market and proportion of benefit claimants etc., we can observe that districts from the South of England and London are systematically under-represented in pilot programmes. The difference with the North of England and the Midlands is particularly large and statistically significant. It is unclear, at this stage of my research, how this choice – to the extent that it is one – can be justified from the viewpoint of the scientific logic. It looks as if London and Southern districts were used as quotas in pilot studies rather than for their true representativeness of the country.

This analysis also showed a positive, significant and consistent correlation between the proportion of ethnic minorities in the population and the odds of being selected as pilot site. Positive correlations were also found for the working age population and the population density.

One possible explanation, mentioned earlier, is that new interventions tend to be piloted in Labour constituencies outside London. Indeed, the map of pilots seems to match to a large extent the map of Labour votes. Given the symbolic property of pilot programmes (Weiss 1979, 1986; Rogers-Dillon 2004), one could argue that pilots are used to give a distributive advantage to some regions, through an early access to new programmes and budgets. To test that claim, I would need to compare the distribution of pilot sites under a Labour and a Conservative government, which goes beyond the scope of this thesis.

6.9. Conclusion

This chapter was designed to examine whether, in the context of employment and welfare pilots, high-performing Jobcentre Plus districts were more likely to be selected as pilot sites than low-performing ones. Such an association would indicate the presence of confirmation bias, as high-performing districts are more likely to generate results supporting the government's initial hypothesis.

Two findings are noteworthy. First, I found that, overall, pilot sites were not selected with the primary aim of warranting representativeness. Indeed, between 1997 and 2010, the 'busiest' Jobcentre Plus district was seven times more likely to be selected as pilot site than the 'idlest' district – a result that is unlikely to be attributable to chance. This finding might appear suspect, especially when one considers that these districts have been shaped, in part, to be comparable in terms of caseload and resource allocation. Some other significant and robust associations are more difficult to explain. Controlling for other variables including population density, labour market characteristics and the ethnic composition of the population, the London and Southern JCPDs were significantly more likely to be selected as pilot sites than the Northern and Midlands districts. Likewise, districts with a higher population, a higher population density and a higher proportion of ethnic minorities were significantly more likely to be sampled. More research is needed to explain why.

Second, the initial hypothesis that high-performing districts would be more likely to be selected as pilot sites than low-performing ones can clearly be rejected. Indeed, the small, insignificant effect of performance on the probability of selection is one of the most robust findings of the study. So even if pilot sites are not representative, they do not appear to be exemplar either, at least not based on the collected evidence. Unsurprisingly given this result, none of the interactions between performance and the political

salience of the intervention had a significant effect on the odds of a given site to be selected. The data shows – convincingly – that the selection of pilot sites was in fact significantly constrained by capacity issues with a negative and significant correlation between the number of pilots already run in a given JCPD and the probability of this district to be sampled for a new pilot.

The main limitation of this study concerns the operationalization of the 'performance' variable. Although local labour market conditions can certainly guide researchers eager to pilot a new policy intervention in 'favourable' conditions, they do not reflect the quality of the local management. In other words, a JCPD with a relatively fluid labour market can still perform poorly if it fails to meet its objectives. Collecting and using data on JCPDs' results against target as well as on JCPD managers' characteristics would be useful in this respect.

7. Effect of policy commitments on outcome reporting

7.1. Introduction

In its most extreme form, reporting bias refers to the non-publication of a study because of inconvenient results (study publication bias). It has received much attention in the field of medical research (Abramson, 2008; Angell, 2005; Avorn, 2005; Goldacre, 2012). Empirical research has consistently shown that published research is more likely to be positive or statistically significant than unpublished research (Easterbrook, Berlin, Gopalan, & Matthews, 1991; Eyding et al., 2010; Song, Parekh, Hooper, et al., 2010). However, this type of investigation is particularly difficult and requires collecting unpublished data from regulators, drug manufacturers and conference papers. It would be even more so in the area of policy research, where protocols and registration are not required. Thus, the following chapter had to pursue a different strategy.

Within-study outcome reporting bias (ORB or 'spin') relates to studies that have been published. It has been defined as a specific reporting strategy, emphasizing the beneficial effect of an experimental treatment (Boutron, Dutton, Ravaud & Altman, 2010) but is equally relevant for the piloting of social interventions. The use of spin in scientific writing can result from ignorance of the scientific issue, unconscious bias, or wilful intent to distract the reader from statistically non-significant results (Boutron et al., 2010; Fletcher & Black, 2007). Spin can take different forms, such as, for example, incomplete reporting, a particular focus on less informative results or an inadequate interpretation of non-statistically significant differences (Boutron et al., 2010). Spin can also occur at later stages, for example in the communication of results to stakeholders and the media (Yavchitz et al., 2012); however this is not addressed here.

The underlying assumption in medical meta-research is that these distortions are a manifestation of confirmation bias (or experimenter's bias), which is a tendency to favour information that confirms prior beliefs or hypotheses (Plous, 1993). The investments made for the development of new drugs is such that pharmaceutical companies can hardly afford reporting on ineffective drugs. In this chapter, I investigate whether a similar risk of bias exists in policy research. The amount of political capital invested in some reforms would justify a more 'hands on' approach to the evaluation. Two related questions will be addressed: What is the prevalence of spin in policy evaluation reports? Is spin more likely when the government expressed a commitment to the policy?

This question matters as the likely bias from spin is to overestimate the effect of the intervention, leading to moral hazard. Firstly, the beneficiaries of these policies will receive interventions which might have an insignificant effect or even cause harm. Secondly, voters using this type of information to appraise government performance will be misled. Thirdly, researchers and policy-makers using these results to inform subsequent policies will also be misguided (see Bailar, 2006; Fletcher & Black, 2007; Marco & Larkin, 2000 for a similar discussion in a medical context).

The following chapter is organised as follows. Section 7.2 reviews the literature on ORB. Section 7.3 presents the data and methods used in the empirical analysis. Section 7.4 introduces the hypotheses to be tested. Section 7.5 reviews two sets of technical specifications for policy evaluations. Section 7.6 presents the results of the content analysis; which are then discussed in section 7.7. Section 7.8 concludes.

7.2. Expected effect of policy commitments

In the UK, where most policy evaluation is carried out by contracted organisations on behalf of ministerial departments, the formal decision of what should be reported and how is shared by the evaluation team and the civil servants managing the project (Boa et al., 2010; The LSE GV314 Group, 2014; Walker, 2001). This situation creates an agency problem. On the one hand, evaluators are recruited based on their reputation for competence and expertise. On the other hand, they might want to reciprocate the favour of having been awarded a contract (Fehr & Gächter, 2000).

7.2.1. Principles of scientific reporting

There is no specific prescription for the reporting of outcomes of policy interventions. In the UK, the reference document for the management of research project, the Magenta Book (2011) is vague. Its recommendations on reporting fit on one page and stress that many policy makers are able to read and understand complicated analysis, but most do not have the time. Consequently, many will want to be given a flavour of the complexities of the analysis but without getting lost in details. Other policy makers may not have the technical background and will want a simpler presentation. So there is a delicate balance between keeping the respect and interest of the more technical while not losing the less technical. ‘Reporting tips’ are provided (based on Vaughan & Buss, 1998).

In contrast, reporting guidelines in the area of medical research are much more thorough. The Declaration of Helsinki states that “authors have the duty to make publicly available the results of their research on human

subjects and are accountable for the completeness and accuracy of their reports" (World Medical Association, 2013). To help enforce this principle, trial registration is required (American Economic Association's registry; Registry for International Development Impact Evaluations) and reporting guidelines are available[14]. Although these reporting requirements go way beyond the practice of policy research, they are still useful as a norm, an objective that researchers should strive to achieve. This chapter assumes that a researcher taking a scientific approach will report findings according to pre-specified research questions, theories and variables. Specifying the method from the outset of the research process means that outcomes cannot be manipulated (for example, in order to present flattering results). Therefore, provided they apply similar methods, different researchers are likely to report the same results, whether these results are positive, negative or nil.

7.2.2. Are evaluation reports spun?

Until recently, the idea that the results of policy evaluations might be spun to produce politically useful results was mainly a speculative one. Building on the research utilisation literature which flourished in the 1970s and 1980s (Barnsteiner & Prevost, 2002; Caplan, 1980; Knott & Wildavsky, 1980; Weiss, 1979), some argued that the strong agency relationship existing between policy researchers (civil servants or consultants) and policy-makers (elected or appointed) made the former vulnerable to the pressure exerted by the latter (see chapters 2 and 4). More precisely, it could shift the purpose of evaluation reports from pursuing 'speaking truth to power' to providing 'political ammunition' to policy-makers already committed to a specific course of action (Bovens et al., 2008; The LSE GV314 Group, 2014). The converse idea that the contractual relationship between the policy-making and research communities could, in the long run, mould 'docile researchers intuitively oriented to producing satisfied funders' has also been put forward (Allen, 2005; Metcalf, 2008). However, evidence of either theory has been equally scant and mostly anecdotal.

A recent survey of academics having completed commissioned research for government has strengthened the evidence base. Researchers found that more than half of respondents reported they were asked to make significant changes to their draft reports (i.e. affecting the interpretation of findings or the weight given to them). The most effective constraint appears to be found when government specifies the nature of the research to be done at the outset. No other form of constraint has as powerful an effect on the degree to which the overall conclusions the researchers reach support government policy (The LSE GV314 Group, 2014).

[14] These guidelines have been listed by the US National Library of Medicine: http://www.nlm.nih.gov/services/research_report_guide.html

Despite its significant contribution, the LSE GV314 study is not without its weaknesses. For example, the finding that "the academics' ability to resist pressure to steer the results appears to be substantial" is debatable given its design. As acknowledged by the authors of the study, "academics have an interest not only in resisting political pressure, but also in appearing to be able to resist it even if they cannot or do not". More objective data is needed.

7.2.3. Policy commitment and level of spin

The ideal design to assess the influence of policy commitments on the level of spin in an evaluation report would be to compare two series of evaluation reports, some conducted by a governmental body, the others by organisations with no vested interest in the success of these programmes. Any significant difference between the two sets of studies, controlling for other variables, would give strong evidence that politics can influence the reporting of policy outcomes. This type of design has already been used in medical research; unfortunately, it is more difficult in a policy context, given that few non-governmental organizations commission evaluations (see chapter 4 for a more detailed discussion). This study is based on a different design, which compares, in a qualitative way, the level of spin across different evaluation reports. These reports were selected in a way that maximizes the contrast between interventions to which the government was strongly committed (high-salience) and interventions to which the government was not or weakly committed (low salience).

The *a priori* relationship between policy commitments and level of spin is unclear. It could be argued that the outcomes of high-salience reforms are more accurately reported because they are more likely to be scrutinized by the media, the research community, watchdogs and interest groups. The opposite case makes just as much sense. The outcomes of high-salience reforms could be subject to more spin, given the high stakes and the blame game and political sanction that could follow the claim that a major reform is a 'failure'. Ministers are rarely neutral about their research. If they are testing a novel intervention, they usually suspect that it is effective otherwise they could not convince themselves, fellow cabinet members, members of parliaments and ultimately the public at large that it is worth evaluating. This lack of equipoise can affect the way they interpret negative results. Ministers having invested a large amount of political capital in developing the policy under evaluation might find it difficult to accept that it may be ineffective. In addition, democratic institutions create strong incentives to 'frame' research findings in a positive way, especially in countries where governments have the responsibility to ensure their citizens' welfare. In such a context, political failures tend to be remembered more than successes, and indeed ministers often turn out to get less credit from the voters for their successes than the blame they get for failures (Hood, 2011; Weaver, 1986).

Although it is thin, the evidence base leans towards the second hypothesis. Several interviewees of the LSE GV314 team indicated that policy-makers were more inclined to try and influence the reporting of outcomes when the reform was perceived as politically salient (The LSE GV314 Group, 2014). Likewise, Rhodes concluded from his observation of the British senior civil service that evidence was used to construct story lines rather than to inform policy decisions (Rhodes, 2013).

7.3. Data and methods

7.3.1. Approach

The design of this study has been shaped by the various constraints pertaining to the research question and the available data. To begin with, the definition of reporting bias presented in the introduction implies that studies not supporting the initial working hypothesis (*i.e.* the intervention has a high probability of having positive and significant effect on the population) are more likely to be spun than others (Hewitt, Mitchell, & Torgerson, 2008). Thus, the studies reviewed in this chapter all reported a primary outcome that was either not statistically significant at the conventional level ($P \geq 0.05$) or in the direction opposed to the initial hypothesis (*i.e.* the intervention has a negative effect).

The availability of data created a number of additional constraints. Firstly, the fact that policy evaluations are overwhelmingly commissioned by the governments which designed and implemented those policies means that it was not possible to compare the amount of reporting bias in studies taking place within the political sphere and outside of it. Such design would have provided a useful counterfactual. Although it has been used in medical research to assess the effect of industry sponsorship on reporting (Bourgeois et al., 2010), it remains difficult to replicate in policy research. Instead, I had to contrast studies with a high level of policy commitment with studies with a lower level of policy commitment, as explained in chapter 2.

Secondly, the absence of formal research protocols for the evaluation of public policy means that it was not possible to estimate the amount of reporting bias through systematic comparisons between the content of published reports and those protocols (Bourgeois et al., 2010) or other documents issued in the planning phase of research such as research proposals (Rising, Bacchetti, & Bero, 2008). In other words, there is no clear baseline against which published results can be benchmarked. Instead, I looked for evidence of research decisions that have previously associated with an intention to spin (Boutron et al., 2010). Those include incomplete statistical outputs (Chan & Altman, 2005; Chan et al., 2004), spurious analyses (Ioannidis & Karassa, 2010; KL Lee, McNeer, Starmer, Harris, & Rosati, 1980; Rothwell, 2005) and biased interpretations of results (Alasbali

et al., 2009; Boutron et al., 2010; Malenka, Baron, Johansen, Wahrenberger, & Ross, 1993). Those variables will be presented in greater details below.

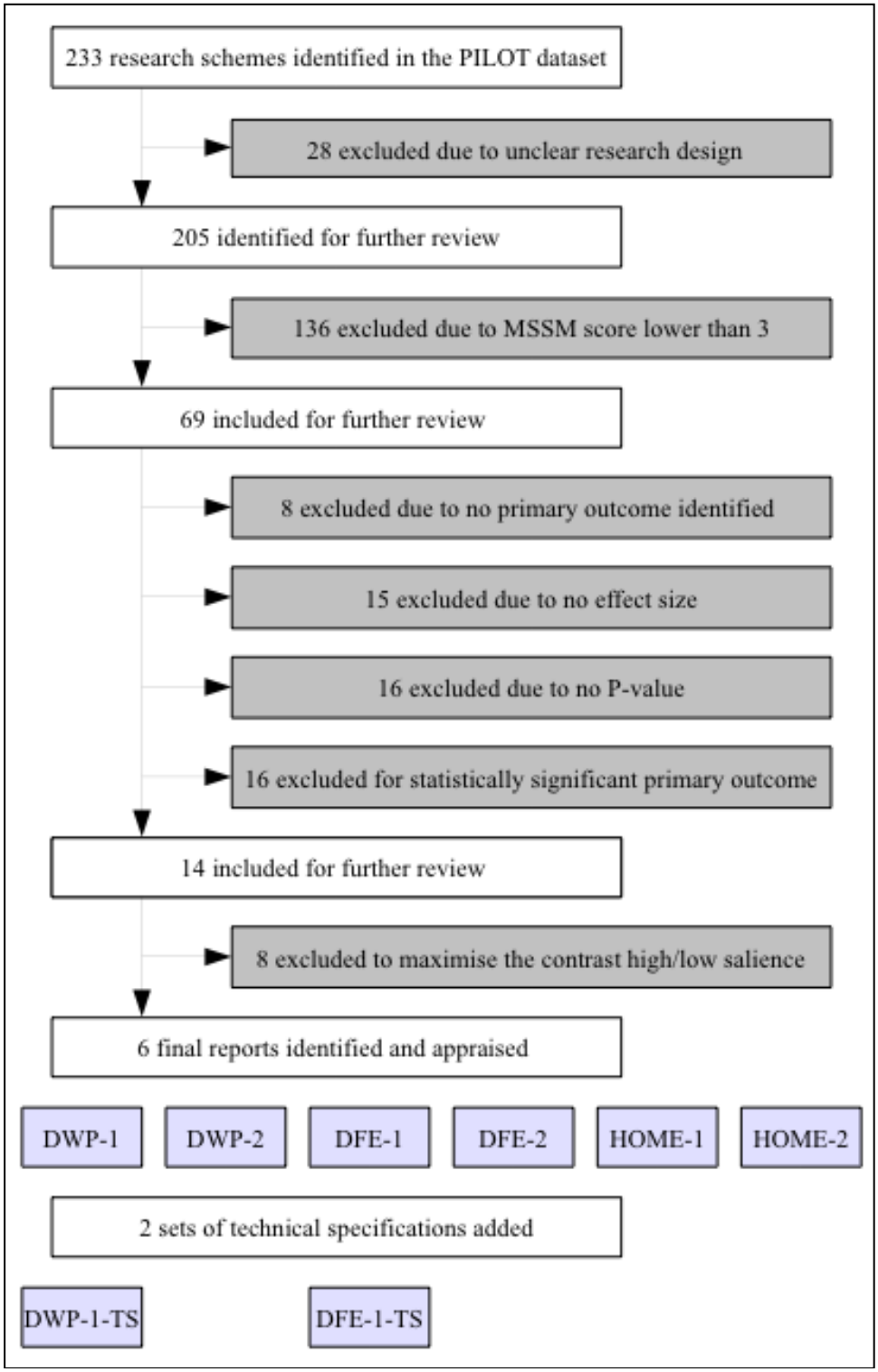
Thirdly, the number of evaluation reports amenable to this kind of research was too limited to allow a quantitative analysis. Instead, a qualitative approach was adopted, focusing on the content of these reports, their claims and the language adopted by evaluators. Two sections were analysed: the 'results' sections and the 'executive summaries' (or, when missing, the conclusion or the 'policy brief' which sometimes accompany the main study). The main implication for this study is that my observations are limited to the chosen sample.

Against this background, this chapter offers a qualitative analysis of the content of six evaluation reports with different levels of political salience. Its objective is to find out, in the context of studies with non-significant primary outcome, whether high-salience interventions are subject to more spin than low-salience interventions.

7.3.2. Selection of studies

The selection process is shown in Exhibit 26. Studies were screened from the PILOT dataset presented in Annex I. The selection process followed a number of steps. First, studies with a score of 3 and above on the Maryland Scale of Scientific Method were included and studies with a 'weaker' design were excluded. Level 3 corresponds to "comparisons between two or more comparable units of analysis, one with and one without the programme" (Sherman et al., 1998). When several studies were available for the same pilot, I selected the one, which seemed to offer the most definitive conclusions regarding the effect of the intervention (*e.g.* final report as opposed to interim report). Cost-effectiveness and cost-benefit outcomes were not considered. From this sample, I then screened the full-text studies and looked for primary outcomes. Only studies showing that the intervention had a non-significant effect were selected ($P \geq 0.05$). The decision to use a P-value of 0.05 or a 95% confidence interval to determine statistical significance is arbitrary but widely accepted. Conversely, I excluded studies for which the primary outcome could not be identified with confidence and studies showing a positive and significant effect of the intervention. In one study, the primary outcome was not identified from the evaluation report itself but from the technical specifications issued by the sponsoring department for the evaluation of the intervention.

Exhibit 26 – Case selection



From this sample, I attempted to select two studies per government department with relatively high/low levels of salience. In this chapter, I used a surrogate measure, which the level of seniority of the 'champion' or 'sponsor' of the reform, based on who made the first announcement. The announcement of a pilot can be seen as a delegation issue, whereby each principal, from the Prime minister to the mid-level bureaucrat can decide whether to be the 'manager', taking direct responsibility for the outcome, or the 'chair of the board' overlooking operations (Hood, 2011). Given politicians' propensity to avoid blame even when that implies not getting credit (Weaver, 1986), I consider a pilot announced by the Prime minister as being more politically salient than pilots announced by any other policymaker (Chancellor, Secretary of State, junior minister, civil servant) or a pilot not announced at all. An ordinal variable reflecting these categories (in this order) was created for my analysis. The announcer is also convenient in that it captures many of the dimensions of political salience including the 'size' of the programme, its visibility, etc. Here, the objective was to maximise the contrast between high-salience interventions and low-salience interventions. When two studies or more at the same level of salience, the final selection decision was made at random. That was possible for the Department for Work and Pensions (DWP) and the Department for Education (DFE) but not for the Ministry of Justice (MoJ) and the Home Office (HO), for which it was not possible to find the desired pairs. In the end, I selected a high-salience intervention evaluated by the HO and a low-salience intervention evaluated by the MoJ. I believe that this decision does not compromise significantly the design of this study for two reasons. Firstly, the two organisations have very similar cultures. The MoJ was formed in 2007 when some functions of the Home Office were combined with the Department for Constitutional Affairs. As a result of this re-organisation, staffs were moved from the HO to the newly created MoJ. Secondly, the two selected interventions were evaluated against the same primary outcome, namely the rate of reconviction. The corpus of this analysis includes six studies, which are presented in Exhibit 27.

In addition, Freedom of Information requests were sent to the relevant government departments to get hold of the technical specifications issued for these evaluations, as well as any interim report not published on their respective websites. Technical specifications were obtained for two studies out of six (DWP-1; DFE-1) and one interim report for one study only (three studies had no interim report). This interim report was screened but no impact analysis was found and so it was decided not to include it in the study corpus. The list of documents that were reviewed can be found in the Appendix at the end of this chapter. The interventions are presented in Exhibit 27.

Exhibit 27 – Study corpus

Pathways to Work (DWP-1)

- Sponsor: Department for Work and Pensions;
- Strong commitment: First announced by the Chancellor.
- Objective: to encourage employment among people claiming incapacity benefits.
- Interventions: The pilot consisted of a series of interventions including mandatory interviews with a personal adviser for new benefit claimants; support to claimant in the management of their health condition; as well as an extra £40 per week credit for the first 12 months of employment.
- Dates: The programme was introduced on a pilot basis in seven Jobcentre Plus districts between October 2003 and April 2004.
- Policy decision: Since then, Pathways gradually expanded to cover more districts, so that by April 2008, all new incapacity benefits claimants in Britain were eligible for Pathways.

Job Retention and Rehabilitation Pilot (DWP-2)

- Sponsor: Department for Work and Pensions and Department of Health
- Weak commitment: Never formally announced
- Objective: To increase the return-to-work rate of those off-work sick for six weeks or more.
- Interventions: The pilot tested three new interventions: a workplace intervention (e.g. ergonomic assessment), a health intervention (e.g. physiotherapy) and a combined intervention, which effect were compared to existing provisions. The JRRP involved over 2,800 voluntary participants who were randomly assigned to one of the four conditions.
- Dates: It was tested in six areas across the UK. The pilot ran from April 2003 to March 2005.
- Policy decision: JRRP was not rolled out.

The Two Year Old Education Pilot (DFE-1)

- Sponsor: Department for Children, Schools and Families (which later became the Department for Education),
- Strong commitment: First announced by the Chancellor in his 2004 Pre-Budget Report.
- Objective: The Two Year Old Education Pilot intended primarily to improve children's social and cognitive development.
- Intervention: The pilot provided free early years education to over 13,500 disadvantaged two year olds.
- Dates: From April 2006 to April 2008.
- Policy decision: The pilot was not rolled out.

The Multidimensional Treatment Foster Care (DFE-2)

- Sponsor: Department for Education
- Weak commitment: Never formally announced
- Objective: To improve the social and cognitive development of children aged 10-16 who were placed in foster care.
- Intervention: MTFC employs multiple methods, including individual and family therapy, social skills training and support with education. MTFC provides young people with a short-term foster placement, usually intended to last around nine months, followed by a short period of aftercare. The study was designed as a randomised controlled trial (RCT), but in anticipation of the potential difficulties, was embedded within an observational study to ensure a sufficient sample.
- Dates: The pilot ran from 2002 to 2006.
- Policy decision: Unknown.

The Alcohol Arrest Referral Pilot (HOM-1)

- Sponsor: Home Office
- Strong commitment: First announced by the Home Secretary in February 2008
- Objective: Its aim was to reduce the number of offences related to alcohol consumption.
- Intervention: AAR involved offering a brief intervention to individuals arrested and deemed by a police officer to be under the influence of alcohol. This intervention consisted of (1) an assessment of the clients' drinking patterns; (2) the provision of information on the risks of excessive alcohol consumption; (3) practical advice for managing the risk of drinking; and (4) a follow-up session.
- Dates: The pilots were located in eight police force areas in England and were funded between November 2008 and September 2010.
- Policy decision: Unknown.

The Restorative Justice Pilots (HOM-2)

- Sponsor: Ministry of Justice (but funded by the Home Office)
- Weak commitment: Never formally announced
- Objective: to reduce re-offending whilst retaining "significant focus on the needs and rights of victims".
- Interventions: Three different types of judicial mediation were tested (direct, indirect and conferencing) in different settings.
- Dates: The three pilots ran between mid-2001 and early 2004.
- Policy decision: Unknown.

7.4. Hypotheses

7.4.1. Missing outcome indicators

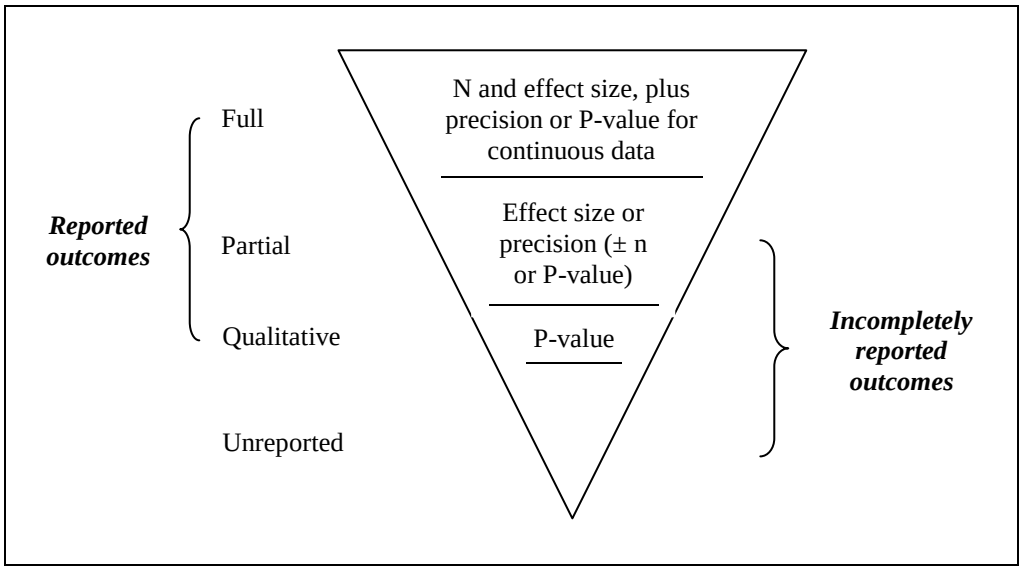
The most obvious form of spin is to 'filter out' the least convenient results. The medical research literature has often commented on the existence of unreported study outcomes, using different strategies to estimate the gap between published studies and what was thought to best represent the original intentions of the researchers. In an early study, the investigator compared the number of statistical analyses reported and the number that were not reported but were very likely to have been undertaken, given the data and variables presented in the study (Tannock, 1996). More recent studies used the research protocols submitted to ethics committees as baseline for their work (Chan et al., 2004; Hahn, Williamson, & Hutton, 2002). However, those are difficult to obtain, even in medical research. One study used primary publications (*i.e.* the first report of final trial results) as a proxy for research protocols and compared the content of subsequent publications to estimate the number and type of unreported outcomes (Chan & Altman, 2005). Unfortunately, for the reasons presented in section 7.3.1, it was not possible to carry out a similar analysis.

7.4.2. Incomplete reporting

Even when outcomes are presented in publications, they may be reported superficially (see Dwan et al., 2008 for a review). Direct evidence of such bias has recently been shown in two cohort studies that compared trial publications with the original protocols (Chan et al., 2004). For each identified outcome, the level of reporting can be recorded as one of three levels based on the amount of data presented in the publication. If sufficient data is provided for inclusion in a meta-analysis, the outcome can be recorded as fully reported. This data includes (a) group numbers; (b) size of intervention effect and (c) a measure of precision/variability (P-value and/or confidence interval). An outcome is considered partially recorded if the publication provides only some of the data necessary for meta-analysis and qualitatively reported if the publication presents only a measure of statistical significance (Chan & Altman, 2005) (see Exhibit 28). Against this background, I formulate the following hypothesis:

H1: In the context of studies with non-significant primary outcome, the risk of incomplete reporting is positively associated with the strength of the government's commitment to the reform.

Exhibit 28 – Hierarchy of levels of outcome reporting



Source: Chan and Altman (2005).

7.4.3. Spurious subgroup analyses

The effects of an intervention on the entire study population are of primary interest in a study. It could be appealing, however, for investigators and research commissioners to identify differential effects in subgroups based on characteristics of trial participants or interventions. This analytic approach, termed 'subgroup analysis', can sometimes be informative – but it is often misleading (Fletcher, 2007; Oxman & Guyatt, 1992; Schulz & Grimes, 2005; Yusuf, Wittes, Probstfield, & et al., 1991). Some have compared them as data-driven 'fishing expeditions', in which investigators perform numerous post-hoc subgroup analyses, seeking statistical significance (Rothwell, 2005; S. Wang, Ou, Cheng, & Dahm, 2010). Clinical research has shown that conducting multiple tests was associated with the risk of false-positive results due to chance. Even when investigators specify a limited number of subgroup analyses *ex ante*, chance can result in the identification of spurious subgroup effects (Rothwell, 2005; S. Wang et al., 2010).

The clinical literature offers criteria that aid differentiation between spurious and real subgroup effects (Guyatt, Wyer, & Ioannidis, 2008; Sun, Briel, Walter, & Guyatt, 2010). The criteria used in this chapter are based on these guidelines (see Exhibit 29). These include whether the hypothesis of a subgroup effect preceded the analysis, and was one of a few subgroup hypotheses that were explored. It is also important that the appropriate statistical test for investigating a subgroup effect is not whether a statistically significant effect is seen in one subgroup and not in another (Wang, Lagakos, Ware, Hunter, & Drazen, 2007). Instead investigators

should use a statistical test of interaction that, assuming that no subgroup effect exists, test the hypothesis of how often one would observe differences in apparent effects as large as or larger than those observed in the study. Against this background, I test the following hypothesis:

H2: In the context of studies with non-significant primary outcome, the risk of spurious sub-group analyses is positively associated with the strength of the government's commitment to the reform.

Exhibit 29 – Guidelines for determining whether differences in subgroup responses are based on real criteria

Design
1. Was the hypothesis specified a priori?
2. Is the subgroup variable a characteristic measured at baseline or after assignment?
3. Is the subgroup difference suggested by comparisons within rather than between studies?
4. Was the direction of the subgroup analysis specific a priori?
5. Was the subgroup difference one of a few hypothesised effects tested?

Analysis
6. Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?

Context
7. Does external evidence support a hypothesised subgroup effect?

Source: Sun et al., (2010)

7.4.4. Spurious within-group comparisons

The essence of a clinical trial or policy pilot is to compare the outcomes of groups of individuals going through different interventions. We expect studies to give us an estimate of the difference (the 'intervention effect') with a confidence interval and a P-value. However, rather than comparing the groups directly, researchers sometimes look *within* groups at the change between the outcome measure from pre-intervention baseline to the final measurement at the end of the trial. They then perform a test of the null hypothesis that the mean difference is zero, separately in each group. They may then report that in one group this difference is significant but not in the other and conclude that this is evidence that the groups, and hence the treatments, are different (Bland & Altman, 2011). To test this idea, I formulate the following hypothesis:

H3: In the context of studies with non-significant primary outcome, the risk of spurious within-group comparisons is positively associated with the strength of the government's commitment to the reform.

7.4.5. Misleading inferences

When a study shows a difference that is not statistically significant there is a risk of interpretive bias (Hewitt et al., 2008; Kaptchuk, 2003). Interpretive bias occurs when authors and readers overemphasise or underemphasise results (Hewitt et al., 2008). For example, authors may claim that the non-significant result is due to lack of power rather than lack of effect, using terms such as 'borderline significance' or stating that no firm conclusions can be drawn because of the modest sample size. In contrast, if the study shows a non-significant effect that opposes the study hypothesis, it may be downplayed by emphasising the results are not statistically significant. For the purpose of this analysis, I define a non-significant result as a regression coefficient with P-value larger than the conventional 5% level. I will test the following hypothesis:

H4: In the context of studies with non-significant primary outcome, the risk of interpretative bias is positively associated with the strength of the government's commitment to the reform.

7.4.6. Upgrading or downgrading outcomes

In a given study, a primary outcome is the outcome of greatest importance. Data on secondary outcomes are used to evaluate additional effects of the intervention (CONSORT Statement 2010). According to Dwan, Kirkham, Williamson and Gamble (2013), 'selective reporting' occurs when, in a given study (1) a primary outcome is downgraded to secondary (downgrade); (2) a secondary outcome is upgraded to primary (upgrade); (3) a new outcome not stated in the protocol is added to the full review (addition); or (4) an outcome stated in the protocol was omitted from the full review (omission). When a change in outcomes occurs, it must be said and justified (Dwan et al., 2013). In this chapter, upgrades/downgrades were identified in two ways: (a) through comparisons between technical specifications or interim reports on the one hand and final reports on the other hands; (b) within studies, by comparing the order of results in the executive summary and the 'results' section. Thus, an outcome coming first in the results section and second in the executive summary will be considered 'downgraded'. Against this background, I formulate the following hypothesis:

H5: In the context of studies with non-significant primary outcome, the risk of upgraded/downgraded outcomes is positively associated with the strength of the government's commitment to the reform.

7.4.7. Conclusion bias

Finally, one can look at the evaluator's final judgement of the merit of the intervention in the conclusion of the report or its executive summary. An overemphasis on positive results will be taken as an indication of interpretive bias. For that purpose, I assessed the level of spin in the executive summaries, updating the classification developed by Boutron, Dutton, Ravaud and Altman (2010). *High spin* was defined as the suggestion that, overall, the intervention was a success despite a non-significant primary outcome. *Moderate spin* was defined as the acknowledgement of the non-significant effect of the intervention, but with an immediate emphasis on spurious analyses meant to distract the reader from the main study outcome. *Low spin* was defined as the acknowledgement of the statistically non-significant result for the primary outcomes and uncertainty in the framing of the study.

H6: In terms of overall conclusion, the risk of conclusion bias is positively associated with the strength of the government's commitment to the reform.

7.5. Analysis of technical specifications

The following section is based on the two sets of research specifications obtained from government (studies DWP-1 and DFE-1). The analysis of the content of technical specifications leads me to make four remarks.

Firstly, the technical specifications issued by commissioning departments provide a clear illustration of the agency problem. On the one hand, tendering evaluators are required to provide evidence of their qualifications for the job. The following excerpt suggests that the most competent candidate will be retained:

> "Tenderers' suggestions for evaluating net impact needs to be of the highest quality, and this will be looked at specifically in addition to a more broad requirement of methodological expertise" (DWP-1 TS, p.27).

On the other hand, the document reminds the candidates that the policy and analysis teams within the commissioning departments will remain the ultimate decision-makers on key research decisions, including reporting:

> "The contractor will be expected to work closely with officials of the Department throughout the research, keeping them informed of progress and involving them in key decisions. Officials in policy and analytical branches in DWP and DH must have the opportunity to

> comment on and approve topic guides and questionnaires, formats for analysis and draft reports" (DWP-1-TS: 22-23).

Secondly, technical specifications suggest that the salience of the reform has an effect on how evaluation outcomes will be reported:

> "This will be a high-profile evaluation and to get full value from it, timely and high quality reporting is essential. To ensure full value of the evaluation tenderers should consider ways in which emerging findings from studies can most appropriately be fed back to policy officials in order to inform further policy development. For example in advance of the production of draft reports, contractors are likely to be asked to present headline findings to core policy officials and analysts" (DWP-1-TS, p.24).

However, it is unclear from the above whether the association is positive (higher salience reports are more spun) or negative (higher salience are less spun). The notion of "high-quality reporting" as that of "policy relevant" is subjective (The LSE GV314 Group, 2014). The following excerpt, also from the technical specifications for the evaluation of a higher salience intervention suggests that the level of spin is limited:

> "It is the expectation that the key outputs from the study will be in the public domain. The Department will aim to publish key outputs within a reasonable period of time following receipt of an agreed final report. The publication of any research articles or other publications based on information collected for this study will be subject to approval from the DfES. However, this will not be unreasonably withheld" (DFE-1-TS, p.4).

Fourthly, the content of technical specifications shows that, despite the fact that they are the closest document to a research protocol one can get, their use remains problematic. Indeed, tendering evaluators are expected to contribute to the design of the study:

> "Tenderers are invited to suggest what further surveys of clients in pilot and other areas would be useful in arriving at an impact assessment" (DWP-1-TS, p.15-16).

Additionally, amendments to the original intervention or to the original design of the evaluation cannot be ruled out:

> "[Tenderers] must also demonstrate a commitment to meet deadlines and yet be sufficiently flexible, should the programme of work require amending" (DWP-1-TS, p.26).

7.6. Analysis of final reports

7.6.1. Incomplete reporting

Hypothesis 1 states that, in the context of studies with non-significant primary outcome, the risk of incomplete reporting is positively associated with the strength of the government's commitment to the reform. All six reports reviewed in this study reported intervention outcomes in a complete way, *i.e.* including group size, effect size and a P-value. A minor presentational flaw was found in the DFE-2 evaluation, which reported P-values in the text and not in the output as customary. However, given these results, no association between completeness of reporting and commitment can be established. On the basis of the limited evidence, Hypothesis 1 seems unlikely.

7.6.2. Within-group analyses

Hypothesis 2 states that, in the context of studies with non-significant primary outcome, the risk of spurious within-group comparisons is positively associated with the strength of the government's commitment to the reform. None of the studies in this review conducted within-group analyses, so no evidence of an association between this type of spin and political salience can be established. Hypothesis 2 seems unlikely.

7.6.3. Sub-group analyses

Hypothesis 3 states that, in the context of studies with non-significant primary outcome, the risk of spurious sub-group analyses is positively associated with the strength of the government's commitment to the reform. To test this hypothesis, I assessed the credibility of the subgroup analyses of each study using the criteria mentioned earlier, with the notable difference that in my study, all subgroup variables are assessed together and not individually for simplification. Exhibit 30 shows the results.

The assessment of the credibility of these sub-group analyses shows a pattern. Firstly, all studies used a small number of sub-group variables, usually between four and six, with one exception (DWP-2: 10 subgroups). Qualitatively, two of these sub-groups seem to be consistently tested: sex and age. Other variables are subject-specific but fairly consistent within a given policy area (socio-economic status, health situation, family situation). Secondly, sub-group analyses undertaken in policy research seem to differ systematically from those conducted in medical research on at least three indicators. Indeed, I found out that all the analyses carried out in these six studies (a) were based on characteristics measured at baseline, (b) were suggested by comparisons of within-studies and (c) were based on tests of

interaction. This suggests that the sub-group analyses conducted in policy research have a high level of credibility.

Given the design of this study and the data available, it is more difficult to assess whether these analyses were underpinned by a theory and whether the direction of the effect was specified from the outset. *Prima facie*, it seems that in most cases, sub-group analysis was exploratory rather than confirmatory. In only two studies, both low-salience (DFE-2 and HOM-2), researchers clearly reported why they were conducting these analyses and what they were expecting. However, the evidence is not strong enough to conclude that, in other instances, sub-group analyses were carried out with the aim to mislead the reader. Instead, experience and an incremental approach to policy development seem to have guided the researcher, as explained in the DWP-2 study:

> "The choice of variables from which to create sub-groups is somewhat arbitrary. The final list is based on a selection of possible variables for which: (a) the sub-groups have large enough sample sizes for at least moderately large impacts to be detected; (b) there is some expectation that impacts may have been different in at least some of the sub-groups" (DWP-2: 49).

Exhibit 30 – Credibility of sub-group analyses

	DWP-1	DWP-2	DFE-1	DFE-2	HOM-1	HOM-2
Number of sub-groups	4	10	4	5	4	6
Number of subgroup variables measured at baseline	4/4	10/10	4/4	unclear	4/4	/6
Number of analyses suggested by comparisons of within *vs.* between studies	4/4	10/10	4/4	5/5	4/4	6/6
Number of sub-group analyses based on interaction	4/4	10/10	4/4	5/5	unclear	unclear
Theoretical justification mentioned	0/4	0/10	0/4	3/5	0/4	2/6
Number of analyses for which the direction of the SG effect was specified *a priori*	0/4	0/10	0/4	2/5	0/4	2/6

Source: Sun, Briel, Walter and Guyatt 2010

Another explanation is that, consciously or unconsciously, evaluators did not report their original intentions with the level of accuracy that would be expected in a peer-reviewed scientific publication. The technical specifications of the DWP-1 study highlight that:

> "A key requirement underpinning sampling is the need to include a discussion on the capability of analysing sub-groups, and any implications for overall samples of the need to estimate impacts of separate components. We would welcome suggestions on types of sub-group analyses" (DWP-1-TS: 17).

The above shows that there is no evidence that the sub-group analyses carried out in these six studies were spurious. Hypothesis 3 seems thus unlikely.

7.6.4. Interpretation of results

Hypothesis 4 states that, in the context of studies with non-significant primary outcome, the risk of interpretative bias is positively associated with the strength of the government's commitment to the reform. The six studies reviewed regarded P-values of 0.05 or less as indicating statistical significance, even though this point was not always explicitly made. For example, the HOM-1 and HOM-2 evaluations made no reference to a cut-off point, however only P-values smaller than 0.05 led to a formal rejection of the null hypothesis. The one exception to this pattern concerns the significance test of the primary outcome of DWP-1. The study reads as follows:

> "The P-value suggests that the impact is statistically significant since there is only a nine per cent probability of finding an effect of this size by chance" (DWP-1, p.48).

This comment is accompanied by the following footnote:

> "By convention, P-values of five per cent or less are regarded as indicating statistical significance. However, this is essentially arbitrary and ignores the continuous nature of P-values. The approach taken in this report is to use the conventional five per cent P-values for the results based on the administrative data but to use ten per cent P-values for the results based on the survey data in view of the smaller sample size available for these estimates" (DWP-1, p.48).

It is useful to mention here that the "smaller sample size" the evaluators refer to is 3,237 – which many will regard as sufficient to yield credible results. The fact that such a bias concerns a pilot with higher political salience suggests that an association between policy commitment and interpretative bias cannot be excluded when considering the 'big picture'.

The other aspect of interpretative bias relates to the attribution of non-significant effects (intervention *vs.* methods). Three types of languages were used across the six studies. In three reports (DFE-1, HOM-1, DWP-2), non-significant effects were unambiguously attributed to the intervention, as shown by the following excerpts:

> "The finding is clear-cut: there is no evidence that, on average, the pilot improved the non-verbal reasoning of children overall" (DFE-1, p.99).

> "The key finding was that overall [the intervention] appeared to be ineffective for the client group in reducing re-offending. There was a higher rate of re-arrest amongst the intervention group, compared with the comparison group" (HOM-1, p.25).

In one study (DFE-2), the non-significant result was attributed to the intervention, however the claim was followed by a caution note on the methodology used in the investigation:

> "Taking the sample as a whole, across both the randomised trial and the observational comparison, there was no evidence that the [intervention] resulted in significantly better functional outcomes than treatment as usual as measured on our primary outcomes. Despite the strengths of the study methods, this conclusion needs to be set against different kinds of limitations for each of the analyses. In the randomised study the sample size was underpowered to detect a plausible effect size. There was also a high proportion of 'crossover' cases". (DFE-2, p.153-154).

In two studies (DWP-1, HOM-2), evaluators strongly suggested that the insignificant effect was due to a lack of statistical power and that the effect would have been significant, had the sample been larger:

> "The small sample size of those in work and with earnings information at the time of the outcome interview reduced the likelihood of detecting an impact on earnings. No statistically significant impact of Pathways on monthly net earnings about a year and a half after the initial incapacity benefits enquiry was found (Table 5.2). It is not possible with the survey data to observe earnings between the time of the initial enquiry and the outcome interview; it is possible that there may have been an earnings effect during this period. In view of the employment effect of Pathways, one would expect a positive impact on earnings" (DWP-1, p.2).

> "The individual restorative justice trials and groups in this study each had relatively small sample sizes and therefore would not, on their own, be expected to have a large enough impact on re-offending to be statistically significant (*i.e.* so that we would know that they were

> unlikely to have been caused by chance). The exception was the Northumbria JRC court property trial which showed such a large impact on the reduced likelihood and severity of re-offending (against a control group) that these results were statistically significant" (HOM-2, p.33).

The above shows that spin might occasionally occur in the interpretation of findings. However, there is no strong evidence that such form of spin be positively associated with the strength of the government's commitment to the reform. Therefore hypothesis 4 seems unlikely.

7.6.5. Upgraded/downgraded outcomes

Hypothesis 5 states that, in the context of studies with non-significant primary outcome, the risk of upgraded/downgraded outcomes is positively associated with the strength of the government's commitment to the reform. I identified three categories of studies. The first category is by far the most representative. It brings together studies with clearly identified primary and secondary outcomes (DWP-2, DFE-1, DFE-2, HOM-2). I found that, for this cluster of studies, outcomes were reported in the executive summaries in the exact same way as in the Results section, *i.e.* according to their relative importance for policy-makers. There is one exception though: one of the secondary outcomes in the HOM-2 study was not reported in the executive summary of the study.

In another study (DWP-1), primary and secondary outcomes were not clearly signposted; however I found that the order in which they had been reported in the Results section was consistent with the objectives of the intervention as presented in the report. Furthermore, the order in which outcomes were presented in the executive summary is the same as in the Results section.

Finally, one study (HOM-1) evaluated the effect of the intervention on just one outcome, making the question of upgrading/downgrading outcomes irrelevant.

In light of the above, there seems to be no association between policy commitment and the risk of upgraded/downgraded outcomes. Hypothesis 5 seems unlikely.

7.6.6. Conclusion bias

Hypothesis 6 states that, in the context of studies with non-significant primary outcome, the risk of conclusion bias is positively associated with the strength of the government's commitment to the reform. In two studies, it was found that the level of spin was high (DWP-1, HOM-2). Indeed, the

executive summary of the DWP-1 evaluation states, despite a primary outcome borderline non-significant at the 10% level, that:

> "Overall, the results are encouraging in that they suggest Pathways continues to have a positive impact on employment and, furthermore, that this impact may be sustained" (DWP-1, p.4).

And although the evidence suggests the opposite, the HOM-2 concludes that:

> "Summed over all three restorative justice schemes, those offenders who participated in restorative justice committed statistically significantly fewer offences (in terms of reconvictions) in the subsequent two years than offenders in the control group" (HOM-2, p.iii).

Two studies were found to be subject to moderate spin (DFE-1; DFE-2). For example, the DFE-1 executive summary does acknowledge the non-significant result for the primary outcome of the study:

> "Taking all those children entering pilot places in aggregate, on average the pilot did not significantly improve the cognitive and social development of the children receiving the free childcare relative to a matched comparison group. The pilot children developed only very slightly further than their matched comparison group over the same period" (DFE-1, p.4).

But this statement is immediately followed by another on the effect of the intervention on one specific subgroup, which I have showed to be an example of spurious analysis:

> "However, this overall lack of a significant impact disguises the fact that for those children who were found places in relatively high quality settings (…) there was an impact on children, at least in terms of child vocabulary. For these children (who between them represent around two-thirds of all pilot children) the effect of the pilots was to significantly improve their language ability scores (from 45.8 to 49.4 on average). This is equivalent to moving a child from the 34th percentile for language development to the 46th percentile. What this suggests is that, had the pilot local authorities been able to secure more places in relatively high quality settings, then the pilot would have had a considerably larger impact overall" (DFE-1, p.4).

The last sentence is particularly interesting, as it suggests that the children centres sampled for this study were excessively representative of the population and that the pilot would have been more effective with higher-quality children centres. A similar pattern is observed for the DFE-2 evaluation.

Finally, I found evidence of low spin in two studies (DWP-2, HOM-1). In one of them (HOM-1), it seems that the results were so consistently negative that the study was 'beyond spinning'. However, spurious subgroup analyses were certainly tried:

> "The regression analyses confirmed that those receiving the intervention were significantly more likely to be re-arrested in the six months post-intervention than those in the comparison group. However, there were no clear subgroups for whom the scheme appeared to be more effective" (HOM-1, p. iii).

The other study (DWP-2) is much more 'sober' in terms of interpretation. No evidence of spin was found, as shown by these two excerpts:

> "It is not entirely clear why the interventions did not impact on employment" (DWP-2, p.7).

> "This report has shown no evidence that offering Job Retention and Rehabilitation Pilot interventions to those off work sick improved their chances of returning to work (DWP-2, p.129).

The fact that I found both high-salience and low-salience pilots in each of these three categories shows that there is apparently no association between salience and level of spin. Hypothesis 6 seems unlikely.

7.7. Discussion

None of the hypotheses predicting outcome reporting bias in connection with policy commitment has been confirmed. Given the theoretical arguments for why we might expect such an association, this section presents a discussion about the circumstances under which the reporting of outcomes from politically salient pilots is not biased. I discuss (1) the type of evaluations that are less spun; (2) the notion of salience; (3) the issue of blame shifting; (4) the stage of the policy process where spin is likely to occur; and (5) the organisational and political context in which reporting is more comprehensive;

Firstly, one could argue that some reports are not 'fit for spinning'. On the one hand, studies reporting overwhelmingly positive results do not need to be spun, as they will offer plenty of good news for their sponsors. On the other hands, studies not reporting a single significant outcome might just be 'beyond spinning'. The political cost of defending a reform showing very meagre results might be greater than that of chucking it altogether since in the latter case, ministers and policy-makers can more easily play the 'Chairman' card (Hood, 2011). Between these two extremes, 'murky' studies mixing a few good findings in a sea of insignificant results could be

better for fishing expeditions. There is some evidence in the studies I reviewed that this could be the case. For example, the authors of the DWP-2 study – which showed very few significant results – warned about the risks of over-interpretation:

> "There is a danger, of course, that in a trial that demonstrates little or no overall impact, too much emphasis may be placed on isolated findings. So, although we believe the findings for those self-assessing they can return to the same job (…) are probably genuine, we should stress that, it may just be statistical 'noise'" (DWP-2, p.50).

To the extent that this true, this means that the political salience of a pilot could have some direct effect on the probability of spinning but that this effect would be stronger in studies showing a mix of significant and non-significant outcomes.

Secondly, one could argue that none of the chosen policy areas was politically salient at the time when these pilots were conducted. Although some of the pilots were more politically salient than others, it could be that, in the broader policy spectrum, these policies were fairly consensual, at least during the New Labour government. This hypothesis is congruent with the idea of a progressive ideological convergence between the Labour and Conservative parties (L. Epstein, 1980; Rae & Gil, 2010). Also, it needs to borne in mind that only six studies were analysed. It cannot be excluded that, had another set of studies been selected, evidence of an association between political salience and spin would have been found.

Thirdly, it could be that the findings of impact studies carry no particular political risk. Indeed, these studies are virtually always accompanied by implementation/process evaluations which allow the government to shift the blame of failure to implementing bodies and frontline workers.

Fourthly, it could be that spin happens at a later stage of the policy cycle, for example in the phrasing of the press release announcing the publication of a given study (Yavchitz et al., 2012) or in subsequent policy documents and communications (Henig, 2008).

Fifthly, it could be that the effect of political salience on the level of spin depends on another variable, such as the minister or the department. That could be explained by the culture or reputation of the department (Carpenter, 2001). Civil servants also have a vested interest in the success in the policy. For example, analysts might want to demonstrate that their predictions regarding the expected effect of the programme were true:

> "It is hoped that the pilot provision will reduce that by approximately 4 percentage points" (DWP-1-TS: 7).

It should be borne in mind that such calculations can also be used to determine the sample size needed for the pilot. Likewise, policy-makers have their own hypotheses; they specify causal chains when they design policy:

> "Personal Advisers will have a central role in helping IB customers prepare for and seek work, and in supporting both individuals and their employers so that employment is sustained. This will be achieved through providing advice related to clients' social and health circumstances, developing their skills and potential, and matching clients with the needs of employers" (DWP-1-TS, p.10).

This echoes what the LSE GV314 Group found:

> "A key distinction in my experience is between commissioners of research and their policy counterparts. It's the latter who are often the trickier to handle, whilst the former sometimes even see themselves as protecting research integrity against the demands of the policy people. This was certainly my experience of doing work… in a politically contentious area" LSE GV314 study

The LSE GV314 Group survey makes a distinction between "those expected to be more sensitive to the political ammunition aspects of the research, above all the policy officials and politicians, taking part in the design of research questions at the beginning and the writing up and reporting at the end", and those more committed to the programme evaluation, above all researchers and research managers.

7.8. Conclusion

This chapter was designed to examine whether, in the context of studies with non-significant primary outcome, politically salient reforms were subject to more spinning than reforms with lower salience. Overall, I found little evidence of spin in the six studies that I reviewed. Out of the seven indicators of spin suggested by the medical literature, one could not be verified given the information available (missing outcomes), four led to a forthright rejection (incomplete reporting, within-group comparisons, spurious subgroup analyses, upgrading/downgrading outcomes) and two found evidence of spin (interpretation of results and conclusion bias). The notion of 'spin' here is not a moral judgement; it indicates that the reporting decisions made in this instance, for whatever reason, diverged from the norms imposed by the scientific method.

The initial hypothesis that high-salience reform would be subject to more spin than low-salience reforms can be clearly rejected – at least on the basis of the present evidence. Indeed, none of the six indicators of spin used in

this study seems to show any form of association. The fact that a 'high-salience' pilot (DWP-1) was more spun than others is not sufficient in itself to validate my theory and looks, on balance, fairly anecdotal. The opposite claim that high-salience reforms would be associated with high-quality reports (in terms of compliance with the principles of scientific reporting) is not supported either.

Unfortunately, I have some reasons to believe that this study did not fully answer the question asked in the introduction. Firstly, the absence of formal research protocols did not allow me to understand what type of information would have been reported, and in what way, had the intervention had a positive and significant effect on its target group. Although technical specifications are useful documents, they cannot be considered as a proxy for research protocol. As a result, the crucial question of missing outcomes could not be answered. Secondly, this research was hampered by the lack of consistency in the presentation of reports as well as the insufficient transparency in research decisions (e.g. no justification for the choice of subgroup analyses and the expected effect). Thirdly, the design of the study did not allow me to make inferences about the vast and ever-increasing amount of evaluation reports commissioned by the UK government. Thus, it is possible that the findings of this study are due to chance and that another set of studies would have yielded different results. A larger and more systematic analysis is needed to test that hypothesis. Finally, a broader research scope would allow me to test whether these findings hold across governments, policy areas and jurisdictions.

Appendix – Reviewed documents

Pathways to Work (DWP-1)

Bewley, H, Dorsett, R, Haile, G (2007). *The impact of Pathways to Work*. Department for Work and Pensions Research Report No 435. Available here.

Department for Work and Pensions (2003). *Evaluation of Incapacity Benefit Pilots. Invitation to Tender*. Unpublished document.

JRRP (DWP-2)

Purdon, S, Stratford, N, Taylor, R, Natarajan, L, Bell, S, Wittenburg, D (2006). *Impacts of the Job Retention and Rehabilitation Pilot*. Department for Work and Pensions Research Report No 342. Available here.

Two Year-Old Education Pilot (DFE-1)

Smith, R, Purdon, S, Schneider, V, La Valle, I, Wollny, I, Owen, R, Bryson, C, Mathers, S, Sylva, K, Lloyd, E (2009). Early Education Pilot for Two Year Old Children. Department for Children, Schools and Families Research Report RR134. Available here.

Department for Children, Schools and Families (2006). *Evaluation of the 2 year old early education pilot: Specification of Requirements*. Unpublished.

Multidimensional Treatment Foster Care (DFE-2)

Biehal, N, Dixon, J, Parry, E, Sinclair, I, Green, J, Roberts, C, Kay, C, Rothwell, J, Kapadia, D, Roby, A (2012). The Care Placements Evaluation (CaPE). Evaluation of Multidimensional Treatment Foster Care for Adolescents (MTFC-A). Department for Education Research Report 194. Available here.

Biehal, N, Dixon, J, Parry, E, Sinclair, I, Green, J, Roberts, C, Kay, C, Rothwell, J, Kapadia, D, Roby, A (2012). The Care Placements Evaluation (CaPE). Evaluation of Multidimensional Treatment Foster Care for Adolescents (MTFC-A). Department for Education Research Brief 194. Available here.

Alcohol Arrest Referral (HOM-1)

McCracken, K, McMurran, M, Winlow, S, Sassi, F, McCarthy, K (2012). *Evaluation of Alcohol Arrest Referral Pilot Schemes (Phase 2).* Home Office Occasional Paper 102. Available here.

Restorative Justice Pilots (HOM-2)

Shapland, J, Atkinson, A, Atkinson, H, Dignan, J, Edwards, L, Hibbert, J, Howes, M, Johnstone, J, Robinson, G, Sorsby, A (2008). *Does restorative justice affect reconviction? The fourth report from the evaluation of three schemes.* Ministry of Justice Research Series 10/08. Available here.

[This page was intentionally left blank]

8. Conclusion

It is now time to conclude. Before I sum up the main findings of this thesis and discuss its implications, it is useful to recall its specifications. The Central Thesis Question was to assess the risk of confirmation bias in government-funded policy evaluations. This goal was broken down into two more definite objectives. First, my thesis assessed the scientific credibility of a sample of government-sponsored policy evaluations. Three common scientific prescriptions were considered: the proportionality of time frames to the scope of the project; the representativeness of pilot sites; and the comprehensiveness of outcome reports (Specific Questions 1a, 1b, 1c). Second, it examined whether the known commitment of the government to a reform was associated with less credible evaluations (Specific Questions 2a, 2b, 2c). The operational map is presented again as a reminder (see Exhibit 31).

This conclusion chapter is organised as follows. Section 8.1 summarises the main findings. Section 8.2 considers the broader implications of this thesis. Section 8.3 draws some conclusions regarding the methodology used in this study. Sections 8.4 and 8.5 provide some recommendations for practitioners and researchers interested in the subject. Section 8.6 presents some final thoughts.

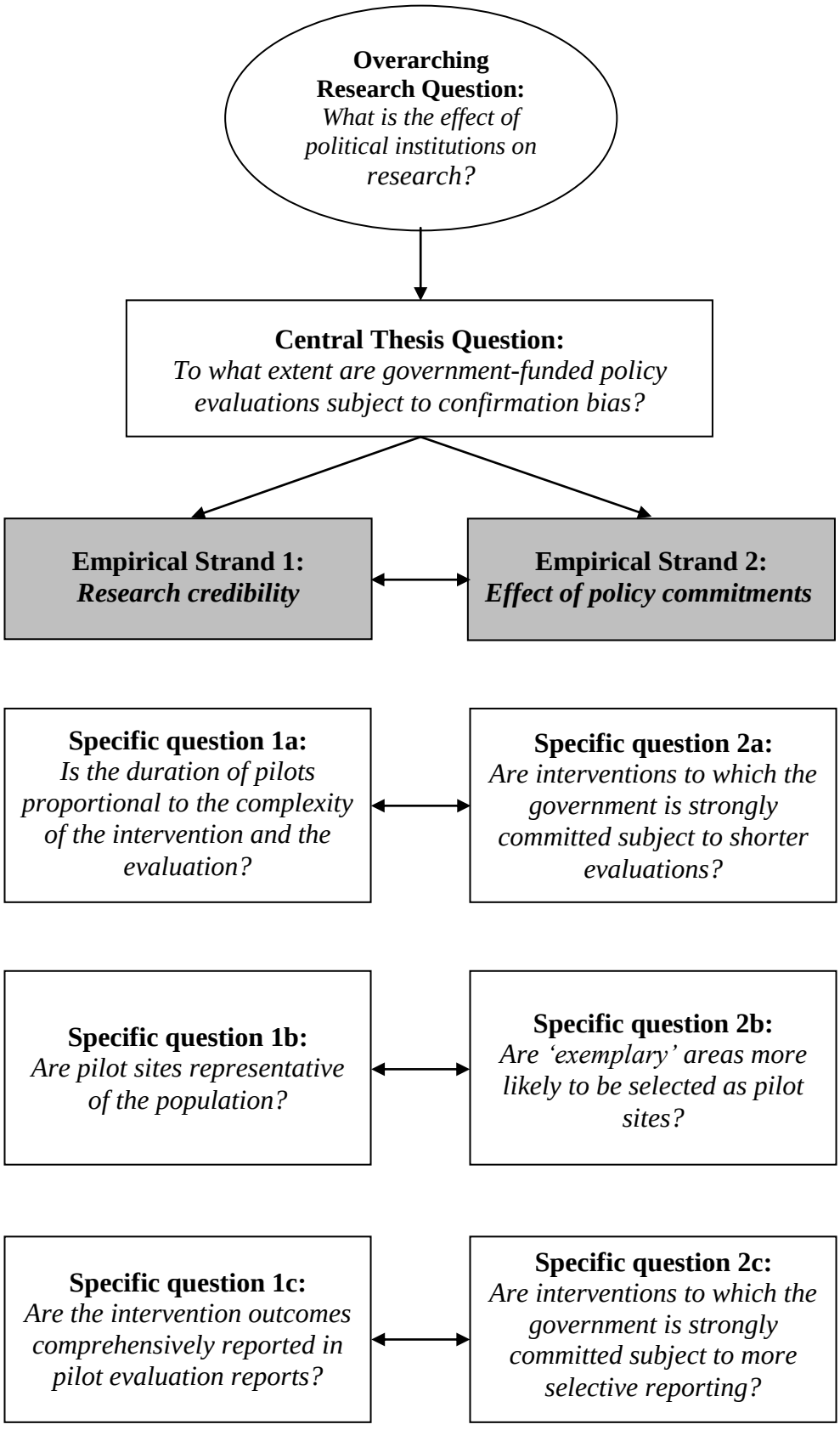
8.1. Findings

8.1.1. Scientific credibility of government-sponsored policy evaluations (Empirical Strand 1)

This thesis addressed two series of questions. The first question was whether, in the context of New Labour's Britain, policy evaluations commissioned by the government were credible from a scientific viewpoint. This question was justified by the government's open commitment to evidence-based policy, by the 'routinisation' of evaluation, and above all by the fact that virtually all policy evaluations were published during that time (and still are). This commitment to transparency could be the single most important difference between policy research and clinical trials, where evidence suggests that a significant number of studies are unpublished (Kerry Dwan et al., 2008; Easterbrook et al., 1991). On the other hand, the poor quality of evaluation reports as well as their inconsistent presentation has been a serious impediment to this research project (see section 8.3.2).

Overall, the evaluation reports I reviewed were found to be, on average, relatively credible from a researcher's point of view (Specific Question 1a).

Exhibit 31 – Operationalisation (reminder)



Results show that, on average, the duration of a pilot was proportional with the scope of the research. Controlling for the department and a number of other variables, evaluations of longer-term effects were associated with longer pilots, in line with scientific norms. Besides, I found that less than 5% of the pilots were truncated. However, I also found a great variation in the duration of those pilots, with a third lasting for 12 months or less and some others lasting for up to four years. These findings suggest that government-sponsored evaluations can serve vastly different purposes, from answering very narrow and 'simple' questions to much more demanding and sophisticated ones. This thesis is the first contribution to our understanding of how resources like time are allocated to policy research.

I also found that government-sponsored policy evaluations presented a limited risk of outcome reporting bias or 'spin' (Specific Question 1c). The qualitative nature of this work does not allow me to make any inference beyond the sample, even though studies were systematically selected. Out of the seven indicators of spin suggested by the medical literature, one could not be verified given the information available (missing outcomes), four led to a forthright rejection (incomplete reporting, within-group comparisons, spurious subgroup analyses, upgrading/ downgrading outcomes) and two found evidence of spin (interpretation of results and conclusion bias). Although there have been some accounts of spin in the literature (LSE GV314), it is the first time that spin is analysed using objective and 'structured' data. The findings suggest that spin is less prevalent in policy research than in clinical trials.

Out of the three research decisions considered, the only one that does not seem primarily driven by scientific considerations is the selection of pilot sites (Specific Question 1b). Chapter 7 shows that pilot sites are almost never sampled using a probability formula and as such, are unlikely to be representative of the UK as a whole. Regardless of the motive, this finding has important implications. Empirically, it suggests that the results of these pilot evaluations cannot be straightforwardly extended to the rest of the territory. In other words, the argument that the intervention will work 'there' because it worked 'here' is flawed. From a theoretical viewpoint, this conclusion supports a point previously made by Cartwright and Hardie regarding the limited external validity of policy experiments (Cartwright & Hardie, 2012).

8.1.2. Effect of policy commitments on the scientific credibility of evaluations (Empirical Strand 2)

The second empirical strand of this thesis concerned the effect of policy commitments on the scientific credibility of evaluations. This question was justified by the strong emphasis on performance and delivery under the Labour government as well as the seniority of generalist bureaucrats over specialists like analysts and various other incentives.

Overall, the effect of policy commitment on the credibility of evaluations was found to be weak.

Policy commitments had no significant effect on the selection of pilot sites (Specific Question 2b). This means that, even if pilot districts were not selected in the way that would warrant representativeness, they were not selected for exemplarity either, at least as measured in this thesis. The fact that each additional pilot being run in a given district significantly reduces the odds of this district being selected again for a new pilot can be interpreted in two different ways. It could be to prevent a risk of 'contamination', which would be advisable from a scientific viewpoint. Or it could be for managerial reasons, to make sure that the administrative burden of implementing pilot interventions be evenly distributed across the territory. Regardless of the motive, this conclusion differs from those of the medical literature.

Policy commitment did not seem to affect the reporting of outcomes either (Specific Question 2c). However, I have already mentioned that the level of spin across the six studies that I reviewed was low, so even if there were an association in the population, it would be difficult to see in the context of this qualitative study. More research is needed to quantify the strength and the significance of policy commitments on spin.

The evidence regarding the effect of policy commitments on the duration of pilots is mixed (Specific Question 2a). On the one hand, I found that reforms to which the government was committed were subjected to significantly shorter pilots, even after controlling for the research question, the department and the prominence of the pilot. On the other hand, I did not find that pilots related to a pre-election pledge were shorter pilots which were not. Moreover, I found that neither the electoral cycle nor the 'salience' of the intervention had a significant effect on the duration of pilots as suggested in the rest of the literature.

8.2. Theoretical implications

8.2.1. How much confirmation bias in policy research (Central Thesis Question)?

What do the above results tell us about the extent of confirmation bias in government-funded policy evaluation? Such an appraisal is a difficult exercise. It is important to bear in mind that the results of a study may in fact be unbiased despite a methodological flaw. Thus, a non-zero risk of bias does not necessarily imply biased conclusions. The design of my thesis did not allow quantifying the 'amount of bias' caused by the fact that the government evaluated its own policies.

Another problem is that there is no real precedent or benchmark. To the extent that the comparison is relevant, I would simply argue that this effect seems to be more *limited* (in terms of scope) than the effect of industry sponsorship on the credibility of clinical trials. My research shows that the effect of policy commitments is only tangible at the beginning of the research cycle, where policy and analytical teams have to work together and make compromises. Later decisions would appear to be more immune to policy commitments. In contrast, the literature suggests that industry sponsorship affects all clinical trial decisions including duration, sampling and outcome reporting.

8.2.2. More confirmation bias than meets the eye?

My research has shown that, *prima facie*, policy commitments do not significantly affect the course of evaluation. However, it has also suggested that confirmation bias might concern research decisions taken both earlier and later than the decisions I chose to analyse.

Earlier, institutions affect the free course of research by imposing strict constraints on *what to study* and *what not to study*. This is in stark contrast with the principle of 'academic freedom', which underpins much of the research conducted in universities. The fact that all pilot interruptions were decided by the new government in May 2010 is a reminder that, in an institutional context, research is not conducted for the sake of knowledge, but to support the government's policy initiatives. In the medical sphere, this bias has for a long time manifested itself in an under-investment in research on rare or neglected diseases (LaMattina, 2012; Rockoff, 2013).

Later, it was suggested that confirmation bias was less visible in the actual studies than in the documents communicating the results to stakeholders and the public at large (see chapter 7).

Last but not least, the effect of confirmation bias is both direct (e.g. by imposing shorter timeframes to some pilots), and indirect – through the allocation of human and financial resources.

8.2.3. Implications beyond the case

To what extent are these findings capable of extension to other jurisdictions? While the direct application of my findings is likely to be limited to policy evaluation in the UK, it does hold implications for other jurisdictions and other research areas, such as clinical trials.

My thesis suggests that there are three main forces that increase the scientific credibility of research and its immunity to policy commitments.

The first is a significant proportion of individuals with a scientific background within the senior management of the organisation. The case of the DWP has been cited as an example of good practice on several occasions. However, I did not find empirical evidence that the studies commissioned by this department were more credible than others. Alternatively, a decent understanding of the scientific method and its requirements and among those – product managers or policy-makers – who commission research.

The second is a high-level commitment to transparency. In the UK, this is enforced through both 'hard' regulation (Freedom of Information Act) and 'soft' regulation (GSR publication guidelines).

The third is the independence of researchers. This independence must be statutory (i.e. researchers cannot be hired, promoted or demoted for reasons other than the quality of their research) and, above all financial, with limited incentives to supply favourable results. Disclosing payments made to researchers has been suggested as a way of making research more accountable (Rao & Sant Cassia, 2012).

8.3. Methodological lessons

Given that, to the best of my knowledge, it is the first time that policy research decisions are studied from a meta-research perspective, it is important to reflect on the added value of the methodology and share my experience.

The *ex-ante* assessment of meta-research showed that the method had two critical advantages over the interviews and questionnaires which have dominated the literature so far. First, it is based on *observed* research decisions as opposed to *accounts* of research decisions, which can often be biased by selective memories and social desirability. Second, this data was systematically collected, which means that there is a limited risk of sampling bias.

Ex post, the PILOT dataset proved to be a useful tool in the study of confirmation bias in policy research. New and important questions have been answered – as shown in sections 8.1 and 8.2.

However, the limitations of the method must also be acknowledged. In particular, the coding of variables has been long and difficult due to the inconsistent quality of government-sponsored evaluation reports. The significant amount of missing information, as well as the occasional factual errors found after cross-checks, suggests that double-coding should be used. Still, the method might be hard to replicate in the future.

8.4. Recommendations for future practice

The above-mentioned results and caveats suggest that no major institutional reform would be justified in the UK. However, some adjustments would dramatically improve the transparency of the evaluation process, and thus, its trustworthiness. Such adjustments include (1) the design and publication of research protocols; (2) the publication of reporting guidelines; and (3) the publication of results in scientific journals and/or research repositories.

8.4.1. Publication of research protocols

Government can hide important information by publishing evaluation reports ex post and 'adjust' its initial intentions to fit the new circumstances. This is known as the 'black box' problem, which has been mentioned several times throughout this thesis.

Government would be wise to require that all research projects be registered by the government *before* their implementation as opposed to after. Protocols specify the time that each phase of the project is likely to take, along with a detailed month by month timeline for each activity to be undertaken. Subsequent modifications to this protocol are then mentioned in the protocol and justified. Furthermore, peer-reviewed research publications indicate the dates when the manuscript was first submitted and published as well as any other interim step such as revisions. Furthermore, technical specifications should be published in the cases where those evaluations are contracted out.

8.4.2. Transparent reporting

The Government Social Research website[15] provides a wealth of professional documents to its members and the general public, including: methodological handbooks, ethical guidelines, publication guides, etc. However, none of these documents address the question of what information a standard evaluation report should contain.

Government should encourage the standardisation of evaluation reports. This could be could achieved in two ways. First, each government department could sign and publish a common statement of their commitment to transparent reporting, such as the CONSORT statement[16]. Second, Government should provide each of its agencies and departments with reporting guidelines such as the CONSORT checklist[17].

[15] http://www.civilservice.gov.uk/networks/gsr/publications
[16] http://www.consort-statement.org/
[17] http://www.consort-statement.org/checklists/view/32-consort/66-title

There is some evidence that reporting guidelines are associated with more comprehensive studies. A 2012 Cochrane systematic review assessed the effect of journal's endorsement of CONSORT on the reporting of trials they publish (Turner, Shamseer, Altman, Schulz, & Moher, 2012). In 50 included studies evaluating the reporting of 16,604 trials, 25/27 CONSORT-related items measured were more completely reported in trials published in endorsing journals than those in non-endorsing journals, five items were significantly better reported. Similar findings were yielded for many items when comparing trials published in journals before and after CONSORT endorsement.

8.4.3. Publication of findings in scientific journals

Studies evaluating the effect of policy interventions are published by the relevant government departments. That has two implications. On one level, this means that these studies are subjected to the different levels of quality control and reporting requirements. On another level, this makes research synthesis more difficult and costlier than if these results were available in a unique location.

Government should encourage researchers to publish their results in journals. Alternatively, studies published by government departments should also be systematically made available in research repositories such as 3ie's *Impact Evaluation Repository*[18]. This would simplify systematic reviews and meta-evaluations.

8.5. Directions for further research

From one point of view, these findings are an important step forward in the study of public administration and public policy. From another point of view, they barely scratch the surface. I see five main directions for future research in this area.

8.5.1. Geographical scope

This thesis focused on the UK during the Labour government (1997-2010). The motivations and the case selection process have been described in section 3.1. This relatively narrow scope has had some benefits. In particular, I believe that it has enhanced the conclusion validity of the findings. Indeed, the research done on the political and institutional context (presented in chapter 4) has been instrumental in the interpretation of findings from the empirical analysis. However, it also means that the results

[18] http://www.3ieimpact.org/evidence/impact-evaluations/impact-evaluation-repository/

can hardly be extended beyond the case. Future research in this area would benefit from a broader scope. Introducing more variation in terms of geography and time would result in larger datasets and greater statistical power. It would also allow comparisons across polities, governments and types of civil service.

8.5.2. Independent variable

In this study, I used the variation in the government's known commitment to a reform (strong/weak) as a proxy for the strength of confirmation bias. Although not entirely satisfying from a construct validity viewpoint, this decision was constrained by the limited number of policy evaluations undertaken by independent entities in the selected geography, time frame and policy areas. The loosening of the time and geographical constraints discussed in section 8.5.1 would make the alternative more feasible. Future research in this area would thus benefit from contrasting policy evaluations sponsored by government with similar evaluations sponsored by non-governmental organisations (academia, think tanks, charities, etc.). As argued repeatedly throughout this thesis, research sponsorship has been a strong predictor of clinical trial outcomes. Using policy commitment as independent variable should be used in last resort, when it is not possible to contrast different types of sponsorship.

8.5.3. Dependent variables

This study examined the effect of policy commitments on the scientific credibility of evaluations based on three research decisions: the time afforded to research, the selection of pilot sites; and the level of spin. Ultimately, though, the extent of confirmation bias is best measured using study outcomes. Future research in this area would benefit from addressing the following lines of inquiry:

- Is government sponsorship of a policy evaluation associated with favourable outcomes?
- Is government sponsorship of a cost-benefit analysis associated with favourable reported ratios?

Other interesting questions include:

- Are international economic development programmes associated with more robust impact evaluations than equivalent domestic interventions?
- How are blame and credit attributed in implementation evaluations?

8.5.4. Using existing data as opposed to collecting new data

This study involved the development of a new dataset of over 230 studies. As explained in section 8.3, data collection has been a painstakingly slow process. This compromises the replicability and the scalability of this type of research. Again, loosening the time and geographical constraint opens up a number of opportunities. Future research in this area could be made more cost-effective by using existing systematic reviews, such as those published by the Campbell Collaboration and by the various 'What Works' centres in the US and the UK (What Works Clearinghouse, Development Experience Clearinghouse, Early Intervention Foundation, What Works Centre for Local Economic Growth, etc.). In the area of social research, these systematic reviews usually include a fairly large number of studies funded by different sources (governmental vs. non-governmental). The second benefit of using existing systematic reviews is that they report many of the variables that I would need for my own research (e.g. effect, sample size, type of research design, etc.).

8.5.5. Why is this so?

Last but not least, there is a need for more qualitative research into the different factors that strengthen or weaken confirmation bias. This thesis has shown that government-sponsored policy evaluations were very diverse in terms of depth and thoroughness. Although trying to identify patterns and means is interesting *per se*, we would also learn a great deal by doing more exploratory research on a number of systematically selected cases. This would be best achieved through comparisons of different organisational contexts (e.g. governmental/ industrial/charitable) and research areas (e.g. social research vs. science and technology studies). Interviews and case studies are seen as particularly desirable methods to address this question.

8.5.6. Use of findings

In addition to conducting more research on confirmation bias, it would be interesting to measure the extent to which the conclusions of policy evaluations inform subsequent policy decisions. This is particularly relevant in the case of pilot evaluations, given that these studies are meant to address a straight-forward question, namely: should these pilot interventions be rolled out nationally?

8.6. Final word

It bears repeating: on the whole, government-funded policy evaluations observe a number of important scientific prescriptions and are only marginally affected by previous policy commitments.

I am aware that this conclusion will disappoint some. The popular assumption among viewers of TV series such as 'Yes Minister' and 'The Thick of It' is that politics (in a derogatory sense) permeates all levels of governance and all steps of the policy cycle, from agenda-setting to lesson-learning. To be clear, I am not suggesting that policy evaluation is immune to politics – indeed, this thesis has shown that higher-level constraints (resources, time, people) were often in the way. However political institutions appear to have a more limited effect on research than the market.

The reader will remember that this thesis began with two opposed views on the role of researchers in the policy process. A first view (David Blunkett) was that social scientists were given enough autonomy to inform policy with their research. A second view (Eric Pickles) was that researchers were 'on tap' and politicians 'on top'. I believe and I hope that readers are now better equipped to make their own judgment.

Annex I – List of pilots included in the PILOT dataset

1. 14-19 Pathfinder Initiative
2. Achievement for All
3. Action Teams for Jobs
4. Activity Agreements Pilots
5. Adult Basic Skills Extension Pathfinder
6. Adult Guidance Pilots
7. Adult Learning Grant
8. Adult Learning Option
9. Adults Facing Chronic Exclusion
10. Aiming High: African Caribbean Achievement Project
11. Alcohol Arrest Referral - Phase 1
12. Alcohol Arrest Referral - Phase 2
13. Ambition programme
14. Anti-social behaviour co-ordinators
15. Automatic Referral to Mediation
16. Better-off In-Work Credit
17. Black Children's Achievement Programme
18. Boarding School Provision for Vulnerable Children Pathfinder
19. Budget Holding Lead Professionals
20. Budget Holding Lead Professionals for Children in Care
21. CONNECT
22. Care First Careers pilot
23. Child Development Grant
24. Child Sex Offender Review (CSOR) Public Disclosure Pilots
25. Childcare Affordability Pilots 2009 - 100%
26. Childcare Affordability Pilots 2009 - Actual Costs
27. Childcare Affordability Pilots 2009 - Disabled Children
28. Childcare Affordability Programme
29. Childcare Taster Pilot
30. Children's Trust Pathfinder
31. Choice Advice Service
32. City Strategy
33. Cognitive Skills Booster (CSB) programme
34. Community Finance and Learning Initiative
35. Community Safety Partnerships
36. Community Support Officers
37. Conditional Cautions scheme
38. Connexions Customer Information System
39. Connexions Direct Pilot
40. Connexions Service Pilots
41. Dedicated Drug Courts
42. Dedicated Sexual Assault Unit

43. Devolution of Education Welfare Services to secondary schools
44. Disabled Children's Access to Childcare
45. Diversity Pathfinders
46. Drug Testing Pilot Programme
47. Drug Treatment and Testing Orders
48. Drug Treatment and Testing Requirements
49. Drug and Alcohol Courts Pilot
50. ESOL Pathfinder
51. Early Education Pilot for Two Year Old Children
52. Early Excellence Centre
53. Early Professional Development
54. Early Support Pilot Programme
55. Early neutral evaluation pilot
56. Education Business Link
57. Education Maintenance Allowance
58. Effective bail scheme
59. Electronically Monitored Curfew
60. Employer Training Pilots
61. Employment Advisers in GP surgeries (Pathways Advisory)
62. Employment Pathfinder - Phase 1
63. Employment Pathfinder - Phase 2
64. Employment Retention and Advancement
65. Empowering Young People Pilots
66. Entry to Learning
67. Ethnic Minority Outreach
68. Evaluating judicial mediation in employment tribunals
69. Every Child Counts
70. Every Child a Reader
71. Every Child a Writer
72. Excellence Fellowship Awards
73. Exit to Work
74. Extended Flexible Entitlement for 3 and 4 YO pathfinder
75. Extended Schools Childcare Pilot
76. Extended Schools Pathfinder
77. Extended Services Subsidy pathfinder
78. Extended Telephone Support Service Pilot
79. Extra Mile
80. Face-to-Face Guidance Pilot
81. Fair Cities
82. Family Nurse Partnership
83. Family Resolutions Pilot Project
84. Family and Young Carer Pathfinder
85. Fast Track to Prosecution for School Non-Attendance Pathfinder
86. Find Your Talent
87. Fine Payment Work
88. Fit For Work Service
89. Forensic Science Service Pathfinder
90. Formalised Peer Mentoring Pilot Evaluation

91. Fortnightly jobsearch review
92. Free School Meals Pilot
93. Helping Children Achieve Trial
94. Helping Families Programme
95. Higher Level Basic Skills Pilots
96. Home Access Programme Pathfinder
97. I-Sign project
98. In-Work Retention Pilot
99. Increasing take-up of formal childcare in BME communities
100. Individual Budgets for families with disabled children
101. Individual Learning Accounts
102. Inform, Persuade and Remind campaign
103. Integrated Domestic Abuse Programme
104. Integrated Domestic Violence Court (Croydon)
105. Integrated Employment and Skills
106. Integrated Offender Management pioneer
107. Intensive Activity Period mandatory for the 50+
108. Intensive Alternatives to Custody pilots
109. Intensive Control and Change Programme
110. Intermittent Custody Pilot
111. Invest to Save Pathfinders
112. JCP Intensive Activity trial for substance misusing customers
113. JRFND [From W25] - Skills Conditionality Pilot
114. Job Retention and Rehabilitation Programme
115. Jobcentre Plus pathfinder
116. Jobseeker Mandatory Activity
117. Justice Research Consortium
118. Key Stage 2 career-related learning pathfinder
119. Leadership and Management Development Programme
120. Learning (Connexions) Card Demonstration
121. Learning Agreement Pilots
122. Link Up
123. Local Authority Child Poverty Innovation Pilot
124. Local Authority Commissioning Pathfinders
125. Local Housing Allowance Pathfinders
126. Lone Parents New Services
127. Lone-parents pilots (IWC, WSP, ESC, ND+fLP)
128. Low Attainers Pilots
129. Making Good Progress
130. Mandatory basic skills pilot - Benefit sanctions
131. Mandatory polygraph testing
132. Mental Health Courts
133. Mentor Points
134. Multi-systemic Therapy Pilot
135. Multidimensional Treatment Foster Care
136. New Deal for the Long-Term Unemployed ND25+ - Pilot
137. New Deal for the Long-Term Unemployed ND25+ - Gateway
138. New Deal for the 50+ - Pathfinder

139. New Deal for the 50+ - Over 50s Outreach Pilot
140. New Deal for Disabled People
141. New Deal for Lone Parents Pathfinders
142. New Deal for Young People Pathfinders
143. New Deal for Young People - Intensive Gateway
144. NOMS Offender Management Model
145. National Reassurance Policing Programme
146. Neighbourhood Agreements Pathfinder Programme
147. New Deal for Partners
148. New Entrepreneur Scholarships
149. New Jobseeker Regime
150. Next Steps test bed Regional Pilots
151. Numeracy/literacy pilots
152. ONE Service
153. Occupational Health Advice Lines pilot
154. Off the streets and into work
155. On-charge drug testing
156. Outcomes for Learners Pathfinders
157. Parent Support Advisor Pilot
158. Parenting Early Intervention Programme pathfinder
159. Partners Outreach for Ethnic Minorities
160. Pathfinder UK Online Centres
161. Pathways to work pilots
162. Penalty notices for disorder on 10- to 15-year-olds
163. Personalised Employment Programme pilots
164. Pilot Beacon Schools
165. Police And Criminal Evidence
166. Post 16 Equal Opportunities Pilots
167. Postal requisitioning
168. Progress
169. Progress File
170. Progression to Work Pathfinders
171. Public Law Outline in family courts
172. Pupil Learning Credits
173. REMEDI
174. Raising the Achievement of Bilingual Learners
175. Raising the Participation Age (RPA) Phased intro
176. Re-Ach Project
177. Referal Orders for 10-17 year olds
178. Referral Orders for 10-17 year olds
179. Repayment of Teachers' Loans Scheme
180. Resettlement Pathfinders
181. Restriction on Bail [DIP]
182. Right2BCared4 pilots
183. Satellite Tracking of Offenders
184. Saving Gateway 1 (SG1)
185. Schema modal therapy in a high-secure hospital
186. School Gate Employment Support (CPP)

187. Schools Linking Network
188. Schools Plus Teams Pilot
189. Second Great Parenting Experiment
190. Secondary Social Emotional and Behavioural Skills
191. Services for Separating Parents
192. Short breaks Pathfinder
193. Single Level Test Pilot
194. Skills Coaching and Passport
195. Small Firm Development Account
196. Social Pedagogy Pilot Programme
197. Soft Skills Pilot
198. Specialist Employment Adviser Programme
199. Stable and Acute' risk assessment pilot
200. Statutory Time Limits in the Youth Court
201. Staying Put: 18 Plus Family Placement Programme pilot
202. StepUP programme
203. Study Plus Pilots
204. Support Childminder Pathfinder
205. Support to victims of road accidents
206. Sure Start Mainstreaming Pilots
207. Sure Start Plus
208. Tackling Knives and Serious Youth Violence Action Programme
209. Targeted Mental Health in Schools
210. Targeted Youth Support Pathfinders
211. Teenage Parent Supported Housing (TPSH) pilot
212. The 'Go-Between' pathfinder projects
213. The impact of debt advice
214. Time to Talk
215. Together Women
216. Transition Information Sessions
217. Travel to Interview Scheme
218. Trust School Pathfinder
219. UK Resilience Programme
220. Understanding Connexions Pilot
221. Unpaid reparative work caution
222. Victims' Advocate Scheme pilots
223. Virtual School Head
224. Virtual courts pilot
225. Welfare Reform Drug Recovery Pilot
226. Women specific condition pilot
227. Work for Your Benefit
228. Work works pilots (Discovery Weeks, ESC+Childcare Tasters)
229. Work-focused services in children's centres
230. Working Neighbourhoods Pilot
231. Young Volunteer Challenge
232. Young witness support
233. Youth Offending Teams

Annex II – PILOT Codebook

Pilot duration (planned)

Definition:	Pilot duration in months, from start to *planned* finish, i.e. taking into account extensions and truncations				
Source:	Evaluation reports				
Type:	Continuous				
Range:	[2, 48]		Unit:	1	
Unique values:	34		Missing:	16/233	
Mean:	19.8		Std dev.:	10.6	
Percentiles:	10%	25%	50%	75%	90%
	7	12	18	24	36

Pilot duration (observed)

Definition:	Pilot duration in months, from start to *actual* finish, i.e. taking into account extensions and truncations				
Source:	Evaluation reports				
Type:	Continuous				
Range:	[0, 60]		Unit:	1	
Unique values:	38		Missing:	18/233	
Mean:	19.9		Std dev.:	11.8	
Percentiles:	10%	25%	50%	75%	90%
	6	12	18	24	36

Pathfinder

Definition:	Whether the government has expressed an explicit commitment to roll out the intervention, or whether a roll-out schedule is mentioned in the evaluation study		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	0/233
Values:	[0] No	[1] Yes	
Tabulation:	Frequency	Value	
	177	0	
	56	1	

Manifesto

Definition:	Whether the intervention and its objective were mentioned in the manifesto of the Labour Party for the previous general election		
Source:	Labour Party manifestos (1997, 2001, 2005)		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	0/233
Values:	[0] No [1] Yes		

Tabulation:	Frequency	Value
	121	0
	112	1

Election

Definition:	Number of months between the start of the pilot and the next general election				
Type:	Continuous				
Range:	[0, 58]		Unit:	1	
Unique values:	53		Missing:	7/233	
Mean:	25.4		Std dev.:	14.8	
Percentiles:	10%	25%	50%	75%	90%
	6	13	24	38	47

Term

Definition:	Labour's term in government		
Type:	Categorical		
Range:	[1, 3]	Unit:	1
Unique values:		Missing:	4/233
Values: Frequency (1):	[1] First term: May 1997 to May 2001 [2] Second term: May 2001 to May 2005 [3] Third term: May 2005 to May 2010		

Tabulation:	Frequency	Value
	38	1
	78	2
	113	3

Type of evaluation

Definition:	Whether the evaluation covers questions pertaining to the process, the outcome or the impact of the intervention. Impact is considered measurable only based on a counterfactual. The dataset records the 'highest' type of design.		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[1, 3]	Unit:	1
Unique values:	3	Missing:	18/233
Values:	[1] Process evaluation [2] Outcome evaluation [3] Impact evaluation		

Tabulation:	Frequency	Value
	112	1
	46	2
	57	3

Department

Definition:	Department sponsoring the evaluation		
Source	Evaluation reports		
Type:	Categorical		
Range:	[1, 4]	Unit:	1
Unique values:	4	Missing:	0/233
Values:	[1] Department for Education [2] Department for Work and Pensions [3] Home Office [4] Ministry of Justice		

Tabulation:	Frequency	Value
	114	1
	58	2
	26	3
	35	4

Mandatory intervention (for DWP studies only)

Definition:	Whether the intervention is mandatory to its target groups and entails a sanction for non-compliance		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	0/54
Values:	[0] No	[1] Yes	
Tabulation:	Frequency	Value	
	34	0	
	20	1	

DWP target group (for DWP studies only)

Definition:	Group targeted by the intervention, according to the DWP Job Outcome classification (highest category recorded)		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[1, 12]	Unit:	1
Unique values:	5	Missing:	4/54
Values:	[1] Employed customers [3] Unemployed customers not claiming benefits [4] Customers claiming JSA for under 6 months [8] Customers on New Deal or claiming JSA for over six months [12] Lone parents, people with a health condition/disability and other inactive benefit customers		
Tabulation:	Frequency	Value	
	2	1	
	1	3	
	10	4	
	22	8	
	15	12	

Chapter 6 – Selection of pilot sites

Pilot

Definition:	Whether Jobcentre Plus District *i* was selected as pilot site for pilot *j*		
Type:	Categorical		
Range:	[0,1]	Unit:	1
Unique values:	2	Missing:	0/2000
Values	[0] No	[1] Yes	

Tabulation:	Frequency	Value
	1589	0
	411	1

Region

Definition:	Region of England where the District is located		
Type:	Categorical		
Range:	[1,4]	Unit:	1
Unique values:	4	Missing:	0/2000
Values	[1] North of England, includes: Cheshire, Halton and Warrington; Cumbria and Lancashire; Manchester (Central); Manchester (East and West); Merseyside; North East Yorkshire and the Humber; Northumbria; South Tyne and Wear Valley; South Yorkshire; Tees Valley; West Yorkshire. [2] Midlands, includes: Birmingham and Solihull; Black Country; Coventry and Warwickshire; Derbyshire; Leicestershire, Rutland and Northamptonshire; Marches; Nottinghamshire and Lincolnshire; Staffordshire. [3] London, includes: Brent, Harrow and Hillingdon; Central; City and East; Lambeth, Southwark and Wandsworth; North East; North; South East; South; West. [4] South of England, includes: Bedfordshire and Hertfordshire; Cambridgeshire and Suffolk; Devon and Cornwall; Dorset and Somerset; Essex; Gloucestershire, Wiltshire and Swindon; Hampshire and Isle of Wight; Kent; Norfolk; Surrey and Sussex; Thames Valley (Berks, Bucks, Oxfs); West of England.		

Tabulation:	Frequency	Value
	550	1
	400	2
	450	3
	600	4

Pathfinder

Definition:	Whether the government has expressed an explicit commitment to roll out the intervention, or whether a roll-out schedule is mentioned in the evaluation study		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	0/2000
Values	[0] No	[1] Yes	
Tabulation:	Frequency	Value	
	1640	0	
	360	1	

Manifesto

Definition:	Whether the intervention and its objective were mentioned in the manifesto of the Labour Party for the previous general election		
Source:	Labour Party manifestos (1997, 2001, 2005)		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	0/2000
Values	[0] No	[1] Yes	
Tabulation:	Frequency	Value	
	1320	0	
	680	1	

Mandatory

Definition:	Whether the intervention is mandatory to its target group and entails sanctions.		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[0, 1]	Unit:	1
Unique values:	2	Missing:	120/2000
Values	[0] No	[1] Yes	
Tabulation:	Frequency	Value	
	1080	0	
	800	1	

JCP Lead

Definition:	Whether Jobcentre Plus leads the implementation of the pilot (as opposed to local authorities or private-sector providers).		
Source:	Evaluation reports		
Type:	Categorical		
Range:	[0,1]	Unit:	1
Unique values:	2	Missing:	200/2000
Values	[0] No [1] Yes		

Tabulation:	Frequency	Value
	1493	0
	307	1

JSA Exit Rate

Definition:	Ratio of the number of individuals terminating their JSA claim because they found a job over the total number of JSA claimants in a given month and Jobcentre Plus District. The value included in the dataset is the annual average JSA exit rate to job of a given district the year before the start of the pilot.				
Source:	NOMIS - Official labour market statistics				
Type:	Continuous				
Range:	[3.98, 31.11]		Unit:	%	
Unique values:	433		Missing:	40/2000	
Mean:	16.34		Std dev.:	3.11	

Percentiles:	10%	25%	50%	75%	90%
	12.56	14.06	16.12	18.12	20.49

Benefit claimants

Definition:	Ratio of benefit claimants (Jobseeker Allowance, Income Support, Incapacity Benefit) in the active population of each Jobcentre Plus District in August 2007.				
Source:	NOMIS - Official labour market statistics				
Type:	Continuous				
Range:	[1.3, 7.3]		Unit:		
Unique values:	26		Missing:	0/2000	
Mean:	3.08		Std dev.:	1.43	
Percentiles:	10%	25%	50%	75%	90%
	1.55	2	2.75	3.8	5.1

Working age population

Definition:	Number of individuals aged 16 to 59 (females) or 64 (males) in 2003 in a given Jobcentre Plus District. The figure in the dataset is expressed in 100,000s.				
Source:	NOMIS - Official labour market statistics				
Type:	Continuous				
Range:	[3.98, 15.37]		Unit:	.001	
Unique values:	39		Missing:	0/2000	
Mean:	7.72		Std dev.:	2.75	
Percentiles:	10%	25%	50%	75%	90%
	4.71	5.45	7.00	9.89	11.35

Population density

Definition:	Total population per hectare in each Jobcentre Plus District				
Source	ONS, Census data 2003				
Type:	Continuous				
Range:	[1, 268]		Unit:	1	
Unique values:	20		Missing:	0/2000	
Mean:	29.2		Std dev.:	53.1	
Percentiles:	10%	25%	50%	75%	90%
	1	2	4	38	82

Ethnic white (%)

Definition:	Ratio of adults identifying themselves as white in the population of a Jobcentre Plus District in 2003.				
Source:	ONS, Census data 2003				
Type:	Continuous				
Range:	[57, 96]		Unit:	1	
Unique values:	20		Missing:	0/2000	
Mean:	85		Std dev.:	10.45	
Percentiles:	10%	25%	50%	75%	90%
	69	79	90	93	94

Capacity

Definition:	Number of pilots already running in Jobcentre Plus District *i* at the launch of pilot *j*	
Source:	Evaluation reports	
Type:	Continuous	
Range:	[0, 6]	Unit: 1
Unique values:	7	Missing: 0/2000
Mean:	0.94	Std dev.: 1.16
Tabulation:	Frequency	Value
	931	0
	591	1
	273	2
	121	3
	55	4
	23	5
	6	6

DWP target group

Definition:	Group targeted by the intervention, according to the DWP Job Outcome classification (highest category recorded)		
Type:	Categorical		
Range:	[1, 12]	Unit:	
Unique values:	5	Missing:	0/2000
Values:	[1] Employed customers [3] Unemployed customers not claiming benefits [4] Customers claiming JSA for under 6 months [8] Customers on New Deal or claiming JSA for over six months [12] Lone parents, people with a health condition/disability and other inactive benefit customers		

Tabulation:	Frequency	Value
	80	1
	200	3
	160	4
	880	8
	680	12

Annex III – List of interviewees

Name	Organisation	Interview date
Alan Marsh	Policy Studies Institute	1 February 2012
Bruce Stafford	University of Nottingham	19 January 2012
Carl Emmerson	Institute for Fiscal Studies	13 December 2011
Claire Crawford	Institute for Fiscal Studies	12 January 2012
Genevieve Knight	Policy Studies Institute	14 September 2011
Jim Hillage and Sarah Dewson	Institute for Employment Studies	6 February 2012
James Riccio	MDRC	2 May 2012
Jo Casebourne	Nesta	23 March 2012
Jonathan Portes	National Institute of Economic and Social Research	10 May 2012
Mike Daly	Department for Work and Pensions	3 October 2013
Rachel Marangozov	Institute for Employment Studies	19 February 2012
Richard Dorsett	National Institute of Economic and Social Research	15 December 2011
Rita Griffiths	CESI	6 February 2012
Stephen Morris	Policy Studies Institute	18 January 2012
Susan Purdon	Bryson Purdon Social Research	18 January 2012
Suzanne King	Freelance evaluator	4 July 2012
Vicky Davies	Ecotec	1 February 2012

Annex IV – Bibliography

Abramson, J. (2008). *Overdosed America: The Broken Promise of American Medicine*. Harper Perennial.

Alasbali, T., Smith, M., Geffen, N., Trope, G., Flanagan, J., Jin, Y., & Buys, Y. (2009). Discrepancy between results and abstract conclusions in industry- vs nonindustry-funded studies comparing topical prostaglandins. *Am J Ophthalmol.*, *147*(1), 33–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18760766

Allen, C. (2005). On the Social Relations of Contract Research Production: Power, Positionality and Epistemology in Housing and Urban Research. *Housing Studies*, *20*(6), 989–1007. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02673030500291132?journalCode=chos20#.U1FcwyiGrfg

Amara, N., Ouimet, M., & Landry, R. (2004). New Evidence on Instrumental, Conceptual, and Symbolic Utilization of University Research in Government Agencies. *Science Communication*, *26*(1), 75–106. Retrieved from http://scx.sagepub.com/content/26/1/75.short

Andersen, L. (2007). *Professional norms, public service motivation and economic incentives: What motivates public employees?* (pp. 1–20).

Angell, M. (2005). *The Truth About the Drug Companies* (2nd ed., p. 319). New York: Random House.

Artes, J. (2013). Do Spanish politicians keep their promises? *Party Politics*, *19*(1), 143–158.

Austad, K., Avorn, J., & Kesselheim, A. (2011). Medical Students' Exposure to and Attitudes about the Pharmaceutical Industry: A Systematic Review. *PLOS Medicine*, *8*(5). Retrieved from http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001037

Avorn, J. (2005). *Powerful Medicines: The Benefits, Risks, and Costs of Prescription Drugs*. Vintage.

Baba, N., Nishioka, S., Oda, N., Shirakawa, M., Ueda, K., & Ugai, H. (2005). Japan's Deflation, Problems in the Financial System, and Monetary Policy. *Monetary and Economic Studies*, *23*(1), 47–111. Retrieved from http://www.imes.boj.or.jp/research/papers/english/me23-1-2.pdf

Bailar, J. (2006). How to distort the scientific record without actually lying: truth, and arts of science. *European Journal of Oncology Nursing*, *11*(4), 217–224.

Barber, M. (2008). *Instruction to Deliver: Fighting to Transform Britain's Public Services*. Methuen Publishing Ltd.

Barnett, A., Van der Pols, J., & Dobson, A. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. Retrieved from http://ije.oxfordjournals.org/content/34/1/215.full

Barnsteiner, J., & Prevost, S. (2002). How to implement evidence-based practice. Reflections on Nursing Leadership. *Sigma Theta Tau International Honor Society of Nursing*, *28*(2), 18–21.

Bartov, E. (1993). The Timing of Asset Sales and Earnings Manipulation. *The Accounting Review*, *68*(4), 840–855.

Bassler, D., Briel, M., Montori, V., & et al. (2010). Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *Journal of the American Medical Association*, *303*(12), 1180–1187.

Bassler, D., Briel, M., Montori, V., Lane, M., Glasziou, P., Zhou, Q., & Heels-Ansdell, D. (2010). Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*, *303*(12), 1180–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20332404

Baumgartner, F., & Jones, B. (1993). *Agendas and Instability in American Politics* (2nd ed.). Chicago: University of Chicago Press.

Beck, U. (1992). *Risk Society: Towards a New Modernity*. London: Sage.

Becker-Brüser W. (2010). Research in the pharmaceutical industry cannot be objective (article in German). *Z Evid Fortbild Qual Gesundhwes*, *104*(3), 183–189. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20608245

Bekelman, J., Li, Y., & Gross, C. (2003). Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *Journal of the American Medical Association*, *289*(4), 454–465.

Benner, M. (2007). The incumbent discount: stock market categories and response to radical technological change. *Academy of Management Review*, *32*(3), 703–720.

Benner, M. (2010). Securities Analysts and Incumbent Response to Radical Technological Change: Evidence from Digital Photography and Internet Telephony. *Organization Science*, *21*(1), 42–46.

Bero, L., Oostvogel, F., Bacchetti, P., & et al. (2007). Factors Associated with Findings of Published Trials of Drug–Drug Comparisons: Why Some Statins Appear More Efficacious than Others. *PLOS Medicine*, *4*(6), 1001–1010.

Beyer, J., & Trice, H. (1982). The Utilization Process: A Conceptual Framework and Synthesis of Empirical Findings. *Administrative Science Quarterly*, *27*(4), 591–622.

Billé, R. (2010). Action without change? On the use and usefulness of pilot experiments in environmental management. *S.A.P.I.EN.S*, *3*(1), 1–6.

Binder, S., Rhodes, R., & Rockman, B. (2008). *The Oxford Handbook of Political Institutions*. Oxford University Press.

Bland, J., & Altman, D. (2011). Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials*, *2*, 264.

BMJ (1996). Pharmaceutical medicine. Webpage: http://www.bmj.com/content/313/7056/S2-7056. Retrieved on 11 December 2014.

Boa, I., Johnson, P., & King, S. (2010). *The impact of research on the policy process* (No. 82). London. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207544/wp82.pdf

Bocking, S. (2004). *Nature's Experts: Science, Politics, and the Environment*. Rutgers University Press.

Boehmer-Christiansen, S. (1994). Global climate protection policy: The limits of scientific advice. *Global Environmental Change*, *4*(2), 140–159. Retrieved from http://www.sciencedirect.com/science/article/pii/0959378094900493

Bourgeois, F., Murthy, S., & Mandl, K. (2010). Outcome reporting among drug trials registered in ClinicalTrials.gov. *Annals of Internal Medicine*, *153*(3), 158–166.

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. (2010). Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes. *Journal of the American Medical Association*, *303*(20), 2058–2064.

Bovens, M., ’t Hart, P., & Kuipers, S. (2008). The politics of policy evaluation. In M. Moran, B. Goodin, & M. Rein (Eds.), *The Oxford Handbook of Public Policy* (pp. 319–335). New York: Oxford University Press.

Box-Steffensmeier, J., & Jones, B. (2004). *Event History Modeling: A Guide for Social Scientists* (p. 234). Cambridge: Cambridge University Press.

Breslau, D. (1997). The Political Power of Research Methods: Knowledge Regimes in U. S. Labor-Market Policy. *Theory and Society*, *26*(6), 869–902.

Brodkin, E., & Kaufman, A. (2000). Policy Experiments and Poverty Politics. *Social Service Review*, *74*(4), 507–532. Retrieved from http://www.jstor.org/stable/10.1086/516423

Bruce Baker, C., Johnsrud, M., Crismon, M., Rosenheck, R., & Woods, S. (2003). Quantitative analysis of sponsorship bias in economic studies of antidepressants. *The British Journal of Psychiatry*, *183*(6), 498–506. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14645020

Bushee, B. (1998). The Influence of Institutional Investors on Myopic R&D Investment Behaviour. *The Accounting Review*, *73*(3), 305–333.

Cabinet Office. (1997). *Your Right to Know: Freedom of Information. White Paper*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/272048/3818.pdf

Cabinet Office. (1998). *Public Services for the Future: Modernisation, Reform, Accountability*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260759/4181.pdf

Cabinet Office. (1999a). *Modernising Government*. Retrieved from http://www.archive.official-documents.co.uk/document/cm43/4310/4310.htm

Cabinet Office. (1999b). *Professional policy making for the twenty first century* (pp. 1–78). Retrieved from http://www.civilservant.org.uk/profpolicymaking.pdf

Campbell, D. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409–429. Retrieved from http://psycnet.apa.org/psycinfo/1969-17253-001

Caplan, N. (1980). What do we know about knowledge utilisation? In L. Braskamp & R. Brown (Eds.), *Utilisation of evaluative information* (pp. 1–10). San Fransisco: Jossey-Bass.

Card, D., Kluve, J., & Weber, A. (2010). Active Labor Market Policy Evaluations: A Meta-Analysis. *The Economic Journal*, *120*(548), F452–F477. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0297.2010.02387.x/abstract

Carpenter, D. (2001). *The Forging of Bureaucratic Autonomy: Reputations, Networks and Policy Innovation in Executive Agencies 1862 - 1928*. Princeton: Princeton University Press.

Carpenter, D. (2002). Groups, the Media, Agency Waiting Costs, and FDA Drug Approval. *American Journal of Political Science*, *46*(3), 490–505. Retrieved from http://www.jstor.org/stable/3088394

Carpenter, D. (2004). Protection without capture: Product approval by a politically responsive, learning regulator . *American Political Science Review*, *98*(4), 613–631. Retrieved from http://www.jstor.org/stable/4145328

Carpenter, D. (2010). *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press.

Carpenter, D. (2012). Is Health Politics Different? *Annual Review of Political Science*, *15*, 287–311. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-050409-113009

Carpenter, D., & Krause, G. (2012). Reputation and Public Administration. *Public Administration Review*, *72*(1), 26–32. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6210.2011.02506.x/abstract

Cartwright, N., & Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Chan, A., & Altman, D. (2005). Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*, *330*(7494), 753. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15681569

Chan, A., Hróbjartsson, A., Haahr, M., & et al. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Journal of the American Medical Association*, *291*(20), 2457–2465.

Chaney, P. (2013). Electoral discourse analysis of state foreign policy development: exploring the party politicization of the Commonwealth in UK Westminster elections. *Contemporary Politics*, *19*(2), 203–220.

Chitty, C. (2000). *Why pilot?* Paper presented to the Royal Statistical Society Conference, 'The evaluation of economic and social policies'. London: RSS, 4 July.

Cialdini, R. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105–109. Retrieved from http://cdp.sagepub.com/content/12/4/105.abstract

Cialdini, R., Cacioppo, J., Bassett, R., & Miller, J. (1978). Low-ball procedure for producing compliance: Commitment then cost. *Journal of Personality and Social Psychology*, *36*(5), 463–476.

Coate, M. (2002). A Test of Political Control of the Bureaucracy: The Case of Mergers. *Economics & Politics*, *14*(1), 1–18. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1468-0343.00097/abstract

Cohen, N. (2011). *Policy Entrepreneurs and the Design of Public Policy: Conceptual Framework and the Case of the National Health Insurance Law in Israel* (No. 7). Raanana. Retrieved from http://www.openu.ac.il/policy/download/maamar-7.pdf

Coleman, J. (1978). The Uses of Social Science in the Development of Public Policy. *Urban Review 10*, 197-202

Davis, L., Cullen, M., Davis, H., & Lindsay, G. (2010). *Evaluation of Time to Talk Community Programme*. London. Retrieved from http://dera.ioe.ac.uk/812/1/DCSF-RR207.pdf

Dechow, P., & Sloan, R. (1991). Executive incentives and the horizon problem: An empirical investigation. *Journal of Accounting and Economics*, *14*(1), 51–89.

Department for Work and Pensions. (2002). *Pathways to Work: Helping People into Employment*. Green Paper Cm 5690.

Department for Work and Pensions. (2012). *Early impacts of Mandatory Work Activity* . Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/222938/early_impacts_mwa.pdf

Dewatripont, M., & Roland, G. (1996). Transition as a process of large-scale institutional change. *Economics of Transition*, *4*(1), 1–30. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0351.1996.tb00159.x/abstract

Dranove, D., & Meltzer, D. (1994). Do Important Drugs Reach the Market Sooner? *The RAND Journal of Economics*, *25*(3), 402–423. Retrieved from http://www.jstor.org/stable/2555769

Dryzek, J. (1990). *Discursive Democracy: Politics, Policy, and Political Science*. Cambridge: Cambridge University Press.

Dunn, W. (2004). *Public Policy Analysis: An Introduction* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., … Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*, *3*(8), e3081. doi:10.1371/journal.pone.0003081

Dwan, K., Kirkham, J., Williamson, P., & Gamble, C. (2013). Selective reporting of outcomes in randomised controlled trials in systematic reviews of cystic fibrosis. *BMJ Open*. Retrieved August 27, 2013, from http://bmjopen.bmj.com/content/3/6/e002709.full

Easterbrook, P., Berlin, J., Gopalan, R., & Matthews, D. (1991). Publication bias in clinical research. *Lancet*, *337*(8746), 867–872. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1672966

Ehrlich, S. (2011). *Access Points: An Institutional Theory of Policy Bias and Policy Complexity*. New York: Oxford University Press.

Epstein, L. (1980). Whatever Happened to the British Party Model? *American Political Science Review*, *14*, 9–22. Retrieved from http://www.apsanet.org/~pop/APSA1950/Epstein1980.pdf

Epstein, S. (1996). *Impure Science: AIDS, Activism and the Politics of Knowledge*. Berkeley: University of California Press.

Ettelt, S., & Mays, N. (2013). Health policy piloting in England in the 2000s: has the drive towards "evidence based policy" resolved long-standing dilemmas in relation to the purpose of piloting? In *Policy Pilots and Evaluation. Health History and Policy Seminar*. London: LSHTM. Retrieved from http://history.lshtm.ac.uk/files/2014/02/Policy-Pilots-report-final-version-2.pdf

Eyding, D., Lelgemann, M., Grouven, U., Härter, M., Kromp, M., Kaiser, T., … Wieseler, B. (2010). Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*, *341*, c4737. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20940209

Fama, E. (1980). Agency Problems and the Theory of the Firm . *Journal of Political Economy*, *88*(2), 288–307.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE*, *4*(5), 1–11.

Fay, R. (1996). *Enhancing the Effectiveness of Active Labour Market Policies: Evidence from Programme Evaluations in OECD Countries* (pp. 1–64). London.

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, *14*(3), 159–181. Retrieved from www.iew.uzh.ch/wp/iewwp040.pdf

Feldman, M., & March, J. (1981). Information in Organizations as Signal and Symbol. *Administrative Science Quarterly*, *26*(2), 171–186. Retrieved from http://www.jstor.org/stable/2392467

Ferraz, C. (2007). *Electoral Politics and Bureaucratic Discretion: Evidence from Environmental Licenses and Local Elections in Brazil* (pp. 1–43). Rio de Janeiro. Retrieved from http://www.cid.harvard.edu/neudc07/docs/neudc07_s1_p20_ferraz.pdf

Fine, C. (2006). *A Mind of Its Own: How Your Brain Distorts and Deceives*. Cambridge (UK): Icon Books Ltd.

Fletcher, J. (2007). Subgroup analyses: how to avoid being misled. *BMJ*, *335*(7610), 96–97.

Fletcher RH, & Black B. (2007). "Spin" in scientific writing: scientific mischief and legal jeopardy. *Medical Law* , *26*(3), 511–525.

Fowler, A., Agha, R., Camm, C., & Littlejohns, P. (2013). The UK Freedom of Information Act (2000) in healthcare research: a systematic review. *BMJ Open*, *3*(11). Retrieved from http://bmjopen.bmj.com/content/3/11/e002967.full

Frye, T., & Mansfield, E. (2004). Timing is Everything: Elections and Trade Liberalization in the Postcommunist World. *Comparative Political Studies*, *37*(4), 371–398. Retrieved from http://cps.sagepub.com/content/37/4/371.abstract

Furubo, J., Rist, R., & Sandahl, R. (2002). *International Atlas of Evaluation*. New Brunswick, NJ: Transaction Publishers.

Gale, M., & Ball, L. (2002). Does Positivity Bias Explain Patterns of Performance on Wason's 2-4-6 task? In W. Gray & C. Schunn (Eds.),

Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society (p. 340). Routledge.

Garrett-Jones, S., Turpin, T., Burns, P., & et al. (2005). Common purpose and divided loyalties: the risks and rewards of cross-sector collaboration for academic and government researchers. *R&D Management*, *35*(5), 535–544.

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, *13*(4), 755–779

Gilinsky, A., & Judd, B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging*, *9*(3), 356–371. Retrieved from http://psycnet.apa.org/psycinfo/1995-01225-001

Gillespie, R. (1991). *Manufacturing knowledge : a history of the Hawthorne experiments*. Cambridge: Cambridge University Press.

Glaser, B., & Bero, L. (2005). Attitudes of academic and clinical researchers toward financial ties in research: a systematic review. *Science and Engineering Ethics*, *11*(4), 553–573. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16279755

Goldacre, B. (2012). *Bad Pharma: How drug companies mislead doctors and harm patients*. London: Fourth Estate.

Government Office for Science. (2012). *Science and Analysis Assurance Review of the Department for Work and Pensions*. London.

Government Office for Science. (2013). Heads of Analysis group. Retrieved August 01, 2013, from http://www.bis.gov.uk/go-science/about/how-we-work/heads-of-analysis-group

Government Social Research Unit. (2010a). *Publishing Research in Government: GSR Publication Guidance*. Retrieved from http://www.civilservice.gov.uk/wp-content/uploads/2011/09/GSR-Publication-Guidance-29-Jan-2010_tcm6-35775.pdf

Government Social Research Unit. (2010b). *Publishing Research in Government: GSR Publication Guidance* (pp. 1–17). London.

Graham, J., Harvey, C., & Rajgopal, S. (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics*, *40*(1), 3–73.

Greenberg, M., & Goldberg, L. (1994). Ethical Challenges to Risk Scientists: An Exploratory Analysis of Survey Data. *Science*

Technology Human Values, *19*(2), 223–241 . Retrieved from http://sth.sagepub.com/content/19/2/223.short

Greenwood, R., Oliver, C., Suddaby, R., & Sahlin-Andersson, K. (2008). *The SAGE Handbook of Organizational Institutionalism*. SAGE.

Grice, A. (2012). Labour tries to outflank Tories on welfare. *The Independent*, p. 9 March. Retrieved from http://www.independent.co.uk/news/uk/politics/labour-tries-to-outflank-tories-on-welfare-7546203.html

Guyatt, G., Wyer, P., & Ioannidis, J. (2008). When to believe a subgroup analysis. In G. Guyatt, D. Rennie, M. Meade, & D. Cook (Eds.), *User's guide to the medical literature: a manual for evidence-based clinical practice* (2nd ed., pp. 571–583). AMA.

Hahn, S., Williamson, P., & Hutton, J. (2002). Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. *J Eval Clin Pract. , 8*(3), 353–359. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12164983

Hallsworth, M., Parker, S., & Rutter, J. (2011). *Policy making in the real world* (pp. 1–105). London: Institute for Government. Retrieved from http://www.instituteforgovernment.org.uk/sites/default/files/publications/Policy%20making%20in%20the%20real%20world.pdf

Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in Practice* (2nd ed.). London: Routledge.

Handley, S., Capon, A., Beveridge, M., Dennis, I., & Evans, J. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking and Reasoning , 10*(2), 175–195. Retrieved from http://philpapers.org/rec/HANWMI-2

Hasluck, C. (2000). *The New Deal for Young People, Two Years On*. Sheffield. Retrieved from http://www2.warwick.ac.uk/fac/soc/ier/publications/2000/hasluck_2000_esr41rep.pdf

Henig, J. (2008). *Spin Cycle: How Research Is Used in Policy Debates, The Case of Charter Schools*. Russell Sage Foundation/The Century Foundation.

Henig, J. (2009). Politicization of Evidence: Lessons for an Informed Democracy. *Educational Policy, 23*(1), 137–160.

Hergovich, A., Schott, R., & Burger, C. (2010). Biased evaluation of abstracts depending on topic and conclusion: Further evidence of a confirmation bias within scientific psychology. . *Current Psychology*, *29*(3), 188–209. Retrieved from http://link.springer.com/article/10.1007/s12144-010-9087-5

Hewitt, C., Mitchell, N., & Torgerson, D. (2008). Listen to the data when results are not significant. *BMJ*, *336*(23): 23-25.

Higgins, J., Green, S., & et al. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Retrieved from www.cochrane-handbook.org

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983). Understanding Robust and Exploratory Data Analysis. Wiley series in probability and mathematical statistics. New York: John Wiley & Sons.

Hobolt, S. B., & Klemmemsen, R. (2005). Responsive Government? Public Opinion and Government Policy Preferences in Britain and Denmark. *Political Studies*, *53*(2), 379–402. doi:10.1111/j.1467-9248.2005.00534.x

Hollister, R. (2008). Hollister Response to Richard Nathan's Opening Statement. *Journal of Policy Analysis and Management 27*(3): 610–615.

Hollister, R. (2009). Reply Comments. *Journal of Policy Analysis and Management 28*(1): 178–180.

Hood, C. (2011). *The Blame Game: Spin, Bureaucracy and Self-Preservation in Government*. Princeton and Oxford: Princeton University Press.

Hood, C., & Jones, D. (1996). *Accident and Design - Contemporary Debates in Risk Management*. London: UCL Press.

Hood, C., & Lodge, M. (2006). *The Politics of Public Service Bargains. Reward, Competency, Loyalty - and Blame* (p. 236). Oxford: Oxford University Press. Retrieved from http://ukcatalogue.oup.com/product/9780199269679.do#.UeayqG3ZXTo

House of Commons. (1999a). *Fifth Special Report of the Social Security Committee*. Retrieved from http://www.publications.parliament.uk/pa/cm199899/cmselect/cmsocsec/855/855s02.htm

House of Commons. (1999b). *Sixth Special Report of the Social Security Committee, Session 1998-99, .* Retrieved from http://www.publications.parliament.uk/pa/cm200001/cmselect/cmedump/206/20619.htm

House of Commons. (2000). *Employment and training programmes for the unemployed*. Retrieved from http://www.parliament.uk/briefing-papers/RP00-81/employment-and-training-programmes-for-the-unemployed

House of Commons. (2003). *Employment and Training Programmes for the Unemployed*. Retrieved from http://www.parliament.uk/briefing-papers/RP03-13/employment-and-training-programmes-for-the-unemployed

House of Commons. (2005a). *Employment and Training Programmes for the Unemployed Volume II: Other programmes and pilots*. Retrieved from http://www.parliament.uk/briefing-papers/RP05-62/employment-and-training-programmes-for-the-unemployed-volume-ii-other-programmes-and-pilots

House of Commons. (2005b). *Employment and Training Programmes for the Unemployed. Volume I: recent developments and the New Deal programmes*. Retrieved from http://www.parliament.uk/briefing-papers/RP05-61/employment-and-training-programmes-for-the-unemployed-volume-i-recent-developments-and-the-new-deal-programmes

House of Commons. (2007). *Scientific Advice, Risk and Evidence Based Policy Making: Government Response to the Committee's Seventh Report of Session 2005–06* (HC 307.). Science and Technology Committee. Retrieved from http://www.publications.parliament.uk/pa/cm200607/cmselect/cmsctech/307/307.pdf

House of Commons. (2012). *Debates.* (p. 6 Nov : Column 558W). Retrieved from http://www.publications.parliament.uk/pa/cm201213/cmhansrd/cm121106/text/121106w0003.htm#121106117000115

Hull, D. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press.

Huss, A., Egger, M., Hug, K., Huwiler-Müntener, K., & Röösli, M. (2007). Source of funding and results of studies of health effects of mobile phone use: systematic review of experimental studies. *Environ Health*

Perspect., *115*(1), 1–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17366811

Innvaer, S., Vist, G., & Trommald, M. (2002). Health policy-makers' perceptions of their use of evidence: a systematic review. *Journal of Health Services Research & Policy*, *7*(4), 239–244.

Institute of Medicine. (2009). *Conflict of Interest in Medical Research, Education, and Practice*. (B. Lo & M. Field, Eds.). Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12598

Ioannidis, J. P. A., & Karassa, F. B. (2010). The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. *BMJ (Clinical Research Ed.)*, *341*(sep13_1), c4875. doi:10.1136/bmj.c4875

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-Telling. *Psychological Science*, *23*(5), 524–532. Retrieved from https://www.cmu.edu/dietrich/sds/docs/loewenstein/MeasPrevalQuestTruthTelling.pdf

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *J Psychiatry Neurosci.*, *37*(3), 149–152. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3341407/

Jowell, R. (2003). *Trying It Out. The Role of "Pilots" in Policy-Making.* London. Retrieved from http://www.civilservice.gov.uk/wp-content/uploads/2011/09/Trying-it-Out_tcm6-36824.pdf

Jureidini, J., McHenry, L., & Mansfield, P. (2008). Clinical trials and drug promotion: Selective reporting of study 329. *The International Journal of Risk and Safety in Medicine*, *20*(1-2), 73–81.

Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *BMJ (Clinical Research Ed.)*, *326*(7404), 1453–5. doi:10.1136/bmj.326.7404.1453

Keat, R., Whiteley, N., & Abercrombie, N. (1994). *The Authority of the Consumer*. London: Routledge.

Keitner, G., Posternak, M., & Ryan, C. (2003). How many subjects with major depressive disorder meet eligibility requirements of an antidepressant efficacy trial? *The Journal of Clinical Psychiatry*, *64*(9), 1091–1093.

Kelley, J. (2007). Who Keeps International Commitments and Why? The International Criminal Court and Bilateral Nonsurrender Agreements. *American Political Science Review*, *101*(03), 573. doi:10.1017/S0003055407070426

Kelly, R., Cohen, L., Semple, R., & et al. (2006). Relationship between drug company funding and outcomes of clinical psychiatric research . *Psychological Medicine*, *36*(11), 1647–1656.

Kingdon, J. (1984). *Agendas, Alternatives and Public Policies* (p. 240). Boston: Little Brown.

Klingemann, H., Hofferbert, R., & Budge, I. (1994). *Parties, Policies and Democracy*. Westview Press.

Knorr, K. (1977). Policy-makers' use of social science knowledge: symbolic or instrumental? In C. Weiss (Ed.), *Using social research in public policy making* (pp. 165–182). Lexington: Lexington Books.

Knott, J., & Wildavsky, A. (1980). If dissemination is the solution, what is the problem? . *Knowledge: Creation, Diffusion, Utilization*, *1*(4), 537–578.

Koehler, J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28–55. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1469652.

Kokis, J., Macpherson, R., Toplak, M., West, R., & Stanovich, K. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*, 26–52. Retrieved from http://www.keithstanovich.com/Site/Research_on_Reasoning_files/RNjecp02.pdf

Krause, G., & Douglas, J. (2005). Institutional Design versus Reputational Explanations for Agency Performance: Evidence from U.S. Government Macroeconomic and Fiscal Projections. *Journal of Public Administration Research and Theory*, *15*(2), 281–306. Retrieved from http://jpart.oxfordjournals.org/content/15/2/281.short

Kunda, Z. (1999). *Social Cognition: Making Sense of People*. MIT Press.

LaMattina, J. (2012). How Committed is Big Pharma to Rare Diseases? *Forbes* , p. 5 February. Retrieved from http://www.forbes.com/sites/johnlamattina/2012/05/02/how-committed-is-big-pharma-to-rare-diseases/

Lavertu, S., & Weimer, D. (2011). Federal Advisory Committees, Policy Expertise, and the Approval of Drugs and Medical Devices at the FDA. *Journal of Public Administration Research and Theory*, *21*(2), 211–237. Retrieved from http://jpart.oxfordjournals.org/content/21/2/211.short

Lavis, J., Robertson, D., Woodside, J., & et al. (2003). How can research organizations more effectively transfer research knowledge to decision makers? *The Milbank Quarterly*, *81*(2), 171–172.

Lee, K., Bacchetti, P., & Sim, I. (2008). Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS Medicine*, *5*(9), e191. doi:10.1371/journal.pmed.0050191

Lee, K., McNeer, J., Starmer, C., Harris, P., & Rosati, R. (1980). Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation*, *61*(3), 508–515. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7353241

Lexchin, J. (2012). Sponsorship bias in clinical research. *Int J Risk Saf Med.*, *24*(4), 233–242. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23135338

Lexchin, J., Bero, L., Djulbegovic, B., & et al. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *British Medical Journal*, *326*(7400), 1167–1170. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12775614

Lindblom, C. (1959). The Science of "Muddling Through." *Public Administration Review*, *19*(2), 79–88. Retrieved from http://faculty.washington.edu/mccurdy/SciencePolicy/Lindblom Muddling Through.pdf

Lipsky, M. (1980). *Street-level Bureaucracy: The Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. Retrieved from http://psycnet.apa.org/psycinfo/1981-05421-001

Lowi, T. (1972). Four systems of policy, politics, and choice. *Public Administration Review*, *32*(July/August), 298–310.

Lowndes, V., & Roberts, M. (2013). *Why Institutions Matter: The New Institutionalism in Political Science*. Palgrave Macmillan.

Luke, A. (2011). Generalizing across borders policy and the limits of educational science. *Educational Researcher*, *40*(8), 367–377.

Lynn, J., & Jay, A. (1984). *The Complete "Yes Minister": The Diaries of a Cabinet Minister*. . London: BBC Books.

MacDonald, J. (2010). Limitation Riders and Congressional Influence over Bureaucratic Policy Decisions. *American Political Science Review*, *104*(04), 766–782. Retrieved from http://dx.doi.org/10.1017/S0003055410000432

MacDonald, J., & Franko, W. (2007). Bureaucratic Capacity and Bureaucratic Discretion: Does Congress Tie Policy Authority to Performance? *American Politics Research* , *35*(6), 790–807. Retrieved from http://apr.sagepub.com/content/35/6/790.abstract

Maguire, M. (2004). The Crime Reduction Programme: Reflections on the Vision and the Reality. *Criminal Justice*, *4*(3), 213–238. Retrieved from http://crj.sagepub.com/content/4/3/213.short?rss=1&ssource=mfc

Mahoney, M. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175. Retrieved from http://link.springer.com/article/10.1007/BF01173636

Malenka, D., Baron, J., Johansen, S., Wahrenberger, J., & Ross, J. (1993). The framing effect of relative and absolute risk. *J Gen Intern Med.*, *8*(10), 543–548. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8271086

Marco, C., & Larkin, G. (2000). Research ethics: ethical issues of data reporting and the quest for authenticity. *Academic Emergency Medicine*, *7*(6), 691–694.

Martin, S., Butzin, C., Saum, C., & Inciardi, J. (1999). Three-Year Outcomes of Therapeutic Community Treatment for Drug-Involved Offenders in Delaware: From Prison to Work Release to Aftercare . *The Prison Journal* , *79*(3), 294–320. Retrieved from http://tpj.sagepub.com/content/79/3/294.short

Martin, S., & Sanderson, I. (1999). Evaluating Public Policy Experiments Measuring Outcomes, Monitoring Processes or Managing Pilots? *Evaluation*, *5*(3), 245–258. Retrieved from http://evi.sagepub.com/content/5/3/245.refs

Melander, H., Ahlqvist-Rastad, J., Meijer, G., & et al. (2003). Evidence b(i)ased medicine--selective reporting from studies sponsored by

pharmaceutical industry: review of studies in new drug applications. *British Medical Journal*, *326*(7400), 1171–1173.

Merton, R. (1942). A note on science and democracy. *Journal of Legal and Political Sociology*, *1942*(1), 115–126.

Metcalf, C. (2008). Threats to independence and objectivity of government-supported evaluation and policy research. *Journal of Policy Analysis and Management*, *27*(4), 927–934.

Mintrom, M. (1997). Policy Entrepreneurs and the Diffusion of Innovation. *American Journal of Political Science*, *41*(3), 738–770. Retrieved from http://www.jstor.org/stable/2111674

Mitroff, I. (1974). Norms and Counter-Norms in a Select Group of the Apollo Moon Scientists: A Case Study of the Ambivalence of Scientists. *American Sociological Review*, *39*(4), 579–595.

Monaghan, M. (2010). The Complexity of Evidence: Reflections on Research Utilisation in a Heavily Politicised Policy Area. *Social Policy and Society*, *9*(1), 1–12.

Montori, V., Devereaux, P., Adhikari, N., & et al. (2005). Randomized trials stopped early for benefit: a systematic review. *Journal of the American Medical Association*, *294*(17), 2203–2209.

Montori, V., Jaeschke, R., Schünemann, H., & et al. (2004). User's guide to detecting misleading claims in clinical research reports. *British Medical Journal*, *329*(7474), 1093–1096.

Morçöl, G. (2001). Positivist Beliefs among Policy Professionals: An Empirical Investigation. *Policy Sciences*, *34*(3), 381–401.

Mulkay, M. (1976). Norms and ideology in science. *Social Science Information*, *15*, 637–656.

Murphy, K., & Fafard, P. (2012). Taking power, politics, and policy problems seriously: the limits of knowledge translation for urban health research. *Journal of Urban Health*, *89*(4), 723–732.

Nagel, S. (2002). *Handbook of Policy Evaluation*. Thousand Oaks, CA: Sage.

Nakazono, Y., & Ueda, K. (2013). Policy commitment and market expectations: Lessons learned from survey based evidence under Japan's quantitative easing policy. *Japan and the World Economy*, *25-26*, 102–113. doi:10.1016/j.japwor.2013.03.004

Nathan, R. (2008a). The role of random assignment in social policy research. *Journal of Policy Analysis and Management*, *27*(2), 401–415.

Nathan, R. (2008b). Nathan Response to Robinson Hollister's Opening Statement. *Journal of Policy Analysis and Management 27*(3): 607–10.

Nathan, R. (2009). Reply Comments. *Journal of Policy Analysis and Management 28*(1): 180–81.

National Audit Office. (2010). *Support to incapacity benefits claimants through Pathways to Work*. London. Retrieved from http://www.nao.org.uk/report/support-to-incapacity-benefits-claimants-through-pathways-to-work/

National Audit Office. (2013). *Evaluation in Government*. Retrieved from http://www.nao.org.uk/report/evaluation-government/

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nunn, A., & Jassi, S. (2010). *Jobcentre Plus Jobseeker's Allowance off-flow rates: Key Management Indicator Post-Implementation Review*. London. Retrieved from http://webarchive.nationalarchives.gov.uk/20130314010347/http://research.dwp.gov.uk/asd/asd5/rports2009-2010/rrep661.pdf

Oakeshott, M. (1996). *The Politics of Faith and the Politics of Scepticism*. (T. Fuller, Ed.). New Haven: Yale University Press.

Olson, M. (1997). Firm Characteristics and the Speed of FDA Approval. *Journal of Economics and Management Strategy*, *6*(1), 377–401.

Ostrom, E. (1986). An Agenda for the Study of Institutions. *Public Choice*, *48*, 3–25.

Oswald, M., & Grosjean, S. (2004). Confirmation Bias. In R. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory* (pp. 79–96). Hove, UK: Psychology Press.

Oxman, A., & Guyatt, G. (1992). A consumer's guide to subgroup analyses. *Annals of Internal Medicine*, *116*(1), 78–84.

Page, E., & Jenkins, B. (2005). *Policy Bureaucracy. Government with a Cast of Thousands* (p. 244). Oxford: Oxford University Press.

Paun, A., & Harris, J. (2013). *Accountability at the Top: Supporting Effective Leadership in Whitehall*. London. Retrieved from

http://www.instituteforgovernment.org.uk/sites/default/files/publications/Accountability at the top - final.pdf

Pelz, D. (1978). Some Expanded Perspectives on Use of Social Science in Public Policy. In J. Yinger & S. Cutler (Eds.), *Major Social Issues: A Multidisciplinary View* (pp. 346–357). New York: The Free Press.

Penman, S., & Zjang, X. (2002). Accounting Conservatism, the Quality of Earnings, and Stock Returns. *The Accounting Review*, *77*(2), 237–264.

Peters, G. (2011). *Institutional Theory in Political Science: The New Institutionalism* (3rd ed.). Continuum Publishing Corporation.

Pharr, S., & Putnam, R. (2000). *Disaffected Democracies: What's Troubling the Trilateral Countries?* (p. 360). Princeton NJ: Princeton University Press.

Pinch, T. (1986). *Confronting Nature: The Sociology of Solar-Neutrino Detection*. (G. Bohme, Ed.). Dordrecht: Dordrecht Reidel.

Plous, S. (1993). *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill.

Pomper, G., & Lederman, S. (1980). *Elections in America: Control and Influence in Democratic Politics.* . London/New York: Longman.

Pratt, C., & Moyé, L. (1995). The cardiac arrhythmia suppression trial: Casting suppression in a different light. *Circulation*, *91*(1), 245–247.

Rae, N., & Gil, J. (2010). *Party Polarization and Ideology: Diverging Trends in Britain and the United States* . Retrieved from http://www.britishpoliticsgroup.org/documents/rae-gilbpgpaperfinal.pdf

Rallings, C. (1987). The Influence of Election Programs: Britain and Canada 1945-1979. In I. Budge, D. Robertson, & D. Hearl (Eds.), *Ideology, Strategy and Party Change* (pp. 1–14). Cambridge: Cambridge University Press.

Rao, J., & Sant Cassia, L. (2012). Ethics of undisclosed payments to doctors recruiting patients in clinical trials. *BMJ*, *325*(36). Retrieved from http://www.bmj.com/content/325/7354/36.1

Redelmeier, D. A., & Tversky, A. (1996). On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences*, *93*(7), 2895–2896. doi:10.1073/pnas.93.7.2895

Regan, D. (1971). Effects of a favor and liking on compliance. *Journal of Experimental Social Psychology*, *7*(6), 627–639. Retrieved from http://www.sciencedirect.com/science/article/pii/0022103171900254

Repenning, N., & Henderson, R. (2010). *Making the Numbers? "Short Termism" & The Puzzle of Only Occasional Disaster* (No. 11-033) (pp. 1–36). Retrieved from http://www.hbs.edu/faculty/Publication Files/11-033.pdf

Rhodes, R. (2013). *Political Anthropology and Civil Service Reform: Prospects and Limits*. Retrieved from http://www.cbs.dk/files/cbs.dk/cbs_lecture_rod_rhodes.pdf

Riddell, P., Gruhn, Z., & Carolan, L. (2011). *The Challenge of Being a Minister: Defining and developing ministerial effectiveness* . London. Retrieved from http://www.instituteforgovernment.org.uk/sites/default/files/publications/The Challenge of Being a Minister.pdf

Riley, R., Bewley, H., Kirby, S., Rincon-Aznar, A., & George, A. (2011). *The introduction of Jobcentre Plus: An evaluation of labour market impacts*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/214567/rrep781.pdf

Rising, K., Bacchetti, P., & Bero, L. (2008). Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. *PLoS Medicine*, *5*(11), e217; discussion e217. doi:10.1371/journal.pmed.0050217

Ritter, A., & Lancaster, K. (2013). Measuring research influence on drug policy: A case example of two epidemiological monitoring systems. *International Journal of Drug Policy*, *24*(1), 30–37.

Rockoff, J. (2013). Drug Makers See Profit Potential in Rare Diseases. *Wall Street Journal*, p. 30 January.

Rogers-Dillon, R. (2004). *The Welfare Experiments: Politics and Policy Evaluation* . Stanford: Stanford University Press.

Rose, R. (1980). *Do Parties Make a Difference?* London: Macmillan.

Rossi, P., Wright, J., & Anderson, A. (1983). *Handbook of Survey Research*. Orlando and London: Academic Press.

Rothstein, B., Samanni, M., & Teorell, J. (2012). Explaining the welfare state: power resources vs. the Quality of Government. *European Political Science Review*, *4*(1), 1–28.

Rothwell, P. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet*, *365*(9453), 82–93.

Royed, T. (1996). Testing the mandate model in Britain and the United States: Evidence from the Reagan and Thatcher eras. *British Journal of Political Science*, *26*(1), 45–80.

Rutter, J. (2012). *Evidence and Evaluation in Policy-Making: A Problem of Supply or Demand?* London. Retrieved from http://www.whatisscience.info/files/evidence_and_evaluation_in_template_final_0.pdf

Sackett, D. (1979). Bias in analytic research. *J Chronic Dis.*, *32*(1-2), 51–63. Retrieved from http://www.epidemiology.ch/history/PDF bg/Sackett DL 1979 bias in analytic research.pdf

Salisbury, C., Stewart, J., Purdy, S., Thorp, H., Cameron, A., Lart, R., … Calnan, M. (2011). Making the Most of Evaluation: A Mixed Methods Study in the English NHS. *Journal of Health Services Research and Policy*, *16*(4), 218–225. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21878444

Sanderson, I. (2002). Evaluation, Policy Learning and Evidence-Based Policy Making. *Public Administration*, *80*(1), 1–22.

Sapp, J. (1990). *Where the Truth Lies: Franz Moewus and the Origins of Molecular Biology*. New York: Cambridge University Press.

Schattschneider, E. (1960). *The Semisovereign People: A Realist's View of Democracy in America*. Holt, Rinehart and Winston.

Schulz, K., & Grimes, D. (2005). Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet*, *365*(9471), 1657–1661.

Science and Analysis Review of the Department for Children, Schools & Families (now Department for Education). (2010) (pp. 1–68). Retrieved from http://www.bis.gov.uk/go-science/science-in-government/reviewing-science-and-engineering/completed-reviews/department-for-education

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaughnessy, A. (2003). What happened to the valid POEMs? A survey of review articles on the treatment of type 2 diabetes. *British Medical Journal*, *327*(7409), 266.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1998). *Preventing Crime: What Works, What Doesn't, What's Promising*. Retrieved from https://www.ncjrs.gov/pdffiles/171676.PDF

Sismondo, S. (2008). How pharmaceutical industry funding affects trial outcomes: Causal structures and responses. *Social Science and Medicine, 66*, 1909–1914.

Smyth, R., Kirkham, J., Jacoby, A., & et al. (2011). Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *British Medical Journal*, *342*(c7153), 1–12.

Song, F., Parekh, S., Hooper, L., & et al. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment*, *14*(8), 1–193.

Stanovich, K., & West, R. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247. Retrieved from http://psycnet.apa.org/psycinfo/2007-11908-001

Steneck, N. (2003). The role of professional societies in promoting integrity in research . *American Journal of Health Behaviour* , *27*, 239–247.

Sterling, T. (1959). Publication decision and their possible effects on inferences drawn from tests of significance - or vice versa . *Journal of the American Statistical Association*, *54*(285), 30–34.

Stiegler, S. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, *6*(2), 103–114.

Sun, X., Briel, M., Walter, S. D., & Guyatt, G. H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ (Clinical Research Ed.)*, *340*(mar30_3), c117. doi:10.1136/bmj.c117

Sutton, J. (1984). Organizational Autonomy and Professional Norms in Science: A Case Study of the Lawrence Livermore Laboratory . *Social Studies of Science*, *14*(2), 197–224.

Svallfors, S. (2012). *Welfare attitudes in Europe: Topline Results from Round 4 of the European Social Survey* . Retrieved from http://www.europeansocialsurvey.org/docs/findings/ESS4_toplines_issue_2_welfare_attitudes_in_europe.pdf

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*(3), 755–769. doi:10.1111/j.1540-5907.2006.00214.x

Tannock, I. (1996). False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *J Natl Cancer Inst.*, *88*(3-4), 206–207. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8632495

Tavare, A. (2012). Scientific misconduct is worryingly prevalent in the UK, shows BMJ survey. *British Medical Journal*. Retrieved from http://www.bmj.com/content/344/bmj.e377

Tetlock, P. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton (NJ): Princeton University Press.

The future of the Civil Service: Making the most of scientists and engineers in government. (2013). London. Retrieved from http://www.bis.gov.uk/assets/goscience/docs/r/bis-13-594-review-science-engineering-in-civil-service.pdf

The LSE GV314 Group. (2014). Evaluation under contract. Government pressure and the production of policy research. *Public Administration*, *92*(1), 224–239. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/padm.12055/abstract

The Magenta Book: Guidance for evaluation. (2011) (pp. 1–141). London.

Tonry, M. (2004). *Punishment and Politics: Evidence and Emulation in the Making of English Crime Control Policy*. London: Willan .

Travers, J., Marsh, S., Williams, M., & et al. (2007). External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, *62*(3), 219–223.

Trotta, F., Apolone, G., Garattini S, & Tafuri, G. (2008). Stopping a trial early in oncology: for patients or for industry? *Annals of Oncology*, *19*(7), 1347–1353.

Trotta, F., Apolone, G., Garattini, S., & Tafuri, G. (2008). Stopping a trial early in oncology: for patients or for industry? *Annals of Oncology*, *19*(7), 1347–1353. Retrieved from http://annonc.oxfordjournals.org/content/19/7/1347

Turner, C., & Spilich, G. (1997). Research into smoking or nicotine and human cognitive performance: does the source of funding make a difference? *Addiction*, *92*(11), 1423–1426. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1360-0443.1997.tb02863.x/abstract

Turner, L., Shamseer, L., Altman, D., Schulz, K., & Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting

of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews , 1*(60). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23194585

Van Staa, T.-P., Leufkens, H., Zhang B, & et al. (2009). A Comparison of Cost Effectiveness Using Data from Randomized Trials or Actual Clinical Practice: Selective Cox-2 Inhibitors as an Example. *PLOS Medicine*, *6*(12), 1–9.

Vaughan, R., & Buss, T. (1998). *Communicating Social Science Research to Policy-Makers* . Thousands Oaks, CA: Sage.

Vedula, S., Bero, L., & Scherer, R. (2009). Outcome Reporting in Industry-Sponsored Trials of Gabapentin for Off-Label Use. *The New England Journal of Medicine*, *361*(20), 1963–1971.

Vogel, D. (1990). When Consumers Oppose Consumer Protection: The Politics of Regulatory Backlash. *Journal of Public Policy*, *10*(4), 449–470. Retrieved from http://www.jstor.org/stable/4007452

Walker, R. (2001). Great Expectations: Can Social Science Evaluate New Labour' . *Evaluation*, *7*(3), 305 – 330.

Wang, R., Lagakos, S., Ware, J., Hunter, D., & Drazen, J. (2007). Statistics in medicine--reporting of subgroup analyses in clinical trials. *N Engl J Med.*, *357*(21), 2189–2194. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18032770

Wang, S., Ou, Y., Cheng, C., & Dahm, P. (2010). Evidence-based urology in practice: when to believe a subgroup analysis?. . *BJU International*, *105*, 162–164. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1464-410X.2009.09053.x/abstract

Weaver, R. (1986). The Politics of Blame Avoidance. *Journal of Public Policy*, *6*(4), 371–398.

Weber, M. (1946). Science as a Vocation. In H. Gerth & W. Mills (Eds.), *Essays in Sociology* (pp. 129–156). New York: Oxford University Press.

Weiss, C. (1979). The Many Meanings of Research Utilization. *Public Administration Review*, *39*(5), 426–431.

Weiss, C. (1980). Knowledge creep and Decision Accretion. *Science Communication*, *1*(3), 381–404.

Weiss, C. (1998). Have We Learned Anything New About the Use of Evaluation? *American Journal of Evaluation*, *19*(1), 21–33.

Westen, D., Blagov, P., Harenski, Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: an FMRI study of emotional constraints on partisan political judgment in the 2004 U.S. Presidential election. *J Cogn Neurosci.*, *18*(11), 1947–1958. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17069484

White, A., & Dunleavy, P. (2010). *Making and breaking Whitehall departments: a guide to machinery of government changes*. London: Institute for Government. Retrieved from http://www.instituteforgovernment.org.uk/sites/default/files/publications/making_and_breaking_whitehall_departments.pdf

Wilson, J. (1989). *Bureaucracy: What Government Agencies Do and Why They Do it*. Basic Books.

Winner, L. (1997). The handwriting on the wall: Resisting technoglobalism's assault on education. In M. Moll (Ed.), *Tech high globalization and the future of Canadian education: A collection of critical perspectives on social, cultural and political dilemmas* (pp. 167–188). Halifax NS: Fernwood.

Wood, B. (1988). Principals, Bureaucrats, and Responsiveness in Clean Air Enforcements. *The American Political Science Review*, *82*(1), 213–234.

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Medicine*, *9*(9), e1001308. doi:10.1371/journal.pmed.1001308

Yusuf, S., Wittes, J., Probstfield, J., & et al. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*, *266*, 93–98.

Zimmerman, M., Chelminski, I., & Posternak, M. (2004). An illustration of how a self-report diagnostic screening scale could improve the internal validity of antidepressant efficacy trials. *Journal of Affective Disorders*, *80*(1), 79–85.

Zuckerman, H. (1988). The sociology of science . In N. Smelser (Ed.), *Handbook of sociology* (pp. 511–574). Newbury Park: SAGE Publications.