Causal Reasoning in Economics

A Selective Exploration of Semantic, Epistemic and Dynamical Aspects

Causaal redeneren in de economische wetenschap

Een selectieve studie naar semantische, epistemische en dynamische aspecten

Thesis

**to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus**

Prof.dr. H.G. Schmidt

**and in accordance with the decision of the Doctorate Board
The public defence shall be held on**

Thursday the 13th of December 2012 at 15.30 hours

by

François Simon Claveau

born in Dolbeau, Canada

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral Committee

Promotors:

Prof.dr. J.J. Vromen
Prof.dr. K.D. Hoover

Other members:

Prof.dr. H.C.M. de Swart
Prof.dr. J. de Koning
Prof.dr. J. Williamson

Copromotor:

Dr. J. Reiss

Contents

List of Figures

List of Tables

Acknowledgments

I came to the Erasmus Institute for Philosophy and Economics (EIPE) four years ago because I was looking for people with whom to have intelligent discussions over topics at the intersection of philosophy and economics. My desire has been amply fulfilled. I have an immense intellectual debt to the EIPE community; to the faculty members, to my fellow students, and to all the ones who visited us during my four years in Rotterdam. I write these lines away from Rotterdam, back at my origin in Montreal. I am fully aware that a main challenge for me in the months and year to come will be to find a community here able to supply me with a dose of intellectual stimulation which rivals the one I was steadily receiving in Holland.

Among all the people with whom I have been interacting in the last four years, Julian Reiss is without doubt the one to whom I owe the most. Julian introduced me to serious thinking about causality and he enormously influenced my approach to philosophy of economics. He has been a great supervisor, not only by abundantly commenting on my work but also by being a source of inspiration for me. I realize now that many of my ideas have been (unwittingly, I assure you) stolen from him. I also much benefited from all the seminars that Julian ran in the last years (on causality, current affairs, and the foundations of statistics) and from his active engagement in our EIPE Reading Group. Thanks Julian.

My second greatest debt go to my other supervisor, Kevin Hoover, who kindly accepted to direct me at a distance. In the last two years, I have mostly lived under the fear of receiving one of his lengthy reports on my work. This fear has been extremely productive. I knew that Kevin would not accept mediocrity, and the expectation of his critical eye made me strive for excellence (though I certainly fall quite short of the goal). It is only a pity that I could not have more face-to-face discussions with him; there is so much more I could learn from Kevin. I sincerely hope (and implicitly ask him now) that our relationship survives my PhD.

Among the other established academics with whom I have had the

chance to interact, special thanks go to the director of EIPE, Jack Vromen. Jack has been the supervisor of my Research Master thesis, and he generously accepted, in spite of tragic personal circumstances and due to the caprice of academic bureaucracy, the role of official supervisor of my PhD research. I have had many enriching discussions with Jack throughout the years and I am grateful to him for having given me such an enjoyable position at EIPE.

I attended a big bunch of classes during the last four years, all of them by interest, and only a few to also get a grade. I thank all the ones responsible for these classes for accepting me even though I was rarely registered, and especially for their patience regarding my talkativeness. The ones not already mentioned are Roger Backhouse, John Davis, John Groenewegen, Alain Guay, Conrad Heilmann, Geoffrey Hodgson, Arjo Klamer, Rogier de Langhe, Fred Muller, Jan van Ours, Matteo Picchio, Michael Sampson and Harrie de Swart.

EIPE would be nothing without its graduate students. I much enjoyed and benefited from the countless discussions I have had with my fellow travelers. I have been especially lucky to sit next to two other PhD candidates, Luis Mireles-Flores and Attilia Ruzzene, working on topics highly related to mine. I also want to thank Sine Bagatur for being such a great office mate; Tyler DesRoches, Tom Wells and Luis for making me join the editorial team of the *Erasmus Journal for Philosophy and Economics*; and David Bassett, Marion Collewet (though not an EIPE student), Deren Olgun, Johanna Thoma, Melissa Vergara Fernández and Philippe Verreault-Julien for especially rewarding discussions.

I am also grateful to all the scholars who commented on my work throughout the years at conferences, due to journal submissions or out of pure generosity. The scholars not already mentioned include: Nancy Cartwright, Stephan Hartmann, Harold Kincaid, Conor Mayo-Wilson, Phyllis McKay Illari, Luca Moretti, Peter Rodenburg, Federica Russo, anonymous referees for and the editors of *Erasmus Journal for Philosophy and Economics*, *Journal of Economic Methodology*, *Philosophy of Science*, *Philosophy of the Social Sciences*, *Studies in History and Philosophy of Biological and Biomedical Sciences*, and many more scholars attending the numerous conferences to which I participated. Thanks also to James Kelleher for proofreading chapter 4, to Roebin Lijnis Huffenreuter for translating in Dutch my “samenvatting”, and to Myriam Van Neste for the beautiful cover art.

In the last four years, I have also been living the Dutch life. For making this experience enjoyable, I thank Ticia Herold who prepared the ground for my family and advised me throughout my stay, Stella Maaswinkel-Thoeng who taught me Dutch, the whole team at my kids' school and daycare which made it so easy for them to become real Dutch children, and the members of my rowing team (Alfons, Douwe, Fokke, Karel, Wiebe) with whom I could pretend (and feel like) being Dutch while sporting. Special thanks go also to Raymonde 'Croquette' Leblanc and Louise 'Melon Miel' Lefebvre for crossing the ocean to care after my kids while I was working hard to wrap up my thesis.

My graduate studies would not have been possible without the financial support of numerous bodies. On the Canadian side, I must thank the *Fonds de recherche du Québec – Société et culture* and the *Social Sciences and Humanities Research Council*. On the Dutch side, I am grateful to *Nuffic* and the Faculty of philosophy, Erasmus Universiteit Rotterdam.

I have always feared becoming what Paul Feyerabend (1999, pp. 112-13) called an expert: "a man, or a woman, who has decided to achieve excellence, supreme excellence in a narrow field at the expense of a balanced development." If I have not become such an 'expert' yet, all the credits go to my two children, Arthur and Ève, and to my partner, Sophie. I will be eternally grateful to them for moving enthusiastically to Rotterdam with me such that I could pursue my graduate studies, and for keeping my mind and body busy with other things than the philosophy of economics. Above all, I thank them for their affection and support.

Introduction

There is a long tradition of great thinkers for whom the decision to study the economy sprang from moral considerations. They believed that by inquiring into the causes of persistent poverty—and its opposite, prosperity—they could improve the lot of mankind. It is undeniable that Adam Smith had such a goal in mind for his *Inquiry into the Nature and Causes of the Wealth of Nations* (1776). Similarly, Alfred Marshall believed it to be of the greatest moral importance to devote himself to "the study of the causes of the degradation of a large part of mankind" (1920, § I.I.5). It seems that the same could be said of most (perhaps all) great economists up to this day: it was to give a better future to mankind that they decided to study the causes of economic distress. They believed that a knowledge of these causes would be a significant step toward averting their tragic effect.

The study of economics has proven hard and frustrating. Despite the efforts of generations of great thinkers, it is a matter of debate whether we currently know much more about the causes of economic distress. To be sure, there has been an undeniable improvement in the welfare of populations in industrialized countries. But this improvement might have little to do with the intellectual efforts of economists. The fact that armies of economists have failed, by and large, to export prosperity to other countries may lead us to doubt that prosperity—where it occurred—is to be imputed to our improved causal understanding.

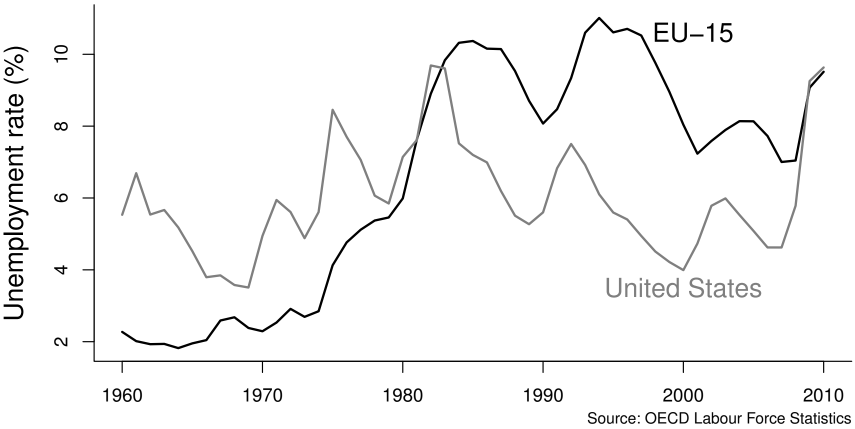
Economic distress is still present in industrialized countries. One form it takes is high unemployment. The intellectual heirs of Smith and Marshall are thus actively inquiring into the causes of unemployment. The evolution of ideas on unemployment can illustrate how hard the study of the economy is.

Back in the 1960s, economists in the United States looked enviously at Europe:

> European countries set their full employment goal at a lower unemployment rate than does the United States, and they have in recent years been more successful in achieving the goal. (Gordon, 1965, p. 42)

Figure 1 shows indeed that U.S. unemployment was above 5% in the early 1960s while it was around 2% in countries of Western Europe.[1] Back then,

[1] 'EU-15' stands for the group of countries that were part of the European Union between 1995 and 2004—i.e. Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom.



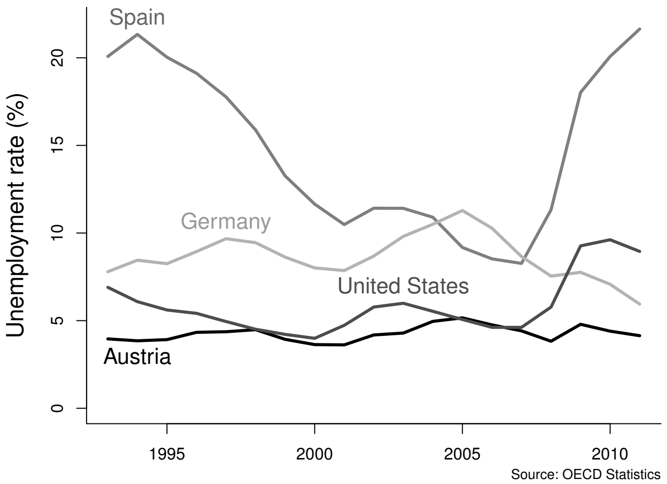
Source: OECD Labour Force Statistics

Figure 1: *Unemployment rates in the United States and in the European Union*

economists credited diverse aspects of European interventionism for this success (see Myers, 1964).

History kept moving and, by the early 1980s, there was nothing more to envy about European unemployment. It had departed from its low points of the 1960s to peak above 10% in the mid-1980s. By the end of the decade, it was European economists who looked at the United States with envy, and everyone started pointing the finger at European interventionism. While interventionism was seen as the solution to (relatively) high U.S. unemployment in the 1960s—U.S. economists then emphasizing that "there will have to be more extensive economic planning in the future" (Ross, 1964, p. 209) and that "the mental barriers stemming from our political philosophy and from our economic priorities are severe obstacles" (Lester, 1964, p. 198)—interventionism became the purported problem later on. And still today, the idea that pervasive labor-market regulations cause high unemployment in Europe is endorsed by most economists.

But history keeps going and ideas might be forced to adapt to it. The first historical fact which might force a reconsideration of widespread causal beliefs among economists has been around for a while: European countries exhibit extreme variations in their unemployment performances. This fact is evident from figure 2. At one extreme, we find countries like Spain with unemployment rates sometimes above 20%, and showing huge fluctuations through time. At the other extreme, countries like Austria have managed to consistently preserve their unemployment rates close or below the one of the United States. The troubling thing with this sharp



Source: OECD Statistics

Figure 2: *Recent unemployment trajectories of selected countries*

contrast is that it is far from clear that countries like Spain are, in any meaningful sense, more interventionist than countries like Austria.

The other fact which contributes to a revision in causal beliefs is that, since the 2008-09 economic crisis, the unemployment performance of the United States is not at all impressive. As figure 1 shows, U.S. unemployment caught up with Europe's in 2009. Furthermore, U.S. unemployment does not behave has it did after the previous recessions—i.e. rapidly shrinking back to its pre-recession level. Instead, it is now slowly and hesitantly drifting downward.[2] In contrast, the unemployment figures of countries such as Germany and Austria have been mildly modified by the recent crisis, and are now more inspiring that the U.S. performance (see figure 2). These countries might thus become attractive models of wise labor-market interventionism.[3]

In this thesis, I will not investigate what are the real causes of high unemployment. You are reading a thesis in philosophy of science, not in applied economics. The above story about unemployment was meant both to introduce the case study which runs throughout the thesis—i.e. economic *research* on unemployment—and to drive three points home. First, economists are studying matters of great importance. Given the economic distress generated by situations of high unemployment, one social objective should be to keep unemployment low. If economics can

[2] The last data from the Bureau of Labor Statistics report a U.S. unemployment rate of 8.2% in May 2012, well above its pre-recession trough of 4.4% (May 2007).

[3] This process has already begun, as I will partly document in chapter 5.

help us with this goal (as it promises), it is a highly important inquiry. Second, if there is knowledge generated by economic research, this is primarily *causal* knowledge. Economists generate claims about what causes high unemployment, and try to justify these claims. Third, it is not obvious that knowledge on issues like the causes of unemployment is steadily accumulating. Economists entertain various beliefs on this topic, these beliefs change through time, and it is not clear—at least not at first sight—that the current beliefs are significantly better justified than the previous ones.

These three points are meant to motivate the project of this thesis, which is a philosophical study of causal reasoning in economics. The hope is that, by taking a step back, we come to better understand this important project pursued by generations of great thinkers. This improved understanding should, in turn, equip us to better assess the practice of causal reasoning, and to suggest ways of reforming this practice if it is found wanting. The reader should note, however, that this thesis is far more about improving our understanding than about performing the assessment and reaching prescriptions. I am wary of premature assessment based on a faulty understanding. If my philosophical investigation contributes to a better *understanding* of causal reasoning in economics, I will humbly judge my efforts to have been worthwhile.

Some readers might find my project surprising: don't we already know the ins and outs of causal reasoning in disciplines like economics? In fact, a red thread going through my thesis is that we still have limited knowledge on this topic. Much of philosophy of science has not been developed with an eye on economics. The result is that most off-the-shelf ideas in philosophy of science shed little light on economics (Hands, 2001). Furthermore, economics turns out to be extremely rich. It is what I will call here an eclectic science, a science characterized by variety and combination. This type of science can certainly not be well understood overnight.

I do not want to exaggerate the singularity of my project. A strength of it is, in fact, that it inserts itself in a strong revival of interest for the concept of causation. It is thus a timely project which can draw on and contribute to lively discussions. The causal revival is obvious both in philosophy and in economics. In philosophy, a profusion of approaches to the metaphysics and semantics of causality are now on offer. The debate over the 'correct' approach to causality continues unabated, but with a twist: the debate is now partly about whether one can (or ought to) be pluralist (and in what sense) about causality.[4]

In economics, journals are now filled with articles sporting "causal effects" and similar phrases in their titles. Economists seem to have now totally recovered from the aversion to explicit causal language that they for some time shared with most social scientists (Hoover, 2004; note that most social scientists outside economics still seem to be 'causality averse'). Although there is some debate in economics about the *meaning* of causal claims,[5] it is fair to say that the most urgent problem for economists is epistemic: How can we learn about causal relations? How can we gather compelling evidence for a causal claim?[6]

For scholars like me straddling philosophy and economics, the causal revival in both fields brings the opportunity to make distinctive contributions to the conversations both among philosophers and among economists.[7] My thesis attempts to make such contributions. I believe that my efforts to better understand causal reasoning in economics can be of value both to philosophers inquiring about 'causation', and to economists busy finding out 'what causes what' in the social world.

The three parts of my thesis look at different aspects of causal reasoning—i.e. semantics, epistemology, and dynamics. One can think of each

[4] One way to organize the literature is to rely on Hall's (2004) distinction between causation as dependence and causation as production. *Dependence* accounts are then subdivided in regularity accounts (e.g. Mackie, 1974), probabilistic accounts (e.g. Suppes, 1970; Cartwright, 1979), counterfactual accounts (e.g. Lewis, 1973, 1979), and interventionist accounts (Woodward, 2003); while *production* accounts include process accounts (Salmon, 1998; Dowe, 2000) and (complex-system) mechanistic accounts (e.g. Glennan, 1996, 2002; Machamer et al., 2000; Bechtel and Abrahamsen, 2005). This way of classifying approaches is not totally satisfactory. On the one hand, the approaches are not really mutually exclusive—e.g. the recent trend to mix mechanistic and interventionist accounts (Glennan, 2011). On the other hand, the classification is not exhaustive. One must also take note of recent inferentialist accounts of causality (e.g. Williamson, 2005; Spohn, 2006; Reiss, 2012) and of approaches emphasizing pluralism (e.g. Cartwright, 2007; Longworth, 2010). For a discussion of the different ways to be pluralist, see Hitchcock (2007).

[5] For a review of positions in this debate, see Hoover (2006).

[6] Recent general discussions about causality in economics include Heckman (2000), Hoover (2008), and Granger (2007). Two great contending approaches to causal inference—but by no means the only ones—are the structural approach (Heckman, 2005, 2008) and the design-based approach (Imbens and Wooldridge, 2009).

[7] Prominent scholars at the intersection of philosophy and economics who are already making such contributions include Nancy Cartwright (2007), Daniel Hausman (1998), Kevin Hoover (2001), and Julian Reiss (2007). Note that the boundaries are necessarily fluid between pure philosophy, pure economics, and the field at the intersection of the two; many more scholars interested in causation connect philosophy and economics to some extent.

part as giving elements of an answer to three broad questions. First, what are the meanings of causal claims? Second, how can a causal claim be adequately supported by evidence? Finally, how are causal beliefs affected by incoming facts?

There are many ways to contribute to the ongoing conversations on these questions. My approach is characterized by *selectiveness* in two senses. I explain these senses in the next section. I then give a foretaste of what appears in the three parts of this thesis.

A *selective* exploration

As I just said, while my thesis considers extremely broad questions, my approach to answering them is, in two ways, selective.

First, the answers given are not comprehensive. In part I and II, I do *not* maintain that *all* causal claims must mean this or that, nor do I present an allegedly *exhaustive* list of methods to provide adequate evidential support for a claim. Instead I argue that the meaning of *some* claims is distorted by a leading causal semantics, and give the contours of an alternative semantics. Similarly, I argue that *one* widespread way to gather strong evidence for a claim is too often overlooked. The same pattern appears in part III. I first restrict my inquiry into belief dynamics to situations where the incoming information is deviant. I thus focus on deviant-case research—i.e. the attempt to account for cases deviating from what was expected. I argue that a widespread philosophical story about deviant-case research does more harm than good in helping us understand and assess *one* instance of deviant-case research. Although I work out the skeleton of an alternative story, I do not commit myself to the thesis that the original story and my alternative story span the space of relevant narratives for deviant-case research.

Comprehensiveness is highly valued, and sometimes for good reasons. My reason for refraining from upholding comprehensive answers to my questions has to do with my understanding of philosophy of science and its object. This understanding is best discussed after introducing the second way in which my contributions are selective, to which I now turn.

This second way of being selective is that my contributions are case-based. More specifically, the thesis mainly draws on one instance of scientific practice to shed light on fundamental questions about causal reasoning. As I already said, my case study is the literature on the causes of aggregate unemployment. This literature forms a narrow subset of economics, and I do not pretend that it is ‘representative’ even of what economists do (and certainly not of science in general).

Much of contemporary philosophy of science is case-based. The great advantage of case-based philosophy of science is thus widely recognized: the risk is smaller that the philosophical account of science so produced is widely off the mark. One can indeed interpret the focus on case studies in contemporary philosophy of science as a reaction to previous accounts which are now judged to be misleading depictions of science. There are however risks associated with case-based philosophy of science.

The obvious danger is unwarranted generalizations. Studying a case might lead us to a better understanding of *this* scientific practice, but there seems to be a major leap involved in interpreting this narrow study as providing a philosophical lesson for science in general. It would however be too quick to conclude that general philosophical lessons cannot result from a case study. The first thing to note is the power of a single counterexample—i.e., the power of *modus tollens* as emphasized by Karl Popper ([1959] 1992, § 18). Since many philosophical accounts are presented as universal generalizations, exhibiting a case where the generalization appears to break down is already an achievement; it is evidence against the purported universality of an account. Some of what I do here can be interpreted as providing counterexamples to purported universal accounts.

There is also another, and more important way in which a case-based philosophy of science can be relevant to our understanding of science beyond the cases that are actually studied: conceptual innovation (Chang, 2011; a similar point is made by Burian, 2001). Much of what is provided by philosophers of science are abstract conceptual frameworks to understand and assess scientific practice. In the process of studying a concrete case, one can come to the conclusion that the available philosophical concepts tie the case to a Procrustean bed.[8] The case study can thus be used as a concrete platform to develop new abstract frameworks. The first task of the concepts proposed is to foster understanding for the case under study, but these abstract concepts also enrich the philosophy of science more generally by giving us additional lenses to interpret and assess other instances of scientific practice.

Even if the reader were to grant that philosophy of science might im-

[8] Examples of philosophers reaching this conclusion abound in the history of the philosophy of science, from the rejection of the deductive-nomological model of explanation for history (e.g., Scriven, 1959), to the turn away from the Lakatosian framework as a way of understanding New Classical Macroeconomics (e.g., Hoover, 1991).

prove as we add to it conceptual frameworks appropriate to understanding some cases of actual scientific practice, a risk of the case-based approach would remain: extreme locality. The whole exercise of conceptual innovation would have little value if the new concepts are likely to be unhelpful beyond the case from which they arise. Philosophy of science would end up in a sorry state if it turned into an enormous toolbox in which each tool can be used on a single occasion only.

I state it up front: I am not in a position to dispel completely the locality worry. For the key concept in the epistemology part—i.e., evidential variety—there is ample evidence of relevance beyond the economics of aggregate unemployment. But the reason why this evidence exists is that the concept is far from new (though often overlooked, as I argue). In the two other parts where I am compelled to be more innovative, I cannot rely on much preexisting work to support the claim that my conceptual frameworks can travel to many other cases. I do, here and there, reach out to other scientific examples, which should alleviate the worry about *extreme* locality. However, I do not attempt in any systematic way to assess the likely scope of applicability of my ideas. I hope to be able to better assess this scope in future work.

In any case, I do not expect my ideas to be enlightening for all instances of scientific practice. This expectation comes from the commonplace that science is not monolithic, that there is a wide variety of scientific practices. There is thus little that one can said about 'science in general', and the specific things I want to say about science are unlikely to fit the whole of science. This obviously does not mean that all accounts will have an extremely local scope—we should still aim to have conceptual frameworks that can travel from cases to cases.

In short, the reasons why I am happy with *non-comprehensive* and *case-based* answers to my questions come down to my view of science as plural, and my view of philosophy of science as providing abstract conceptual frameworks to help us understand and assess this plural science. Philosophy of science is foremost there to give us—'us' also including laymen and practicing scientists, not just philosophers—the vocabulary to reflect on our scientific practices. To fulfill this function, philosophy of science should be attentive to concrete scientific practices because one misses much by disserting about 'science in general'. Furthermore, when the study of a case prompts one to innovate conceptually, it would be too bold to maintain that the conceptual innovation is useful for the study of all scientific practices. I even tend to think that the single-minded quest for comprehensive answers would most likely do disservice to philosophy of science. A reflection-enabling discipline should guard itself against being overtaken by Procrustes.

Foretaste

The five chapters of this dissertation are meant as timely contributions to ongoing discussions about causality in science. They are written as self-standing research papers and should thus need little introduction. In this section, I give a short summary of what happens in each part as a teaser for the reader.

In part I, I concentrate on the meaning of causal claims. Economists with a methodological bent and many other scholars (some sociologists, philosophers, statisticians, computer scientists) are converging on an answer to this issue: the manipulationist-counterfactual account of causation. To be sure, these scholars do not give the exact same answer, but the shared ground is large. They all agree that the relata figuring in a causal claim denote entities[9] that are asymmetrically related such that some ideal (usually counterfactual) manipulation of the purported cause will change the (probability of the) purported effect. This semantic analysis is non-reductive in the sense that it uses the overtly *causal* concept of counterfactual manipulation to analyze the meaning of a causal claim. Furthermore, this account coheres with a metaphysics according to which causal claims are about a *causally-structured* world—i.e. causal relations between entities actually exist, and our causal verbs pick out these relations.[10]

In being non-reductive, the counterfactual-manipulationist account breaks with a standard goal of philosophical theories of causation, i.e. analyzing causal claims with non-causal concepts (e.g., regular association, possible worlds). By focusing squarely on the goal of making more

[9] 'Entities' is used here in the most liberal sense. Proponents of this account are not particularly concerned with restricting the set of admissible causal relata to some kinds of entities (e.g. events versus facts).

[10] For instance, Kevin Hoover (2001, p. 23) explicitly endorses such a metaphysics. The central thesis of his *Causality in Macroeconomics* is that "*Causal structures are fundamental*". And these structures are understood as existing "externally and independently of any (individual) human mind". I am careful in writing that this account of causality *coheres with*, instead of *comes necessarily with*, a realist metaphysics of causation because some proponents of this account avoid explicit metaphysical commitments. For instance, James Woodward (2003, p. 7) "leave[s] it to the reader to decide whether [his work] counts as discovering 'what causation is.'"

precise our causal talk, it has already greatly contributed to our understanding. It is helping us better communicate our causal propositions, and is also tightly linked with powerful methods of causal discovery. But does this account capture the meaning of *all* causal claims?

In chapter 1 (co-written with Luis Mireles-Flores),[11] we argue that, despite its great merits, the manipulationist-counterfactual account leads one to misinterpret the meaning of some causal claims. The problem lies in a fundamental presupposition of the account: a referentialist approach to meaning (Reiss, 2012).[12] As should already be clear from my short introduction to the account, the meaning of a causal claim is analyzed by pairing components of the sentence (i.e., the causal relata, and the causal connection) to worldly entities. The account is particularly strong at specifying different species of causal relations which can hold between the entities denoted by the relata. When analyzing the meaning of a specific causal claim, the questions are thus: Which causal systems are referred to? Which entities in these causal systems are denoted by the relata? And which causal relation is asserted to hold between these relata?

We argue that these questions lead to unsatisfactory answers when applied to (at least) one type of causal claims: generalizations in policy-oriented social science such as 'more generous unemployment benefits increase unemployment' or, even more general, 'the main cause of high unemployment in OECD countries is the low flexibility of their labor markets'.[13] A manipulationist-counterfactual analysis of the meaning of such generalizations leads to the conclusion that they are highly implausible, ambiguous or confused. This conclusion leaves one wondering why they are so highly valued in the communities where they are found.

We argue that the puzzle disappears when we drop referentialism and opt for an alternative, inferentialist approach to semantics. From the

[11] The work on this chapter was done in multiple stages. A first draft was written by me in the spring of 2011. We then worked on this draft in turn, each of us producing a new version before the other one took it over. A major change occurred at version 7 (spring of 2012): having grown dissatisfied with our referentialist semantics, we decided to oppose it to an inferentialist semantics. I wrote this expanded version and the current chapter is a close descendant of this version.

[12] This presupposition is no exotic characteristic of the manipulationist-counterfactual account. Referentialism (also called truth-conditional semantics) has always been the dominant way to analyze meaning in linguistics and the philosophy of language. Take for instance the first sentence of an influential textbook on semantics by Irene Heim and Angelika Kratzer (1998, p. 1): "To know the meaning of a sentence is to know its truth-conditions." The authors go on to explain that the meaning of a sentence as to do with "what the world would have to be like for it to be true".

[13] These are the actual causal claims analyzed in chapter 1.

standpoint of inferentialism (also known as conceptual-role semantics), the key question to ask is not the word-world questions above, but rather: What is the role of the statement in the inferential practices of the language users? In answering this question, one comes to realize that claims that seemed to border meaninglessness from a referentialist perspective are, in fact, central to the inferential network of the language users. We maintain that the fact that an inferentialist approach can rationalize the widespread practice of demanding and supplying policy-oriented causal generalizations counts in favor of inferentialism, and against referentialism, as an approach to the semantics of *this* type of causal claims. The general lesson of this chapter is thus that forcing all causal utterances to fit the mold of a referentialist semantics would be akin to playing the role of Procrustes.

In part II, I turn to an epistemic question: How can a causal claim be adequately supported by evidence? There is a strong bias in methodological discussions toward single-method assessment. Typical questions include: How to design an experiment such that it is a silver-bullet test of a hypothesis? How to insure that our regression parameters are unbiased when using observational data? The hope seems to be that, once our preferred method is based on a sound methodology, we can generate strong evidence for or against our causal claims by using this *single* method. My contribution starts from the recognition that, in many epistemic situations, a single source is not likely to generate evidence powerful enough to entitle someone to strongly belief a hypothesis. The reason is simple: there is no principled way to guard oneself against *all* potential sources of unreliability of a method. A sound methodology can increase the reliability of a method, but it cannot guarantee it.

What if one has different methods available? It is intuitive to think that pulling together different sources of evidence can be conducive to confirmation. Part II is a development of this intuition. It contains three of my five chapters, which come after some preliminaries. In chapter 2, I concentrate on an ongoing debate in economics that can be framed by the question: What is the most effective way to seek *credible* causal inference in policy-oriented economics? Two approaches take center stage: the design-based approach and the structural approach. I argue that the answers given by these two approaches are characterized by the single-method bias just hinted at. I further maintain that single-method approaches are not likely to be successful in (at least) some epistemic contexts like the one of the economists working on the causes of aggregate unemployment. Luckily, these economists can draw on more than one method. The main goal of the chapter is thus to show that evidential variety—i.e. combining evidence from multiple sources—is both *actually* conducive to credible inference in this epistemic community, and has sound epistemic credentials.

Chapter 3 uses the same notion of evidential variety to contribute to a debate in the philosophy of causality. Federica Russo and Jon Williamson (2007) draw many implications from their observation that *health* scientists rely on both difference-making and mechanistic evidence to establish causal claims. In other words, these scientists have a dualist causal epistemology. I argue that, while a similar practice can be found in my case study from the *social* sciences, two implications drawn by Russo and Williamson are unwarranted. The fundamental reason why their conclusions are inappropriate is that the epistemic practice observed is best interpreted—at least in my case study—as a use of evidential variety. It follows that what I label the first Russo-Williamson Thesis is incorrect: it is not the case that the two types of evidence *must* be present in order to establish a causal claim. I indeed find that one claim lacks difference-making evidence, but is nevertheless consensual in the relevant epistemic community.

My second and most important point in chapter 3 is that the economists in my case study should not be worried that their (typically) dualist causal epistemology is incompatible with the counterfactual-manipulationist *semantics* of causality which, as I already said, is popular in economics. Against what I label the second Russo-Williamson Thesis, I argue that this well-known (monist) account of causality can perfectly make sense of the dualist epistemology observed in my case study. Though how we *seek* causes might inform us about what we *mean* by our causal claims, I find no support for the particular link drawn by Russo and Williamson between epistemology and semantics.[14]

Although I attempt to make the notion of evidential variety more precise in chapters 2 and 3, there is nevertheless much more that can be said about it. Chapter 4, by far the most formal chapter of this thesis, attempts to get a better understanding of ‘variety’ and its implication for

[14] As should be obvious from my summary of part I, my argument here should not be read as committing me to the counterfactual-manipulationist account. I simply argue that the considerations put forward by Russo and Williamson do not seem to be good grounds to reject this account. There is no inconsistency in thinking simultaneously, as I do, that the counterfactual-manipulationist account gives an inappropriate semantic analysis of some causal claims—this is the argument of chapter 1. If we come to reject this account for some (or all) causal claims, we should do it for the good reasons.

confirmation. What do we mean when we say that a body of evidence is more or less varied? Is the variety-of-evidence thesis true—i.e., confirmation increases with variety, *ceteris paribus*? As is pointed out earlier in this part (section 2.5), a promising analysis of variety when it comes to evidential sources is in terms of reliability independence: there are reasons why a given source would be an unreliable indicator of the correct hypothesis, and the less these reasons overlap between sources, the more varied the body of evidence is. Using this analysis of evidential variety, Luc Bovens and Stephan Hartmann (2002; 2003) offer a Bayesian model with a counterintuitive result: in some epistemic situations, *more* variety implies *less* confirmation, *ceteris paribus*. At first glance, this result threatens my argument that evidential variety is a useful tool to learn about causes and, most importantly, it casts doubt on the epistemic credentials of the actual scientific uses of evidential variety. Unfortunately, Bovens and Hartmann say next to nothing on the implications of their formal result for evidential variety *in practice*.[15]

Chapter 4 takes a second look at Bovens and Hartmann's model and finds it wanting. The specific way in which they model reliability dependence seems a far cry from how scientists typically think about the relation between their evidential sources. But what would be the result if the model did appropriately capture reliability independence? I try to answer this question by modifying the model of Bovens and Hartmann. The outcome of my modeling effort is neither totally compatible with Bovens and Hartmann's result nor totally at odds with it. Contrary to them, I find that having *fully* independent sources is always more conducive to confirmation than having *fully* dependent ones, but there are in my model special epistemic situations where having *some more* independence is detrimental to confirmation. The implications of these results for scientific practice are unfortunately not transparent. There is obviously the legitimate worry that such a model cannot be taken seriously given the grand idealizations required in building it. If we do grant its adequacy, I maintain in the conclusion of the chapter that it leaves the variety-of-evidence thesis in a pretty good state after all. The thesis cannot be taken as true for all epistemic situations, but it might well remain a good, though fallible, methodological guideline given that the model

[15] Hartmann (personal communication) maintains that the main goal of their model was to generate a possibility result: it is indeed possible to conceive of an (artificial) epistemic situation where the variety-of-evidence thesis (under a specific interpretation of variety) is turned upside down.

shows it wrong only in extreme epistemic situations.[16]

There is a picture of economics—or at least the subset of economics on which I focus—emerging from the first two parts of this thesis. One might label this picture 'eclectic science'.[17] It is, first, semantically eclectic. It is not only that some of its statements might be appropriately interpreted with a referentialist semantics while an inferentialist semantics seems more appropriate for other statements. It is also that the inferential connections of the latter statements are multifarious and cannot be organized in a neat theory with a few axioms and a deductive structure.[18] Second, economics is epistemically eclectic. In an attempt to make a stubborn world speak, economists probe it from multiple angles and bring together the different outputs of these fallible investigations.

We have much to learn about the functioning of such eclectic science and much careful work to do in order to formulate appropriate methodological guidelines for it. Part III of the thesis is a contribution to this project. I focus on understanding and assessing the dynamics of an eclectic science when it faces a deviant case. Deviant cases—sometimes called puzzles or anomalies—are central to a common post-positivist story about scientific dynamics. This story has its roots in the work of Karl Popper, Imre Lakatos and Thomas Kuhn among others, but I am not particularly interested in exegesis here. The post-positivist story I have in mind understands a case to be deviant if it is *inconsistent* with an empirical proposition deduced from a set of theoretical and secondary propositions. Dynamics is set in motion in order to restore consistency by finding out which proposition(s) to give up.

In chapter 5 I look at a specific instance of deviant-case research in economics: the research on the deviant behavior of the German unemploy-

[16] 'Extreme' in the sense that the evidential sources must be *highly distrusted* for the thesis to break down.

[17] I am still unsure what the most appropriate term is. One could follow Wimsatt (2007a) in using 'piecemeal', but this word emphasizes disconnection between elements while I want connectedness to be front and center. It also seems to me that 'piecemeal' is pointing to a view of science in which the telos is still the 'grand theory of everything' but we have to get there step by step. Cartwright's 'dappled' is also deemphasizing the 'bringing together' and seems anyway a better term for the world than for a science. I could perhaps use 'heterogeneous' but this one does not say much. I thus select 'eclectic' more from a lack of better ideas (linked to the poverty of my English lexicon) than from a belief that the term expresses compactly exactly what I want. I welcome suggestions.

[18] Note how my picture of economics diverges from a well-known one where general equilibrium theory is the centerpiece and economics is the only axiomatized *social* science.

ment rate in the 2008-10 economic crisis. I argue that the post-positivist story is unhelpful in interpreting and assessing this case, and I develop the skeleton of an alternative story. The reason why the post-positivist story is inappropriate is that Germany's deviance did not appear on the background of a neat theory which generates empirical propositions deductively. Germany deviated from expectations which were generated by a more eclectic process than the story presumes. After answering this question about what makes a case deviant in an eclectic science, I offer answers to two other questions. What is the epistemic goal of deviant-case research? And how should this research proceed? The alternative story that I flesh out shows some of the peculiarities of causal inference in an eclectic science.

This teaser should give the reader a good idea of my arguments in the five core chapters. I hope that it also transmits my own astonishment at the richness of causal reasoning in economics. To me, it is no wonder that we still have a rather poor philosophical understanding of causal reasoning in sciences like economics. We are trying to understand scientific strategies developed to grapple with a really difficult subject matter. Although analyzing these complex strategies is hard, achieving a better understanding of them might enable us to develop more self-reflective scientific practices in the future. My (almost foolish) hope is that my own work, by contributing to enable such reflection, will foster the effective pursuit of economic research.

Part I

Semantics

Chapter 1

Semantic Analysis of Causal Generalizations in Policy-Oriented Social Sciences

1.1 Introduction

Causal generalizations are one product of economic research. A reason why economists formulate and attempt to justify causal generalizations is that there is a demand for them. Policy makers expect, and often request, that economists supply these knowledge claims. Take these requests from the final communiqué of the Ministerial Meeting of the Organisation for Economic Co-operation and Development (OECD) in May 1992:

> Ministers invite the Secretary-General to initiate a comprehensive research effort on *the reasons for and the remedies to* the disappointing progress in reducing unemployment [...] (OECD, 1992, emphasis added)

The Ministers wanted the economists of the OECD to tell them why a host of countries were struggling with persistently high unemployment, and they wanted them to formulate policy recommendations. The first request was not read by the OECD research team as a demand for a specific explanation for each country, but rather for causal generalizations that would explain all these disappointing performances.

Economists working at the OECD met the Ministers' requests by producing the landmark *OECD Jobs Study* (1994a; 1994b). This report presents the following causal generalization as its main result regarding the request for explanation:

> [I]t is an inability of OECD economies and societies to adapt rapidly and innovatively to a world of rapid structural change that is *the principal cause* of high and persistent unemployment. (OECD, 1994a, part I, p. vii; emphasis added)

In short, the report maintains that inflexibility is the main cause of high unemployment. In addition to this broad generalization, the report also claims to have identified many more specific causes of high unemployment, including government-imposed barriers to wage flexibility, limited geographic mobility, employment protection legislation, lack of proper training, taxation, and unemployment benefits programs. Each of these causes finds its place in a causal generalization such as: "[r]elatively high unemployment benefit entitlements tend eventually to increase unemployment" (OECD, 1994b, p. 38). The *Jobs Study* thus supplied a broad causal generalization about inflexibility and narrower ones about more specific causes, for instance unemployment benefits.

Since the two causal generalizations already mentioned are our[1] main examples throughout this chapter, let us highlight them for future reference:

1. In OECD countries, the inflexibility of labor markets causes high unemployment.

2. More generous unemployment benefits cause higher unemployment.

What do generalizations such as these two mean? There is a long tradition of philosophical worries about the meaning of generalizations in the social sciences. One nagging worry is that, in not being the *strict* law statements we purportedly find in physics, social-scientific generalizations would be semantically defective. They would be implicitly hedged with a *ceteris paribus* clause, and the meaning of this clause would remain elusive.[2] The unresolved issue of the meaning of social-scientific generalizations greatly contributed to the pervasive doubts about the scientificity of the social sciences in 20th-century philosophy.

[1] Since this chapter is the result of joint work with Luis Mireles-Flores, I will use the first person plural here. I use the singular elsewhere in the thesis.

[2] For a review of the competing semantics of the (usually implicit and perhaps inexistent) *ceteris paribus* clause, see Reutlinger et al. (2011).

The contemporary philosophical literature on causality promises finally to pin down the meaning of these generalizations. The work of James Woodward is emblematic of this literature. His book *Making Things Happen* is explicitly about "capturing or clarifying [the] 'content' or 'meaning'" of causal claims (Woodward, 2003, p. 7, see also p. 38). We want to stress at the outset that we see this literature as a great contribution to our understanding of causation. In particular, it provides a host of careful distinctions which have already enriched our causal language.

In this chapter, we aim first to take stock of this literature by formulating a procedure to identify the meaning of causal generalizations such as the ones formulated in the *OECD Jobs Study.* We think of this procedure as our first contribution to the literature. Our second aim is however to provide an alternative to this procedure because we consider that it falls short of its goal. In short, it fails to identify the meaning of the causal generalizations that we consider because it relies on an inadequate semantics, i.e. a referentialist semantics. We provide evidence that an inferentialist semantics is more appropriate by showing how illuminating it is when applied to the OECD's generalizations.

For our purposes, an important characteristic of the contemporary literature on causality is its explicit desire to be relevant to practicing social scientists by suggesting ways to improve practice. Again, Woodward can function as an exemplar:

> [M]y project has a significant *revisionary* or *normative* component: it makes *recommendations* about what one ought to mean by various causal and explanatory claims, rather than just attempting to describe how we use those claims. It recognizes that causal and explanatory claims sometimes are confused, unclear, and ambiguous and suggests how these limitations might be addressed. (Woodward, 2003, p. 7, emphasis in the original)

This revisionary project is a noble one to which we are sympathetic. It is however a dangerous project. One is reminded of the numerous attempts by philosophers to press social scientists to comply with a standard of good science, which turned out, in retrospect, to be wrongheaded. Are the recommendations coming out of the recent causal literature to which Woodward and many others participate part of this set of wrongheaded suggestions? This is indeed what our inquiry leads us to fear. Because we think that these recommendations are based on too limited a view of what a causal claim can express, social scientists would wind up impinging on the expressive power of their causal language if they were to follow them closely.

We stress that our argument applies only to *part of* causal reasoning and to *some* causal generalizations. It might well be that a referentialist semantics is the appropriate approach for some other types of causal claims. If this speculation happens to be correct, the conclusion would be that the current literature is faulty of overgeneralizing a semantic approach which is fruitful for some claims, but misguided for others.

1.2 On the meaning of causal generalizations: a referentialist approach

The dominant approach to semantics is referentialist (Heim and Kratzer, 1998; Speaks, 2011; Peregrin, 2012, p. 3).[3] To be more precise, the dominant approach to a theory of meaning is to start with a theory of reference. Most thinkers then add another layer to deal with the fact that two extensionally-identical expressions might intuitively have different meanings—e.g. Frege's (1892) famous evening star versus morning star. Fortunately, we don't need the intension/extension distinction for our purposes, and will thus treat an entirely referentialist theory as a full theory of meaning.[4]

According to a referentialist approach, the meaning of words is constituted by what these words stand for, by what they *refer* to. Nouns stand for objects, predicates stand for properties and relations, and the meaning of compound statements depends entirely upon the meaning of their constituents. Reaching the level of a full sentence, meaning is truth-conditions; it is given by "what the world would have to be like for [the sentence] to be true." (Heim and Kratzer, 1998, p. 1) For example, 'snow is white' is true if and only if all the elements that are part of the set of objects referred to as 'snow' are also part of the set of objects with the property referred to as 'white'. By pairing in this fashion the subject and the predicate of our simple sentence with objects and properties in the world, we state the meaning of 'snow is white' according to referentialism.

Referentialism is the standard semantics used in the recent philosophical literature which attempts to spell out the meanings of causal claims. Note that there are a few exceptions—the approach of Julian

[3] This approach takes other names: truth-conditional, representationalist, extensional.

[4] For different ways to add intension to a theory of reference, see Speaks (2011).

Reiss (2011b, 2012) being the one to which our analysis of section 1.4 is most indebted.[5] But for the most part, the meaning of sentences of the form 'X causes Y' is identified, on the one hand, by specifying the referents of the causal relata X and Y and, on the other hand, by providing an analysis of the causal verb in terms of what relation it stands for. The meaning of the whole causal proposition is then equated to its truth-conditions: what would the world have be like for the referents of the causal relata to stand in this relation?

In this section, we build on this literature to propose a systematic, *referentialist* procedure to identify the meaning of a causal generalization. We start by proposing the following schematic form of a causal generalization to identify the elements in need of semantic explication:

$$\text{(For } P\text{,) } X \hookrightarrow Y \tag{1.1}$$

In this formula, 'For P' specifies the relevant population. This clause is often left implicit (for instance, in our previous paragraph), hence we put it here in parentheses. X and Y are the causal relata, and '$\hookrightarrow$' stands for a causal relation, where the causal influence goes from X to Y.

This formula can be used to express compactly our two main examples highlighted in the introduction to this chapter. Let U be the unemployment rate, which is the purported effect in both generalizations. Let *Inflex* be the degree of inflexibility of the labor market, and let B be the degree of generosity of unemployment benefits. Finally, let the superscript $\iota = \{+, -\}$ apply to $\hookrightarrow$ such that $\overset{\iota}{\hookrightarrow}$ expresses either 'positive cause' or 'negative cause'. The inflexibility claim and the claim about unemployment benefits can thus be expressed as

1. (For OECD countries,) *Inflex* $\overset{+}{\hookrightarrow} U$
2. (For P,) $B \overset{+}{\hookrightarrow} U$

Our main goal in using examples (beyond illustrating the analysis) is to demonstrate that causal generalizations such as the ones from the OECD have a wide variety of potential meanings under a careful referentialist analysis. We use the term *semantic complexity* to refer to this property of a statement to have a wide variety of potential meanings, a variety which is not obvious from looking at the surface structure of the statement.

In this section, our analysis will mainly focus on claim 2. If our point about semantic complexity comes out clearly by using this claim, it

[5] Other, broadly inferentialist approaches include Williamson (2005), Spohn (2006), and Beebee (2007).

must be obvious that the conclusion should also hold for the *broader* claim 1. Note also that there does not seem to be anything special about these generalizations—as generalizations in policy-oriented social science—which would make our conclusion hold *only* for these generalizations. Numerous generalizations seem amenable to the same analysis. The following two generalizations—which will play a secondary role in this section—are part of the lot:

3. Short-time work schemes cause lower unemployment in periods of crisis.[6]

4. Free trade causes economic gains.

Claim 3 came to prominence recently in the attempt, by the OECD and others, to explain the different employment performances of countries during the 2008-9 economic crisis (OECD, 2010, ch. 1).[7] Its existence suggests that the work on unemployment by the OECD still results in the formulation of generalizations that are much like the ones of the 1994 *Jobs Study.* Claim 4 is widespread and suggests that our analysis could extend beyond research on unemployment.[8]

To identify the meaning of causal generalizations, we propose to answer four main questions:

i. What do the different values of the causal relata X and Y refer to? In other words, what possible changes in the world are meant to be captured by a variation in the relata?

ii. What are the units composing the relevant population?

iii. What relation does the causal verb refer to? This broad question can be further divided in two:

    a) What relation is referred to in the underlying *unit* causal claims?
    b) Which sets of unit causal claims are entailed by the causal generalization?

iv. In sum, what are the truth-conditions for the causal generalization?

We take each question in turn in the following subsections, and show how answering them reveals the semantic complexity of our examples.

[6] Short-time work schemes are public schemes inciting employers to temporarily reduce the number of working hours of their employees instead of laying them off.

[7] This claim comes back in chapter 5 below.

[8] Luis Mireles-Flores (2013) analyzes this claim about free trade.

1.2.1 The meaning of the causal relata

Following a strong trend in the philosophy of causation and in conformity with general usage in economics, we take X and Y to be variables, and use lowercase italics (x and y) to represent specific values of these variables.[9] The first step in our referentialist procedure is to ask what exactly these variables stand for or, more to the point, what actually is meant to change in the world when these variables take different values. By using the generalizations in the *OECD Jobs Study*—mostly (For P,) $B \overset{+}{\hookrightarrow} U$—we now show that there are multiple answers to this question for social-scientific variables such as U and B.

The unemployment rate U was the effect variable under scrutiny in the *Jobs Study*. So what is the meaning of this variable? What does a change in the unemployment rate refer to? The unemployment rate of an economy is defined as a ratio between two head counts: the number of participants in an economy having the status to be both ‘active and jobless’ and the number of participants being ‘active’.[10]

The semantic complexity of U comes down to the specification of the two relevant categories—active and jobless. Specifying these two categories—who to include, who to leave out—requires a host of decisions. Precisely for this reason, the International Labour Organization made an effort to provide detailed guidelines on how to define and measure these categories (ILO, 1982, p. 2-5). These guidelines actually helped pinning down a more definite referent of the concept ‘unemployment rate’, but two sources of semantic complexity remain.

First, the guidelines leave some substantial margins of interpretation. For instance, a necessary condition for an individual to be among the ‘active’ is that she is ‘willing to work’, which is translated as being engaged in ‘active job-search’ when the individual is out of work. But ‘active job-search’ is notoriously open to interpretation. The OECD (1994a, part II, p. 186) indeed asserts that the condition “is in some countries interpreted rather widely”.[11]

[9] We take the *values* of a variable to stand for events.

[10] In an equation, the unemployment rate looks like:

$$U = \frac{\#[i|\text{active}(i) \cap \text{jobless}(i)]}{\#[i|\text{active}(i)]} \tag{1.2}$$

where i stands for an individual, active(i) and jobless(i) mean that the individual falls in the category ‘active’ and ‘jobless’ in the relevant economy, and # reports the cardinality of the set of individuals falling in these categories.

[11] The OECD continues:

Second, choices made following the guidelines do not necessarily line up with what one would intend or intuitively expect the unemployment rate to stand for. Consider the 'jobless' category for instance. The set 'employed' (i.e., the complement of the jobless) is interpreted by statistical agencies—if only because of data limitations—as 'employed in the formal sector'. The implication is that the set 'jobless' includes individuals who might actually be employed in the *informal* economy. But one might intend (or expect) a causal claim about unemployment to be about unemployment *tout court*, and not only about the proportion of 'active' not employed in the formal sector (e.g. Schneider and Enste, 2000, p. 106; Tyrowicz and Cichocki, 2011).

Since a referentialist semantics understands the meaning of a sentence as being determined by the meaning of its constituents, the actual referent of 'unemployment rate' obviously matters for the meaning of the causal generalizations in which it occurs.[12]

Semantic complexity is compounded when we consider the cause side of the OECD's generalizations. The semantic analysis of variables such as 'inflexibility of a labor market' (*Inflex*) and 'generosity of unemployment benefits' (B) is more challenging than that of U because they are best understood as multidimensional variables. These variables can indeed be represented as vectors, with each component of the vector standing for a dimension of the relevant concept.[13] For instance, a first attempt to capture the meaning of B would (at least) distinguish between three dimensions: the level of benefits B_l, the duration of entitlement B_d, and the eligibility conditions B_e (see Nickell et al., 2005, p. 4; Boeri and van Ours, 2008, sec. 11.1). Yet, specifying B as a tridimensional vector is still

> [G]reater standardization, for example with a consistently strict interpretation of the notion of 'step of active job-search', could make a significant difference to the level of unemployment reported in labour force surveys for some countries. (OECD, 1994a, part II, p. 187)

One is reminded of the credo among some economists that 'involuntary unemployment' is something like an oxymoron.

[12] As an illustration, take claim 2 about the causal role of the generosity of unemployment benefits. The following story shows that the meaning of the causal claim can shift substantially with the meaning of U. It is plausible that increasing the generosity of unemployment benefits incites more individuals to work in underground markets while they are pretending to be actively searching for a job. In this hypothetical situation, it could be true that more generous benefits increase the *official* unemployment rate, but be false that benefits increase the *overall* unemployment rate (i.e. classifying informal workers as employed).

[13] The same arguably holds for variables capturing 'short-time work schemes' and 'economic gains' (see claims 3 and 4 on p. 26).

an extreme oversimplification since the two first dimensions are in fact multidimensional as well. The semantic complexity of the first dimension, namely, the level of benefits B_l—expressed by the replacement rate (the ratio of unemployment benefits to previous employment earnings)—is highlighted in the following comment by an OECD economist:

> There is no such thing as *the* replacement rate in any OECD country, rather there are a myriad of replacement rates corresponding to the specific personal and family characteristics of the unemployed, their previous history of work and unemployment, and the different structures and entitlements of unemployment insurance (UI) and social assistance (SA) systems in OECD countries and the ways in which these systems interact with tax systems. Once one tries to grapple with these complexities in order to compute replacement rates for the purpose of international comparisons, the task becomes a daunting one. (Martin, 1996, p. 100)

The utterance 'the generosity of benefits increases' is consequently susceptible to a variety of interpretations.[14] To be sure, there are cases when a change in the unemployment benefit system can be unproblematically interpreted. For instance, if the replacement rate for a subset of individuals is increased and the rest of the system stays the same, this change will certainly count for an increase in generosity. Still, for the more complicated cases,[15] there is the need to (implicitly) rely on a transformation of the multidimensional variable into a unidimensional scale if one wants to talk about lower or higher generosity. For the purposes of the *OECD Jobs Study*, a "summary measure of benefit entitlements" has indeed been constructed. Its explicit goal was "to capture the degree of 'generosity' of a country's benefit system" (OECD, 1994a, part II, p. 172). Even though the construction of this measure is an impressive achievement—for each country, it averages the replacement rates across

[14] There is nothing peculiar about unemployment benefits in this respect. If instead one considers the purported cause 'short-time work schemes' appearing in claim 3 above, one finds that recent discussions about short-time work schemes actually decompose them in 14 dimensions (which are then regrouped into four main families of features; see OECD, 2010, Annex 1.A1). When it comes to 'flexibility' (i.e. the posited cause in claim 1), we are not even aware of any serious attempt to spell out the different dimensions of the concept.

[15] These harder cases occur only if the comparison is between two realizations of the multidimensional variable V, say v and v', such that for some dimension i, $v_i > v'_i$, and for another dimension j, $v_j < v'_j$.

18 distinct personal situations—the OECD is of the opinion that it is only "a very approximate indicator" of actual generosity, and discusses many instances where the measure would actually fail to register a change in generosity while intuitions would go in the opposite direction (OECD, 1994a, part II, p. 173-6). In consequence, this summary measure does not solve the issue of identifying B's referent.

In sum, the first question in our referentialist procedure uncovers the semantic complexity stemming from the multiple potential specifications of the relevant categories involved (e.g. 'active' or 'level of benefits') and from the multidimensionality of some concepts (e.g. 'generosity of benefits'). Since a referentialist approach to meaning conceives of the meaning of a sentence—here a causal generalization—as being determined by the meanings of its elements, different interpretations of the causal relata have direct implications for the answer to the general question 'what does this causal generalization mean?'

1.2.2 The relevant units

The second step in our referentialist analysis is to identify the units that are referred to. This is a necessary step because generalizations (at least in the social sciences) are not *universal* generalizations.[16] In uttering '$X \hookrightarrow Y$', one is typically not endorsing the claim that in *all* systems where the variable X can meaningfully be said to be realized, the referent of this variable is causing the referent of Y. For instance, we believe that 'birth-control pills prevent pregnancy', but we implicitly restrict the set of units considered to non-sterile women; similarly, $B \overset{+}{\hookrightarrow} U$ cannot be true of units in which it is impossible to be (officially) jobless—think of contemporary North Korea.

The task of identifying the relevant units can be broken into two parts. First, one determines the spatiotemporal boundaries of the systems. One could, for instance, consider that the unit is a human individual from day 1 until death. Such a system seems to be implicit in many claims such as 'smoking causes lung cancer'. For the OECD's generalizations, the units are certainly not individuals, but most plausibly countries. With this first part comes some sources of semantic complexity. What are the exact *spatial* boundaries of a country—i.e. which entities and activities do

[16] This point is made, for instance, by Daniel Hausmann: "Since the causal role of variables depends on background circumstances, causal generalizations should be relativized to some population P. " (Hausman, 2010, p.50; see also Eells, 1991, p.23-40)

we consider as being in this system? What are the *temporal* boundaries of a country—i.e. do we consider this system far back in the past or far in the future as being the same unit? Nature does not impose answers to these questions; we can cut out systems in the space-time fabric as we see fit.

The second part identifies the *set* of units—i.e. which of the systems of the appropriate type (according to the first part) are included. Taking the example of contraceptive pills again, the first part would plausibly draw boundaries around human individuals, while this second part would select all and only non-sterile women among the systems isolated in part 1. This second part is also a source of semantic complexity for the OECD's generalizations. Do we need to include countries like Mexico, which became a member of the OECD in 1994? Do we include countries that are not part of the OECD (such as any African country) or became part of it only recently (like Slovenia in 2010)? Whether we answer these questions affirmatively or not changes the meaning of the causal generalizations.

In sum, this step in the referentialist procedure renders explicit the population $P = \{u_1, u_2, ...\}$ by first determining which types of units are the u_i's, and then which subset of these units are actually part of P. There are multiple ways to proceed that we don't need to explore here—e.g. extensional versus intensional specifications of the set of units. Our key point in this subsection is that there is some leeway in the identification of the relevant population which contributes to the semantic complexity of causal generalizations such as the ones found in the *OECD Jobs Study*.

1.2.3 Which causal relation?

The two steps discussed above have not much to do with 'causal' in 'causal generalizations'. The third step of the procedure is all about that notion. The goal of this step is to specify the relation which is referred to by the causal verb in a generalization.

Someone aware of the philosophical literature on causality might expect that this subsection will pit against each other the main contenders for a theory of causality—e.g. regularity, probabilistic, counterfactual, interventionist, and process theories. This is not what we offer.[17] The main goal of these theories has been to find necessary and sufficient conditions for a relation to be causal; they focus "on finding criteria that distinguish causal from non-causal relationships" (Woodward, 2010, p. 287).

[17] For an article analyzing the potential contributions of these different theories of causality and concluding with an *inferentialist* flavor, see Reiss (2009).

In contrast, we want to distinguish *among* causal relations. The recent literature on causality has indeed highlighted many distinctions among causal relations. A causal verb could thus refer to various relations. This possibility is the source of semantic complexity for this step.

There is a further characteristic of what we offer here which distinguishes it from a discussion of standard theories of causality. Most of these theories have aimed to offer a *reductive* analysis of causation: the goal has been to formulate criteria for causation which rely on *non-causal* notions—e.g. co-occurrences, probabilities, counterfactuals, and energy transfers. In contrast, our analysis is non-reductive: we rely on a causal notion, i.e. counterfactual manipulation, to spell out distinctions among causal relations. The use of the concept of counterfactual manipulation to this end is widespread nowadays. Woodward (2003) is its most distinguished proponent in philosophy but he is indebted to a rich literature which predates him. Woodward relies, for instance, extensively on Pearl (2009), who finds roots for his approach in structural econometrics (e.g. Haavelmo, 1944; Simon, 1957). The potential outcome framework—first developed in parallel but now integrated to the structural approach—also uses a notion of counterfactual manipulation (Holland, 1986; Morgan and Winship, 2007; Imbens and Wooldridge, 2009). This vast literature shares a non-reductive account of causation which we label the counterfactual-manipulationist account. This account provides a powerful toolbox that we propose to use in our semantic analysis of causal generalizations.

The counterfactual-manipulationist account conceives of the world as being made of *causally-structured* systems. A causal *generalization* is asserting that a causal relation holds between the referents of the relata *in a population of systems.* To analyze the meaning of a generalization, our procedure inquires first about the causal relation referred to *for a single unit.* It moves only in a second stage to the population level.

Causing: unit level

A causal generalization is about a population of units. An extreme case is when the population includes only one unit. Talking about a generalization in this case seems inappropriate; we rather wish to say that such a claim is about unit causation, it is a unit causal claim.[18] We start

[18] A claim about unit causation is not necessarily about actual causation. Actual causation is when the causing indeed occurs—e.g. the cue ball indeed hit the 8 ball. Unit causation is about causal relations, either actual or not, for a single unit. The two dimensions are orthogonal. See Hitchcock (2001a, p. 219-20) for an enlightening

our semantic analysis of the causal verb by looking at unit causal claims. Paralleling formula (1.1), the schematic form of a causal claim for unit u_i would be

$$\text{For } u_i,\ X \hookrightarrow Y \tag{1.3}$$

For the manipulationist-counterfactual account, the causal structure of a single system is what makes some causal claims true of this system. In general, a causal claim like (1.3) is true of system u_i if and only if some *ideal* manipulation of the referent of X would change the value or the probability distribution of the referent of Y. Proponents of the manipulationist-counterfactual account do not agree on the exact characterization of an ideal manipulation, but their goal for this characterization is the same: they want that the referent of X be surgically manipulated. This manipulation does not have to be actually implemented, nor does it have to be humanly feasible. The manipulation is a hypothetical contrivance to reveal what is asserted about the *asymmetric* structure of the system, asymmetric in that the referent of relata Y is claimed to depend on the referent of X but not the other way around.[19]

For unit causal claims, the semantic complexity due to the causal verb is that there is a vast panorama of ways that the referent of Y can *depend* on the referent of X. In this subsection, we do not explore this panorama in its entirety but we attempt to give an impression of its richness. We start by defining two possible interpretations of the causal relation picked out by '$\hookrightarrow$'. These two interpretations are polar extremes: we will call them minimal and maximal causations. We then focus on two directions one can take to move away from maximal causation and toward minimal causation.

Minimal causation. X minimally causes Y in u_i if and only if there are two values x_0 and x_1 of X, and there are some values z of some other variables Z which have well-defined referents in u_i, such that the probability distribution of Y conditional on $Z = z$ changes as

discussion.

[19] Note that one of the main objections to the manipulationist-counterfactual account is about the modularity requirement implicit in this use of hypothetical manipulations (Cartwright, 2007). One can doubt that it is even metaphysically possible that an ideal manipulation exists for the cause in *all* causal relations. The nature of the modularity requirement depends on the precise characterization of the *ideal* manipulation, but all versions of the manipulationist-counterfactual account require some sort of modularity requirement. Hoover (2012b) claims that a strength of his version is that the modularity requirement is weaker, perhaps even "trivial".

the referent of X is switched by an ideal manipulation from x_0 to x_1.[20]

It should be clear why these truth-conditions capture only a *minimal* dependence of the referent of Y on X. They say that, in at least one background state z, at least one manipulation of the referent of X will change the likelihood of the different values of Y.

Our two main examples—*Inflex* $\overset{+}{\hookrightarrow}$ U and $B \overset{+}{\hookrightarrow} U$—are claims about *positive* causation.[21] We thus want a definition of minimal positive causation. The definition of minimal causation just given is in probabilistic terms, but minimal *positive* causation can be expressed in a far more intuitive manner if we switch to structural equations. Indeed, a common strategy among proponents of the manipulationist-counterfactual account is to think of propositions like 'for u_1, $X \hookrightarrow Y$' as only a shorthand for expressing some characteristics of the causally-interpretable functional dependence of Y on X, which can be referred to by the structural equation

$$Y \Leftarrow f(X; Z). \tag{1.4}$$

Note that the symbol $\Leftarrow$ distinguishes this formula from a standard equation with an equality sign (Hoover, 2001). In the present case, the formula expresses that Y depends causally on X (and Z) but not the other way around. The implication is that, if we were to invert function $f(\cdot)$ to express the values of X in terms of Y and Z, the resulting function could not be given a causal interpretation. X can be represented as functionally depending on Y and Z, but only function $f(\cdot)$ can be interpreted causally.

[20] Note, first, that the wording of this definition and the ones below fits better Pearl's (2009) and Woodward's (2003) versions of the counterfactual-manipulationist account than the one of, say, Hoover (2001, 2012b).

Second, some readers might prefer a more formal characterization of minimal causation. Define V to be a vast set of variables including all variables that have values with clear referents in u_i. Obviously, $X, Y \in V$. Define $Z \subseteq V \setminus \{X, Y\}$ which explicitly allows Z to be empty. I use lowercase x, y, z to denote values of the associated variables. Define a well-behaved probability distribution $P(X, Y, Z)$. Following Pearl (2009), define a *do* operator $do(W = w)$ which is interpreted as forcing the referent of a variable W to take a specific value w by an ideal manipulation on this variable.

Truth-condition of a *minimal* causal claim. 'For u_1, $X \hookrightarrow Y$' is true as a claim about *minimal* unit causation if $\exists\, x_0, x_1$ of X, $\exists\, Z$ and $\exists\, z$ of Z, such that $P(Y|do(x_0, z)) \neq P(Y|do(x_1, z))$.

[21] This idea is explored, for instance, by Steel (2008, p. 22-23) with his "monotonic interpretation" of positive causal relevance, and by Hausman (2010, p. 48-49) with his notion of "causal role".

One might object that interpreting 'for u_1, $X \hookrightarrow Y$' as being about characteristics of $f(\cdot)$ in (1.4) could change the meaning of the proposition, and that it would be preferable to use a probabilistic framework in which it is the general probability distribution of Y (not its actual values) which depends on X. But our goal is not here to present all the potential referents of the causal verb; we want only to illustrate some of the possibilities, and this is easier in a functional framework.[22] We start, as promised, by minimal positive causation.

Minimal positive causation. X is a minimal positive cause of Y in u_i if and only if, for two values x_0 and x_1 of X with $x_0 < x_1$, and for some values z of some other variables Z, we have $f(x_0, z) < f(x_1, z)$.

This can be expressed in words: in at least one background state referred to by z, the resulting value of Y is higher when the referent of X is fixed by a manipulation to a specific value x_1 compared to the value of Y associated with a specific lower value x_0. Interpreted in this way, a causal claim says little about the functional dependence of Y on X. It would be surprising that proponents of causal generalizations such as the ones in the *OECD Jobs Study* want to say so little when they utter them.

Starting with this characterization of minimal positive causation, there are many ways to be more informative. In fact, these ways are only limited by one's imagination regarding the form of the dependence of the referents of Y on the referents of X, conditional on the background Z. We now turn to the other extreme.

Maximal positive causation. Assuming that $f(\cdot)$ is differentiable for expository purposes, X is a maximal positive cause of Y in u_i if and only if $df/dX > 0$ for all values of X.

The referentialist analysis would thus be that the magnitude of the referent of Y is always increasing as the magnitude of the referent of X is increased by an ideal manipulation. Note that this dependence is claimed to hold irrespective of the initial state of the rest of the system (which is captured by Z). While minimal causation said little about the dependence between the referents of Y and X, maximal causation says a lot.

[22] There are also reasons to always favor a functional framework in causal analysis, see Pearl (2009, p. 26-27).

We now discuss two ways to weaken the interpretation of '$\overset{+}{\hookrightarrow}$' away from maximal positive causation: restricting the changes of X that are characterized and conditioning on the background state of the system.

Firstly, it might well be that the causal claim does not assert that the positive relation holds for *all* changes in the values of X. One possibility is to rule out the causal efficacy of *marginal* changes. Take $B \overset{+}{\hookrightarrow} U$. The claim might not assert that a tiny change in the generosity of benefits has an effect (even tiny) on unemployment. For instance, one might plausibly think that some threshold must be passed for labor-market agents to see a difference in generosity and act on it. Another (and more important) possibility is that an enormous change in generosity might also not be the subject of the claim. Is the event of a dismantlement of the unemployment benefit system covered by the claim? Given that such a reform would plausibly have a few extra effects—e.g. social unrest—it might be implicitly ruled out. So which changes in the referent of X are meant by the claim? Here lies some semantic complexity.

Secondly, the interpretation in terms of total differentiation in our definition of maximal positive causation omits the background Z. It maintains that manipulating the referent of X is *sufficient* for the referent of Y to change in the stated direction. This sufficient causality which is about *directions of change* in X and Y must be contrasted to a case where the *value* of X is sufficient to determine the *value* of Y. To illustrate the difference between the two notions of sufficiency, note that if the function for Y is additively separable, X will typically fulfill the conditions to be a sufficient cause about directions of change but not about values. Being additively separable, the function will look like

$$Y \Leftarrow f(X; Z) = g(X) + h(Z). \tag{1.5}$$

It is easy to see from this equation that the value of X does not suffice to determine the value of Y since we also need values for Z; but we can typically induce a change in Y by wiggling X irrespective of Z.[23] From now on, we will use 'sufficient cause' to refer to sufficiency for directions of change, not value.

A claim about sufficient causation is quite a strong one to make. It implies that if policy makers were able to actually implement the ideal

[23] One must add the condition that Z is not, in turn, an effect of X such that wiggling of X both affects Y directly (through $g(X)$) and indirectly (through $h(Z)$). Furthermore, causal sufficiency in the sense use here does not necessitate that the function is additively separable. For instance, the function could be $f(X; Z) = XZ$ and X would still be a sufficient, positive cause of Y if Z could only take strictly positive values.

manipulation (which is admittedly doubtful), they could change the magnitude of Y in the stated direction (in comparison to a state where X is left untouched) without having to worry about any other factor in the system. For the case of $B \overset{+}{\hookrightarrow} U$, there are reasons to think that the OECD nevertheless means sufficient causation. The main reason is that one source of evidence for the claim is *linear* cross-country regressions, which indeed assumes (for unbiasness) that the causal structure is such that its appropriate functional representation is additively separable.

There are also reasons to think that the OECD does *not* mean sufficient causation, but something weaker like an INUS condition (an insufficient but non-redundant part of an unnecessary but sufficient causal set; Mackie, 1974). An increase in the referent of B is an INUS condition for an increase in the referent of U only if the former increase results in the latter in *some*, but not all, background circumstances, i.e. in states of the system corresponding to a strict subset of the potential values of Z.[24] One reason why this reading is plausible is that the OECD does not hesitate to exclude some countries in its (linear) regression analysis. Political instability is one justification for this exclusion according to the OECD (1994a, part II, p. 178). A plausible interpretation of this move is that the OECD excludes countries that have values for Z which are implicitly ruled out by the claim. B would thus be causally effective for U only when the causal system is in some appropriate state. Another reason in favor of the INUS reading is the explicit recognition by the OECD that "institutional factors influence whether or with what lag new benefit entitlements affect unemployment." (OECD, 1994a, part II, pp. 211) The OECD thus seems to believe that some institutional factors (referred to by Z) can affect *whether* the referent of B is indeed causally related to the referent of U. So, is the unit causal claim 'for u_i, $B \overset{+}{\hookrightarrow} U$' an assertion about sufficient or INUS causation? If it is the latter, what are the additional contributing factors that are not explicitly present in the statement?

In sum, the causal verb in a *unit* causal claim is a source of semantic complexity. We put forward two extreme interpretations of this verb in our examples about positive causation: minimal and maximal positive

[24] As an illustration take the structural equation $Y \Leftarrow XZ_1 + Z_2$, where variables X, Z_1, Z_2 can only take values 0 or 1. X is not sufficient for the causal effect $\Delta Y = 1$ since switching X from 0 to 1 will fail to affect Y when $Z_1 = 0$. Not sticking to $X = 0$ given that one manipulates Z_1 and not Z_2 is however sufficient and necessary for $\Delta Y = 1$. Furthermore, $\Delta Y = 1$ can occur through a change of Z_2 irrespective of the values of X and Z_1—these last two are not necessary.

causation. Since neither of these extremes is much plausible, two directions to weaken the interpretation of the causal verb have been discussed: which changes of X are to be considered and which, if any, background factors have to be in an appropriate state. Note that these two dimensions are not the only sources of semantic complexity at this stage; we gloss over other sources.[25]

Causing: population level

Now we want to consider a claim about a population including more than one unit, i.e. a proper causal *generalization*. Our procedure suggests at this step to analyze a population-level assertion in terms of the sets of unit causal claims into which it can be translated. This strategy of analysis is why we made an apparent detour by unit causal claims. Now, we want to say that the population-level statement should be interpreted as a disjunction among sets of unit causal claims. If this suggested strategy does not sound intuitive, we hope that its relevance will become clear as we proceed with our discussion of two sources of semantic complexity: the possibility of heterogeneity and of interaction effects.

Possible heterogeneity. The first interpretation of a causal generalization 'for P, $X \hookrightarrow Y$' that probably comes to mind assumes causal homogeneity. Under this interpretation, accepting the generalization commits one to asserting the existence in every unit u_i of the causal relation denoted by $\hookrightarrow$. In other words, saying that a causal relation exists for a population is saying nothing else than that the same causal relation exists for each unit. Causal homogeneity thus maintains the following equivalence between population-level and unit-level claims:

$$\text{`For } P,\ X \hookrightarrow Y\text{'} \ \equiv\ \{\text{For } u_i,\ X \hookrightarrow Y\}_{\forall i \in \{1,\ldots,n\}} \tag{1.6}$$

For our example about unemployment benefits, this interpretation implies that all countries in the implicit population P are asserted to share the same causal relation—i.e. the referent of U in each country depends in the same way on the referent of B.

[25] Another interesting source of semantic complexity is considerations about the timing of the changes. The last quotation by the OECD talks indeed about a "lag" in the production of the effect. How long and varying is the implicit lag? Yet another source is the distinction between component and net effects (Hitchcock, 2001b; Woodward, 2003, p. 50).

If this identity of causal relations is relaxed, we have multiple options for the meaning of causal heterogeneity. For instance, one might read the claim as a 'in most cases' generalization: in most units of P, the same relation denoted by $\hookrightarrow$ is claimed to exist. But the most popular interpretation of heterogeneity is the average-effect interpretation, which is central to one variant of the manipulationist-counterfactual account, i.e. the potential outcome framework (Holland, 1986).

Since the average-effect interpretation is affected by the *magnitudes* of the causal effects for each unit, we need a bit more notation. Keeping with our example of unemployment benefits, imagine that the claim implicitly considers only two levels of generosity b_i^{low} and b_i^{high}. In the spirit of the potential outcome framework, define the associated values of unemployment u_i^{low} and u_i^{high}, which are interpreted as denoting the unemployment rate in country i when benefits are (ideally) manipulated to be low or high. For i, the causal effect on U of moving from low to high benefits is denoted by $\Delta u_i = u_i^{\text{high}} - u_i^{\text{low}}$. With this additional notation, we can compactly expressed different interpretations of the population-level claim 'for P, $B \overset{+}{\hookrightarrow} U$'. The three that we have introduced are:[26]

Causal homogeneous. $\Delta u_i > 0$ for all $i \in \{1, ..., n\}$

Most cases (median). $\#(i \in P | \Delta u_i > 0) > n/2$

Average effect. $\frac{1}{n}\sum_{i=1}^{n} \Delta u_i > 0$

Note that the average-effect interpretation establishes a peculiar relation between the generalization and unit causal claims. Strictly speaking, the generalization does not necessarily pick out the direction of the effect in *the majority* of units u_i, since a skewed distribution of the unit-level causal effects could make the median and the average interpretations diverge.

Possible interaction effects. While our brief discussion using u_i^{low} and u_i^{high} was meant to reveal one source of semantic complexity—i.e. the distinction between causal homogeneity and various causal heterogeneities—it obscured another source. In defining these two states of system i, we indeed relied implicitly on a major simplification that Donald Rubin (1980, 1986) labeled the 'stable-unit-treatment-value assumption' or SUTVA:

[26] There are other possibilities including a Pareto-dominance interpretation, see Hitchcock (2001a).

> SUTVA is simply the a priori assumption that the value of Y for unit i when exposed to treatment x will be the same no matter what mechanism is used to assign treatment x to unit i and no matter what treatments the other units receive. (Rubin, 1986, p. 961, notation slightly changed)

As is clear from the quotation, SUTVA has in fact two parts.[27] The first part is about the invariance of the unit-level causal effect to the *actual mechanism assigning* the treatment. We will not discuss this issue here.[28] We focus on the second part which is about the invariance of the causal effect for a given unit to what values of the cause the other units happen to have.

When the second part of SUTVA fails, one says that we are in presence of interaction effects.[29] The point can easily be illustrated by the following example outside labor economics. Take the causal generalization 'free trade causes economic gains'. This claim is semantically complex in numerous respects but let us focus on the following question: What if country i is the only one to adopt free-trade policies? That is, all the other countries keep high tariffs and other trade barriers while this single country decides to revoke all its barriers. It is plausible to say that 'economic gains' for i are less likely to be caused by this isolated policy than if the other countries were also engaged in trade liberalization. We could thus amend the generalization to something roughly like: 'free trade causes economic gains *conditional* on the other countries also liberalizing'.

What this hand-waving example shows (regardless of its empirical credibility) is that, in presence of interaction effects, there is much work needed to detail which counterfactuals lie behind a given causal generalization.[30] In such cases, the causal relation in each system depends

[27] Note that many authors (e.g. Morgan and Winship, 2007, sec. 2.4; Imbens and Wooldridge, 2009, sec. 2.3) —though not all (e.g., Heckman, 2005, p. 35fn)— reduce SUTVA to the condition about the stability of treatment effect to the treatment assignment of other units.

[28] The notion of ideal manipulation that we use (from the manipulationist-counterfactual account), in being ideal, rules out this variability in the treatment assignment mechanism. As argued in Reiss (2007, ch. 10) and Cartwright (2007, ch. 16), this is at the cost of making the causal claims less directly relevant to policy. Indeed, even if one grants a causal claim, it can still be argued that the effect will not follow from a real-world implementation of the cause simply because this manipulation fails to be of the ideal type.

[29] This class of effects are often labeled 'interaction and general-equilibrium effects'. We write interaction effects for short.

[30] Note that Rubin (1986) seems to be of the opinion that SUTVA is a necessary condition for a causal claim to be meaningful.

on the state of the other systems. The relevant unit causal claims must consequently include information about the state of other systems. One interesting contrast is then between:

'Isolated reform' counterfactual: the causal effect on Y_i of manipulating X_i when the X_j's of all the other units $j \neq i$ are not being manipulated.

'Generalized reform' counterfactual: the causal effect on Y_i of manipulating X_i when the X_j's of many or all other units are also being manipulated.

When interaction effects are suspected—such as in the free trade case—choosing between these two interpretations of the counterfactual will matter a great deal to the meaning of the causal generalization.

In sum, semantic complexity takes different forms at the step of analyzing the causal verb. Our key distinction is between semantic complexity at the unit level—the range of changes in X considered and the potential dependence on background factors—and at the population level—the possibility of heterogeneity and interaction effects.

1.2.4 Truth-conditions of the generalizations

Now that we have given a referentialist procedure to analyze the meaning of all the parts in a causal generalization of the form '(For P,) $X \hookrightarrow Y$', the meaning of the whole causal generalization is only a small step away. Remember that, for a referentialist semantics, the meaning of a sentence is its truth-conditions, i.e. what the world must be like for the sentence to be true. We can identify the truth-conditions of a generalization in the following way. '(For P,) $X \hookrightarrow Y$' is true if and only if all the unit causal claims in one of the sets into which the generalization is translatable are true. In turn, each unit causal claim is true if and only if the referents of X and Y in this system are indeed related by the causal relation to which $\hookrightarrow$ refers.

The procedure shows that causal generalizations are semantically complex: although it might not be apparent from their surface structure, these statements have a wide variety of potential meanings. Here is a summary of the sources of semantic complexity:

a) There are many possible ways to translate a generalization in terms of sets of unit causal claims.

b) The choice of the units considered matters.

c) There are many potential referents for common social-scientific variables.

d) There are many possible causal relations between the referents of variables.

It should be obvious that different choices on these four dimensions can lead to the same utterance having different truth-*values* for our world—the utterance might be true under some interpretations, but false under others.

1.3 A challenge for the referentialist procedure

The referentialist procedure of the previous section leads one to the conclusion that there is a tremendous variety of potential interpretations of seemingly simple causal generalizations such as ‘more generous unemployment benefits cause higher unemployment’. The fact that our referentialist procedure allows us to see this menu of potential meanings is a *prima facie* advantage of the procedure. One might argue that the procedure allows us to access properties of a sentence that were initially hidden. We think, however, that the semantic complexity identified by our referentialist procedure should instead cast doubt on the procedure itself. We argue here that the most plausible interpretation of the semantic-complexity conclusion is that the referentialist procedure looks at the wrong place for the meaning of (our type of) causal generalizations.

The identification of semantic complexity raises a question. Which, among the menu of *potential* meanings identified, is the *actual* meaning of a given generalization? There are two options: either the generalization has one specific meaning or it is ambiguous, i.e. the sentence equivocates among various meanings. We discuss each option in turn. Our general line of argument is that, under both of these options, there is no plausible way to *rationalize* the widespread practice of demanding and supplying causal generalizations in policy-oriented social science. We thus end up with a dilemma: either this widespread practice is unreasonable or the referentialist procedure misses the meaning of (our type of) causal generalizations.

The first option is that generalizations such as '$B \overset{+}{\hookrightarrow} U$' and '*Inflex* $\overset{+}{\hookrightarrow}$ U' pick out one meaning from the menu generated by our referentialist procedure. The puzzle now is why the ones formulating or transmitting a claim are not more explicit about its actual meaning. Indeed, one finds these claims all over policy-oriented reports such as the *OECD Jobs Study*, usually with little additional information that would count as clarifying their meanings from the perspective of a referentialist semantics. This lack of explicit clarification is puzzling because our referentialist procedure identifies so many potential meanings that the transmitters of a claim might well fear that it will be misinterpreted on the receiver's side. We can think of many tentative explanations of this puzzle, none of them plausible.

Firstly, one could try to explain the lack of explicit clarification by the existence of background information already shared by the senders and the receivers. Now, reports like the *OECD Jobs Study* are instances of communication from expert economists to policy makers (and the lay public). There is certainly *some* shared background information between these two groups but, given the quantity and nature of the information required according to our referentialist procedure, it is unlikely that this implicit knowledge will be sufficient for the receivers to get the meaning right. In other words, identifying the referentialist truth-*conditions* requires many decisions, and the information plausibly shared by the senders and receivers will not be sufficient for the receivers to follow the senders in this labyrinth. It is thus unlikely that meaning would be preserved in the transmission.

Secondly, one might argue that what needs to be preserved in the transmission are truth-*values* not truth-*conditions*; the senders want the receivers to come to believe, thanks to them, a true proposition, not a proposition with the exact same meaning as the one they believed in. It turns out that one potential meaning of a causal generalization has this property. We will use the generalization about unemployment benefits as an example. Imagine that the initial meaning of the claim has the following properties:

a) The appropriate translation in terms of unit causal claims is that the causal relation holds for each unit (homogeneity) with no interaction effects.

b) All countries (back then, now and in the future) are in the relevant population.

c) All plausible referents for ‘unemployment benefits’ and ‘unemployment rate’ would turn the claim true for our world.

d) The causal relation is sufficient causation (i.e. Z is empty) and spans the entire range of possible changes in the referent of B.

Call a generalization under such an interpretation a ‘maximally-sturdy generalization’. If the senders believe the maximally-sturdy generalization—i.e. if they judge it to be true of our world—there is little to fear about misinterpretation on the receiver’s side since the truth of the claim would not be affected.[31] For instance, the receiver might think that the claim is an average claim, but the average claim will be true if the homogeneous claim is; she might restrict the population, the range of values of B, etc., but the narrower claim will still be true given that the maximally-sturdy generalization is true. It hence seems reasonable to transmit the claim without further ado *provided* that it is believed as a maximally-sturdy generalization.

The problem with this interpretation is that maximally-sturdy generalizations about the social world are most certainly false. It is a well-known fact that all lawlike claims in the social sciences suffer from counterexamples when interpreted strictly—hence the literature on *ceteris paribus* laws. And our maximally-sturdy generalization is even less likely to be true than other lawlike claims. Its truth is, for instance, even invariant to changes in the referents of the relata. It is thus not *reasonable* for the senders to believe their generalizations as being maximally sturdy. A few dogmatic researchers might have such beliefs, but it is not plausible to explain the widespread practice of supplying non-clarified causal generalizations by maintaining that senders believe them to be (almost) maximally sturdy.

The third, and last, potential explanation of our limited-clarification puzzle under the assumption that generalizations have a clear-cut referentialist meaning makes it a matter of convenience. It is cumbersome to be explicit about the meaning of all our utterances, so we use shortcuts. A non-clarified generalization is such a shortcut, but one should not worry too much since the one transmitting the claim has a definite (referentialist) meaning in mind, and is ready to specify it if asked. This meaning can also be quite far from the maximally-sturdy interpretation, which is

[31] The senders might not believe exactly the maximally-sturdy generalization but something a bit less sturdy. The fear that the receiver, through misinterpretation, is led to believe a false proposition should increase as the ‘sturdiness’ of the initial proposition decreases.

nice since we can restore the reasonableness of the experts' beliefs. Taking the unemployment-benefits claim for example, they may mean that it holds only on average, for a specific understanding of 'unemployment' and 'benefits', for some narrow range of changes in the cause variable, and if some potential preventers are absent. With such a less ambitious meaning, the generalization has a better prospect of being true. The experts will simply tend to keep most of these conditions implicit because it is inconvenient to spell them out.

There is certainly something correct about the argument that convenience affects how much detail we include in our expressions. It would be, for example, surprising that experts had insisted in supplying a full referentialist clarification of their generalizations in the two-page press release which accompanied the publication of the *OECD Jobs Study*. There is, however, a major problem with this explanation: in the 500-page *OECD Jobs Study*, one does not find anything that would qualify as a referentialist analysis of the meaning of these generalizations. It thus means that experts would keep the meanings of these generalizations for themselves, even when they have ample space to spell them out. Since misinterpreted generalizations could well have as a consequence the implementation of ineffective or even harmful policies, the secretive stance of experts makes them irresponsible. If they have in mind well-specified and non-sturdy (referentialist) meanings for their generalizations, the reasonable thing to do is to spell out these meanings at least somewhere in their lengthy reports.

The other option is that experts do not have well-specified meanings for their generalizations. There is thus no limited-clarification puzzle: there is nothing that the experts keep for themselves. This option is delivering Woodward's diagnosis as cited in the introduction to this chapter: these generalizations are "confused, unclear and ambiguous" (Woodward, 2003, p. 7). In uttering them, experts are equivocating (deliberately or not) among many potential meanings. Note that *some* equivocation is certainly always present in science, but *some* equivocation will not be enough to account for the present examples, we would be in a situation with serious (not *mild*) equivocation.

There is a new puzzle coming with this option: if equivocation is a serious issue, why are causal generalizations so much in demand? Why are they so much valued? Remember that we started this chapter by noting that policy makers are fond of causal generalization, and that the OECD paraded its 'inflexibility causes unemployment' as the main achievement of the *Jobs Study*. Since the referentialist diagnostic is that these generalizations are ambiguous, why are all these people content with this loose talk? In particular, if at least some policy makers care about putting in place *effective* policies, it seems that they would be better off being informed about the actual evidence—e.g., the cross-country correlations, the other countries experiences, and so on. What would be the extra value of ambiguous causal generalizations? Call this the 'high-value puzzle'.

We conclude that if the referentialist procedure used in the previous section captures the potential meanings of causal generalizations, the whole business of demanding and supplying causal generalizations for policy purposes is unreasonable. If one adheres to the view that these generalizations are fundamentally ambiguous, prescriptions to practitioners are not far away. Following Woodward, the referentialist procedure might be used to make "*recommendations* about what one ought to mean by various causal and explanatory claims" (Woodward, 2003, p. 7). These recommendations would be that, at last, practitioners utter generalizations which have a well-specified referentialist meaning.

As we mentioned in the introduction to this chapter, one should be careful in requesting reforms since the recommendations might be an artifact of a distorting diagnostic tool. Indeed, it seems to us that the limited-clarification and the high-value puzzles should cast more doubt on our referentialist procedure than on the widespread practice of demanding and supplying causal generalizations. One strong reason why the blame should fall on the referentialist procedure is that there is an alternative semantic approach, an *inferentialist* approach, from the perspective of which one can make sense of this widespread practice of formulating generalizations which do not necessarily have a definite meaning from the point of view of a referentialist semantics. The next section is an exploration of this alternative semantics.

1.4 On the meaning of causal generalizations: an inferentialist approach

Referentialism as an approach to meaning has dominated Western thought. Its dominance is however disputed. Ludwig Wittgenstein, for instance, challenged the referentialist approach to meaning by maintaining that:

> For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language. (Wittgenstein, [1953] 2001, § 43; see also Wittgenstein, 1958, p. 69)

Following Wittgenstein's lead, an alternative semantic approach emerged, which prioritizes *inferential* relations over referential relations. While a referentialist semantics is exclusively focused on the relation between a term and its worldly referent, an inferentialist semantics centers on the inferential connections of an expression to other expressions. The basic idea shared by all variants of the inferentialist approach to semantics is that "the meaning or propositional content of an expression or attitude is determined by the role it plays in a person's language or in her cognition" (Whiting, 2009).[32]

According to inferentialism, the error of a referentialist semantics is that it takes meaning to be only constituted by word-world relations—i.e. referential relations—while meaning is primarily constituted by word-word relations, i.e. by intralinguistic relations. The inferentialist thesis does not entail that referential relations do not participate to the meaning of expressions; after all, a referential relation can always be re-expressed in terms of inferences from the assertion being analyzed to other assertions, i.e. assertions about the worldly existence of some objects, properties, and relations. The inferentialist thesis is however a direct criticism of the referentialist's single-minded quest for the referents of our terms. Applied to our referentialist procedure for causal generalizations (section 1.2), the inferentialist criticism is that the procedure diverts the attention away from where the actual meaning of a generalization lie: the network of inferences which has the generalization as a premise or conclusion.

Our goal in this section is not to develop a *comprehensive* inferentialist semantics. We rather limit ourselves to specifying only a few elements of such an alternative semantics.[33] The selected elements are the ones we deem necessary to our main task: offering an inferentialist analysis of our examples of causal generalizations from the *OECD Jobs Study*. Our conclusion is that this analysis rationalizes the demand and supply of causal generalizations in this case. We take this conclusion to be (non-decisive) evidence *for* an inferentialist analysis and *against* a referentialist analysis of causal generalizations in policy-oriented social science.

[32] There are now many labels and many species of inferentialism. The labels include: conceptual role semantics, inferential role semantics, functional role semantics, procedural semantics, and use theory of meaning. For discussions of the different species of inferentialism, see Block (1998), Whiting (2009), and Peregrin (2012). For an inferentialist semantics of causal claims inspiring our account, see Reiss (2011b, 2012).

[33] As noted in footnote 32, inferentialism is not monolithic. We aim to be as non-committal as possible in our characterization of the approach.

1.4.1 Elements of an inferentialist semantics

Inferentialism takes the meaning of an expression to be constituted by its inferential connections: “If one were to enumerate all the transitions an expression is involved in, one would thereby give its meaning.” (Whiting, 2009) A sentence can either play the role of premise or conclusion in an inference (Brandom, 2007, p. 654). The meaning of a sentence would be exhausted by spelling out the inferences in which it appears.[34]

The notion of inference at play here must be understood in a somewhat liberalized fashion. To start with, the relevant inferences are not limited to logically valid ones, they rather extend to what is called material inferences (Sellars, 1953; Norton, 2003; Brandom, 2007; Brigandt, 2010). An inference is materially correct in virtue of the content of the concepts figuring in its sentences. For instance, if a competent language user accepts the proposition ‘Lightning is seen now’, then she typically accepts the proposition ‘Thunder will be heard soon’, because the concepts of lightning and thunder are inferentially connected (example taken from Brandom, 2007, p. 657).[35] The connection between concepts can also be one of material incompatibility. For instance, by knowing that a figure cannot be a square and a triangle at the same time, one accepts the sentence ‘Figure A is not triangular’ when also accepting ‘Figure A is a square’. The relevant material incompatibilities are part of the meaning of a sentence.

Another liberalizing move for the notion of ‘inference’ is to allow it also for some word-world connections—i.e. ‘language entry’ and ‘language exit’ transitions (terminology slightly modified from Sellars, 1954, pp. 210-11). For instance, the meaning of the sentence ‘Lightning is seen now’ includes the type of circumstances which would make a competent user of the sentence utter it (language entry). The meaning of this same sentence also includes transitions, in the right circumstances, from sentence to act

[34] Spelling out all these inferences is usually an impossible task, but this is no objection to inferentialism because analyzing the *full* meaning of a sentence is not something inferentialists set for themselves. Inferentialism maintains that meaning is constituted by a set of inferential connections; this could well be true even though we cannot, due to practical limitations, enumerate all the elements of the set. If inferentialists are correct, it however implies that the referentialist quest is misguided *even though* one can enumerate all the *potential* referentialist meanings of a sentence (like we tried to do in section 1.2). The problem that we identified in section 1.3 is that meaning does not seem to lie where referentialism looks for it.

[35] The acceptance of ‘Thunder will be heard soon’ is not necessary. For instance, if the lightning is ‘heat lighting’ and the language user has been experiencing this phenomenon, then the inference will not be made.

(language exit). The meaning of ‘Lightning is seen now’ includes, for instance, the act by our competent language user of reaching the lake shore and getting off her canoe. These connections “can be understood to be *inferential* in a broad sense, even when the items connected are not themselves sentential” (Brandom, 2007, p. 658).

Among all the types of sentences we utter, causal generalizations are among the types of sentences for which an inferentialist semantics should be *prima facie* most plausible. This initial plausibility comes from widespread beliefs about ‘causation’ and about ‘generalizations’. We discuss each concept in turn.

In philosophical inquiries about the nature of causation, it is widely accepted “that causation and inference are intimately related” (Reiss, 2011b, p. 914). This view can be found in Hume who famously argued that “the necessity or power, which unites causes and effects, lies in the determination of the mind to pass from the one to the other.” (Hume, [1739] 1975, § 1.3.14.23)[36] More recently, ‘A causes B’ has been inferentially associated with “effective strategies” (Cartwright, 1979). In general, causal claims are widely believed to be inferentially connected to claims about explanation, prediction and outcome of intervention. The further step taken here (as in Reiss, 2011b, 2012) is to maintain that the inferential web is all there is to the meaning of (some) causal claims.

Inferentialism combines also nicely with two recent lines of research on the nature of some *generalizations.* The first line of research stems from the literature on *ceteris paribus* laws that we mentioned in our introduction to this chapter. To dissolve the ‘either-falsity-or-triviality dilemma’ coming from interpreting generalizations as *ceteris paribus* law statements (Lange, 1993, p. 235; Reutlinger et al., 2011, sec. 4), some authors have recently proposed that we should focus on the roles of these statements in our cognitive practices. To be sure, we are not aware of scholars explicitly combining this proposition with an openly inferentialist semantics (the closest one gets to that is Earman et al., 2002, sec. 4). Still, focusing on the cognitive roles of statements has an undeniable inferentialist flavor. Marc Lange (2000), for instance, maintains that laws are inferential rules, and that believing law statements involves believing the *reliability* of these inferential rules.[37]

The second line of research stems from the literature on generics. Generics are statements such as ‘tigers are striped’ and ‘cars have radios’,

[36] For a fully inferentialist interpretation of Hume’s view on causation, see Beebee (2007).

[37] Sandra Mitchell (2009, p. 50-56) defends a similar approach.

which obviously allow for exceptions. Sarah-Jane Leslie argues both that these sentences defy all semantic analyses of a referentialist type, and that “our understanding of generics reflects our default mechanism of generalization” (Leslie, 2008, p. 44).[38] While Leslie does not make this step, it seems promising to offer an inferentialist semantics of generics. After all, the picture emerging from her work is that generics are the linguistic manifestations of our most basic information-gathering mechanism to orient ourselves in the world. While the truth-conditions of generic statements are hard to specify, humans are found to be highly competent in inferring *to* and *from* them. Are causal generalizations not only a special type of generics?[39]

Because of its relation to ideas about causation and generalization, our thesis that the meaning of a causal generalization is given by its inferential connections has some initial plausibility. Since the proof of the pudding is in the eating, our next section offers an inferentialist analysis of causal generalizations in the *OECD Jobs Study.* Though not ‘proven’, we consider that our thesis comes out of this analysis reinforced.

1.4.2 *OECD Jobs Study*: An inferentialist analysis

In this subsection, we attempt to give a faithful, though not comprehensive, rendering of the inferential network of some of the generalizations found in the *OECD Jobs Study.* Note that our analysis is not attempting to justify the OECD’s claims, but rather to make explicit their meanings. In other words, we spell out the inferential connections of the claims, yet it could well be the case that the OECD is not justified to make these connections. We will come back to the idea of justification in the next subsection.

Though not comprehensive, we want our analysis to be systematic. We thus begin by presenting the five types of inferential connections that we uncover:

Principal-cause incompatibility: To accept one principal-cause generalization implies rejecting other principal-cause generalizations.

Wide-narrow reinforcement: A wide generalization is connected by a

[38] Before Leslie, there were also suggestions for a semantic analysis of generics in terms of nonmonotonic inferences (see Krifka et al., 1995, pp. 58-63).

[39] The literatures on *ceteris paribus* laws and on generics ran in parallel to each other. Bernhard Nickel (2010) and David Liebesman (2011, sec. 5) offer recent attempts to connect the two.

link of mutual support to narrower generalizations that can be read as concretizations of it—i.e. accepting one reinforces the commitment to the other.

Evidential-base connection: A generalization is connected to sentences that would constitute evidence for it. The sentences in the evidential base are what 'the data are expected to say' when the generalization is accepted.

Policy implication: The generalization singles out (types of) actions that should be envisaged in priority; it is thus connected to sentences about policy recommendations.

Research implication: The generalization highlights the kind of research worthy of being pursued; it is thus connected to sentences about recommendations for future research.

We will put more flesh around this bare typology as we go through our case study.

By developing our typology, we contribute to enrich Julian Reiss' (2012) inferentialist analysis of causal claims. Reiss distinguishes between two types of sentences connected to a causal claim: the sentences in its inferential base and its inferential target. The inferential base for a causal claim is "given by sentences constituting or describing the evidence for it" (Reiss, 2012, p. 2). Our evidential-base connections are exactly linking a generalization to what Reiss calls its inferential base.[40] The base for a

[40] Note two minor differences between our evidential-base connections and Reiss' links between the causal claim and its inferential base.

First, our characterization of the evidential base uses the subjunctive, i.e. what *would* count as evidence for the claim. Reiss uses the indicative. This difference implies that, while we would say that a sentence about evidence that has not been gathered yet can be part of the meaning of a causal claim, Reiss seems to reject this possibility. For instance, in accepting the claim '(For P,) $X \overset{+}{\hookrightarrow} Y$', one might also endorse the sentence that 'if we were to compute the correlation coefficient between X and Y in population P, we would get a positive value'. This sentence about potential evidence would be part of the meaning of the causal claim according to us.

Second, we want to separate connections between two causal claims from connections between a causal claim and its evidence. Since Reiss does not have our two first categories—principal-cause incompatibility and wide-narrow reinforcement—he might want to include causal claims in its inferential base. Indeed, under some plausible understanding of evidence, a statement rejecting a principal-cause claim or endorsing a narrower claim would count as evidence for a principal-cause generalization. In contrast, our evidential base is meant to include only non-causal statements.

causal claim is closer to language entry than the causal claim itself since it comprises descriptive statements about data.

The inferential target for Reiss is made of the "[s]entences relating to the cognitive, evaluative and practical content of a causal claim" (p. 3), which are typically explanatory sentences and sentences about predictions and policy interventions. The two last categories of connections in our typology are connections to this inferential target. Connections to policy recommendations are already made explicit by Reiss, and we only elevate them to the status of a separate category of connections. Research implications were not identified as constitutive of the meaning of a causal claim by Reiss; this is something we add. Both policy and research recommendations are closer to language exit than the causal claims; they are a link in the chain from the acceptance of some causal claims to action.

Finally, the connections to other principal-cause generalizations and to narrower generalizations—i.e. our first two categories—are our additions. We believe that these two types of connections are crucial both to identify the meaning of causal generalizations in policy-oriented social sciences, and to see why the practice of demanding and supplying causal generalizations is reasonable. The point about rationalizing the practice should becoming clearer as we develop our case study here, and we will return to this point in the next subsection.

We now proceed to analyze, with an inferentialist approach, the meaning of some causal generalizations in the *OECD Jobs Study*. In reading our analysis, some readers might benefit from figure 1.1 at the end of this subsection, which is a graphical representation of the inferential network that we slowly uncover. We start by reproducing (from our introduction to this chapter) what the Secretary-General of the OECD presented as the "central finding of the *Jobs Study*" (OECD, 1995, p. 3):

> [I]t is an inability of OECD economies and societies to adapt rapidly and innovatively to a world of rapid structural change that is the principal cause of high and persistent unemployment. (OECD, 1994a, part I, p. vii)

We came to represent this claim compactly by $\mathit{Inflex} \overset{+}{\hookrightarrow} U$. From the standpoint of a *referentialist* approach, this claim promised to have a vast array of potential meanings. In particular, it is doubtful that one could pin down the referent for the notion of 'ability for rapid adaptation' (or 'inflexibility'). In consequence, many would be tempted to take this claim as loose talk, and to attribute little importance to it. The prospects are quite different when one adopts an inferentialist semantics. In support of the Secretary-General's assertion, this causal generalization can be shown to be central to the whole inferential network of the OECD's study. We proceed to show this centrality of $\mathit{Inflex} \overset{+}{\hookrightarrow} U$.

In accepting the claim that low flexibility potential is the *principal* cause of unemployment, OECD economists rejected other claims because of what we characterized as principal-cause incompatibilities. Three were explicitly rejected in the report (OECD, 1994b, p. 27):

a) "Technology causes rising unemployment" ($\mathit{Tech} \overset{+}{\hookrightarrow} U$)

b) "Imports from low-wage countries cause higher unemployment..." ($I_{l\text{-}w} \overset{+}{\hookrightarrow} U$)

c) "The intensity of competition is to blame" ($\mathit{Comp} \overset{+}{\hookrightarrow} U$)

What do these three principal-cause generalizations mean—i.e. what are their own inferential connections beyond their incompatibility with $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ (and with each other)?

First, part of the meaning of these rejected generalizations is constituted by their evidential-base connections. For instance, individuals putting forward the first claim about technological unemployment had predicted numerous times in the past a permanent increase in unemployment. Such a historical upward trend in unemployment is evidentially connected to the generalization. The OECD believes that no such trend is visible in the data, a belief which coheres with its rejection of $\mathit{Tech} \overset{+}{\hookrightarrow} U$ (OECD, 1994a, part I, p. 124). Similarly, the OECD maintains that there is a tension between accepting the second alternative claim about low-wage countries and recognizing that imports from these countries count for only a tiny share of overall expenditures in OECD countries (OECD, 1994b, p. 28).[41]

Remember that the inferential connections constituting the meaning of a claim go well beyond deductively valid inferences. This should be clear for evidential-base connections, since the sentences inferentially linked to

[41] The argument against the last claim is less clearly evidential. After all, if the problem is "an inability of OECD economies and societies to adapt rapidly and innovatively to a world of rapid structural change" (OECD, 1994a, part I, p. vii), the cause seems to be both too low adjustment potential and *too high pace of change*. One might interpret this pace of change as the intensity of competition (or at least see the two concepts as closely related), and conclude that the intensity of competition is indeed the principal cause of high unemployment. The semantic difference between this third claim and $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ is thus less major when it comes to 'evidential-base connections', but it is major when one considers policy implications.

the causal claim are about what is expected given the causal claim, not what is deductively entailed by the claim. For instance, it is not a logical contradiction to jointly hold the claim $Tech \overset{+}{\hookrightarrow} U$ and the standard economic history of the last 200 years—which can be summarized by 'rapid technological improvements with no long-term trend in unemployment'. It is indeed easy to cook up a story for why the failure of a prediction which seems to follow from the generalization does not make the generalization untenable. In other terms, evidential-base connections are more about what one would expect the data to say, than what they must say.

Second, policy implications also contribute to the meaning of our alternative causal generalizations. These connections are hard to miss in the *OECD Jobs Study* because the three alternative generalizations are listed with their associated policy orientations (OECD, 1994b, p. 27):

a)* "This view holds that the pace of technological change should be slowed"

b)* "Proponents of this view [...] support protectionism to curb what they see as social dumping"

c)* "The response would be to reduce the intensity of competition"

Rejecting the three alternative causal generalizations and accepting the claim $Inflex \overset{+}{\hookrightarrow} U$ is thus also connected to the rejection of these policy orientations. For instance, the OECD judges that reducing the intensity of competition (i.e., the third policy orientation) would be detrimental because it would "cut off economies from the forces that have always been the mainsprings of economic growth and betterment." (OECD, 1994b, p. 29)

Let us recenter our discussion on $Inflex \overset{+}{\hookrightarrow} U$. The idea of flexibility is prevalent in economics. In thinking about the labor market, textbook economics is much about the notion of wage flexibility—i.e., the capacity of the price of labor to adjust to changes in demand or supply. For a student of economics, the idea of an inflexible labor market would immediately bring to mind some factors preventing the wage to adjust to its equilibrium value. The point here is that a lot of connections to abstract notions from what is known as 'economic theory' contribute to the meaning of $Inflex \overset{+}{\hookrightarrow} U$. Beyond the simple story about wage flexibility, a generalized notion of inflexibility connects to all the potential factors preventing *structural* unemployment to be as low as it could be—where "[s]tructural unemployment may be defined as that part of unemployment which is not reversed by subsequent economic upturn" (OECD, 1994a, part I, p. 66). Note that structural unemployment is also an abstract notion, which has no straightforward referent.

Although all these abstract notions have no straightforward referents, they are nevertheless allowing a competent language user to determine evidential-base connections for $\mathit{Inflex} \overset{+}{\hookrightarrow} U$. To start with, accepting this claim makes one expect that different empirical proxies for the abstract notion of structural unemployment would be raising through time—something the OECD does not fail to report (OECD, 1994a, part I, p. 67-8). In addition, one would expect to detect modifications in the structure of labor markets that somewhat predate the rise in these empirical proxies of structural unemployment.

One part of the *OECD Jobs Study*, entitled "The Adjustment Potential of the Labour Market" (OECD, 1994a, part II), is exactly trying to detect these modifications. It covers a wide range of issues including "government-imposed barriers to greater aggregate and relative wage flexibility" (p.52; e.g. minimum wage), geographic mobility, employment protection, training, unemployment benefits, and taxation. The meaning of 'flexibility' thus includes connections to these more specific concepts. 'Flexibility' is a structuring concept in that it allows one to inferentially articulate a host of labor-market dimensions as all being about 'more or less flexibility'.

This articulation of labor-market dimensions around the concept of flexibility allows the inferential connections that we previously labeled 'wide-narrow reinforcement'. The wide $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ is indeed inferentially connected to narrower generalizations. The OECD is explicit about the structuring role of its wide generalization in the quest for narrower claims:

> [T]he main thrust of the study was directed towards identifying the institutions, rules and regulations, and practices and policies which have weakened the capacity of OECD countries to adapt and to innovate, and to search for appropriate policy responses in all these areas. (OECD, 1994a, part I, p. vii)

Our second example of causal claim—i.e. more generous unemployment benefits cause higher unemployment or $B \overset{+}{\hookrightarrow} U$—comes at this point in the inferential network. It is the inflexibility claim transposed 'one level down' in that it focuses on one institution among others which play a role, according to the OECD, in the capacity to adapt of economies.

$\mathit{Inflex} \overset{+}{\hookrightarrow} U$ and $B \overset{+}{\hookrightarrow} U$ support each other. In one direction, the inflexibility claim contributes to the plausibility of the benefits claim because the latter is seen as a concretization of a more general lesson that one is endorsing. In the other direction, $B \overset{+}{\hookrightarrow} U$ is also transferring some plausibility to the more general $\mathit{Inflex} \overset{+}{\hookrightarrow} U$. This happens because the benefits claim has its own evidential-base connections. In particular, one would expect that countries with more generous benefits are also the ones with higher unemployment, and similarly that a country changing importantly the generosity of its benefits experiences in later years a change in unemployment in the right direction. The OECD interpreted the outcome of its empirical research and its survey of the existing literature as being roughly in line with these expectations: cross-country regressions give the expected sign of the key parameter for some specifications, and historical narratives for selected countries are compatible with the belief in a positive effect of benefits on unemployment (but with a long, and hard to predict, time lag; see OECD, 1994a, part 2, chap. 8).[42]

The mutual reinforcement that we identify between $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ and $B \overset{+}{\hookrightarrow} U$ also holds between the inflexibility claim and the other narrower generalizations that we do not discuss explicitly here—e.g. about minimum wage, hour flexibility, employment protection. The claim that inflexibility is the main cause of unemployment thus works as the keystone connecting all these generalizations, and therefore all these inferential connections constitute its meaning. The literature reflecting on the contribution of the *OECD Jobs Study* highlights this unifying function of the report. It presents the report as offering a 'view', a 'perspective', a 'framework', and even a 'paradigm' for unemployment research.[43]

We are now left with the two types of inferential connections that are 'inferentially downstream' in comparison to $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ and $B \overset{+}{\hookrightarrow} U$, and thus closer to language exit. We start with research implications and keep policy implications for the end.

Believing the generalizations has implications on the kind of research worthy of being pursued. The *OECD Jobs Study* fueled research on labor market institutions which were mainly focused on finding "rigidities"

[42] The '*roughly* in line' is important here. In 1994, the evidence was gappy and polyphonic. We cannot do justice to this complextiy here. The interested reader is referred to OECD (1994a, part 2, chap. 8).

[43] All these terms are, for instance, used in the (critical) volume of Howell (2005). The authors also talk about an 'orthodoxy', and make the connection with neoliberal ideology. It seems indeed correct to say that the inflexibility claim is also inferentially connected with even more abstract claims about the purported 'efficiency' of free markets. We will not go down this road in our non-exhaustive semantic analysis, and will leave for another time the inferentialist treatment of 'ideology'.

(Boeri and van Ours, 2008, p. 1). For economists accepting the rigidity view, the main research issue is to increase the resolution of the picture by offering a finer analysis of various institutions and the types of rigidities that they generate. We can thus say that $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ and $B \overset{+}{\hookrightarrow} U$ are inferentially connected to a whole framework about how to analyze labor markets. This framework suggests a direction for subsequent research.

While research implications are concerned with reforming the science, policy implications are about reforming the object of the science. In the *OECD Jobs Study*, causal generalizations play a pivotal role between the compilation of evidence and policy strategies. The structure of the *Study* hints at this role. It is made of two separate reports: a ‘scientific’ report, subtitled ‘Evidence and Explanations’ (OECD, 1994a), which is explicitly presented as the evidential base for a policy-oriented report, subtitled ‘Facts, Analysis, Strategies’ (OECD, 1994b). The transition from the first report to the second report is ostensibly done through the formulation of causal generalizations. Causal generalizations are meant to summarize what we know and point in the direction of what we can do. More specifically, the *Study* ends with nine broad recommendations, which are then subdivided in about 70 narrower statements. Only the first broad recommendation—about growth-enhancing and cycle-smoothing macroeconomic policy—is not directly connected to the inflexibility claim. All the others are meant as ways to “enhance the ability to adjust and to adapt” (OECD, 1994b, p. 43). Among them are recommendations targeting the generosity of unemployment benefits.[44]

The policy output of the *Jobs Study* was holistic: the main thrust was to give a direction to the multitude of policy reforms to come. A decade later, the OECD gave a fair account of the status of the *Jobs Study*’s recommendations:

> The general policy recommendations presented in this study provided an overall framework for reform which has come to be known as the ‘OECD Jobs Strategy’. (OECD, 2006b, p. 24)

[44] The two recommendations most directly concerned with unemployment benefits are:

- “Restrict UI benefit entitlements in countries where they are especially long to the period when job search is intense and rapid job-finding remains likely.
- Reduce after-tax replacement ratios where these are high, and review eligibility conditions where these require little previous employment history before drawing benefits.” (OECD, 1994b, p. 48)

The OECD did not believe in the piecemeal efficacy of its recommendations—i.e. it was not claiming that implementing *one* of its recommendations in a *single* country would reduce unemployment in this country. The OECD's real commitment was that, as policy makers started endorsing its "overall framework", and as they came to act on it, unemployment would decline in OECD countries.

With our analysis of the research and policy implications of $\mathit{Inflex} \overset{+}{\hookrightarrow} U$, we reach the stage of language exit. These two types of implications made the *OECD Jobs Study* the kickoff of a vast policy-oriented research project, which was meant to adapt the broad strategy to the circumstances of each country:

> The general Jobs Strategy framework was subsequently used to derive country-specific policy recommendations – tailored to the institutional, social and cultural characteristics of each member country – in the regular country reviews conducted by the Economic and Development Review Committee. (OECD, 2006b, p. 24)

Indeed, the *Jobs Study* was followed one year later by a report subtitled 'Implementing the Strategy' (OECD, 1995), which paved the way to a chapter entitled 'Implementing the OECD Jobs Strategy' in each country-specific report the next year.[45] Each country thus received its own list of suggested reforms. The actual policy recommendations thus differed across countries, but they were at the same time clear instantiations of the ones in the *OECD Jobs Study.* It is also the case that, in reading the country-specific analyses, one cannot miss the framing role of the 'rigidity view'. In going country-specific, the inferences of the OECD were profoundly guided by the framework set out in the 1994 report.[46]

[45] Starting with the Italian version of the *OECD Economic Surveys* in January 1996 (OECD, 1996a), each country got its own chapter. Some countries, e.g. France, had their chapter published only in 1997. The implementation of the recommendations was further monitored at a country level in later editions of the *OECD Economic Surveys*, and at the cross-country level in many publications (e.g. OECD, 1998, 1999).

[46] A fascinating implication of the rigidity view is that two countries which seemed to have fairly good unemployment performances were treated quite differently if one appeared 'more rigid' than the other. This contrast is stark when comparing the reports of the United States and of the Netherlands. These two countries had similar unemployment rates between 1993 and 1995 (averaging at 6.2%), rates which made other countries envious. However, the 'flexible' United States were offered a light medicine (eight narrow recommendations spanning only two of the nine broad recommendations, see OECD, 1996c, p. 74) while the 'rigid' Netherlands had to act across

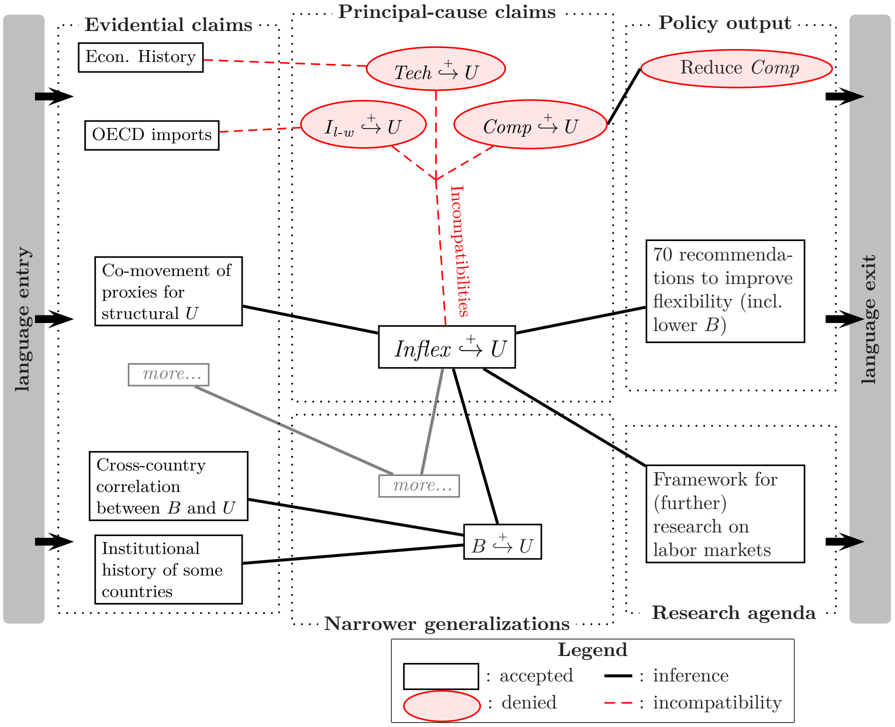

language entry
Evidential claims
Econ. History
OECD imports
Co-movement of proxies for structural U
more...
Cross-country correlation between B and U
Institutional history of some countries
Principal-cause claims
$\mathit{Tech} \overset{+}{\hookrightarrow} U$
$I_{l\text{-}w} \overset{+}{\hookrightarrow} U$
$\mathit{Comp} \overset{+}{\hookrightarrow} U$
Incompatibilities
$\mathit{Inflex} \overset{+}{\hookrightarrow} U$
more...
$B \overset{+}{\hookrightarrow} U$
Narrower generalizations
Policy output
Reduce Comp
70 recommendations to improve flexibility (incl. lower B)
Framework for (further) research on labor markets
Research agenda
language exit
Legend
: accepted
: denied
: inference
: incompatibility

Figure 1.1: *Simplified semantic network of the OECD's main generalization*

Figure 1.1 summarizes the inferential connections that we have discussed. According to inferentialism, the meaning of the claims $\mathit{Inflex} \overset{+}{\hookrightarrow} U$ and $B \overset{+}{\hookrightarrow} U$ is constituted by their inferential connections. We draw the inflexibility claim in the middle of the network and show the five types of propositions to which it is inferentially connected. We do not claim that these connections are exhaustive, and consequently do not believe that our analysis captures the entire meaning of the two generalizations on which we have focused. It should however be clear by now that these generalizations are meaningful in that they play a key role in an *inferential practice*. Someone might want to connect them, in the spirit of referentialism, to propositions about 'what the world should be like for these sentences to be true'. But doing that would only amount to add an extra type of connections to an already rich inferential network.

1.4.3 Discussion

The reader must entertain many questions about our inferentialist analysis of causal generalizations. Inferentialism is in an early stage of development as a framework to analyze causal claims. This framework is promising, but still somewhat immature. In this subsection, we want to address three of what are probably the most pressing worries among readers.

First, the view that meaning is constituted by inferential connections might sound totally counterintuitive to some. Are we not committing a *blatant* category mistake in trying to locate the meaning of a sentence in the inferential network to which it participates? We think not. When it comes to the intuitiveness of a framework, inferentialism might hold the high ground against referentialism. We would speculate that the superiority of inferentialism with respect to intuitiveness would be challenged mainly by individuals trained in philosophy or linguistics; in other words, only by individuals with 'trained intuitions' for referentialism.

Think about the typical answer one would receive to the question: 'What do you mean by this statement?' For concreteness, say that the statement is our now familiar inflexibility claim. When you ask economists about the meaning of this inflexibility claim, typical answers are 'it means that factors such as minimum wages and unemployment benefits are causing unemployment', 'it means that we should stop blaming

the board according to the OECD (19 narrow recommendations subdivided into six classes, see OECD, 1996b, p. 72-73).

developing countries', 'it means that an effective strategy to pull unemployment down is to reduce labor-market frictions', and so on. For a referentialist semantics, all these answers are not going to the point; they do not directly tell us what the world should be like for the sentence to be true. To be sure, a proponent of a referentialist semantics might hope to recover the (referentialist) meaning of the claim by combining information from all these answers. But she has to grant that the answers are directly answering the semantic question as understood by an inferentialist semantics, and only indirectly, if at all, the referentialist variant of this question. In other words, if we ought to be blamed for thinking of meaning as being constituted by inferential connections, the layman (including the typical economist) is also to blame.

Second, we came to be critical of our referentialist procedure because it fails to rationalize the widespread practice of demanding and supplying causal generalizations in policy-oriented social science. Is this popular business really rendered reasonable by an inferentialist semantics? We have here to distinguish between the general practice of demanding and supplying causal generalizations, and a specific instance of this practice (e.g. the demand of the OECD ministers and the supply of the *OECD Jobs Study*). We consider that the *general* practice comes out clearly as reasonable. This practice is generally reasonable because causal generalizations, even though they might not have a clear referential relation, are highly valuable in that *they structure our cognition*—i.e. they help us extract salient elements from the ocean of data, connect elements that might seem unrelated to someone else, collect new elements in a systematic way, and form plans of action. Given that policy makers want tools to cope with the world, it is totally reasonable for them to ask for these great tools that are causal generalizations. What we called the high-value puzzle in section 1.3 is solved.

There is also no limited-clarification puzzle to begin with. It is not that the experts have a definite referentialist meaning for the generalizations that they utter, meaning that they simply omit to spell out for their target audience. In the case of the *OECD Jobs Study*, the experts are spelling out the meaning of their generalizations since they report the propositions from which they infer their generalizations, which additional propositions are inferred from these generalizations, and how generalizations relate to each other. In other words, the inferential network that we depict in figure 1.1 is taken directly from the report, it is not hidden.[47]

[47] There is one interesting point that can be added here about meaning transmission. The inferential network depicted in figure 1.1 is the one of the economists who

While the general practice of demanding and supplying generalizations appears reasonable from the perspective of an inferentialist semantics, it would be problematic if no *instance* of this practice could be *un*reasonable from this perspective. This is the third and last worry we want to address: Are we not losing the prescriptive dimension of a semantic analysis by turning to inferentialism? Remember that the recent literature on the semantics of causality has the explicit goals—exemplified by Woodward in the introduction to this chapter—not only of determining the meaning of causal claims, but also of making recommendations about the practice of causal reasoning. The worry is that we have to give up this last goal if we endorse an inferentialist semantics.

Our semantic analysis of the *OECD Jobs Study* in the previous subsection self-consciously avoided assessment, but there is certainly room for it. In fact, what scientists do daily in questioning specific causal generalizations can serve as a template for how we can combine an inferentialist semantics with an evaluative stance. Although the meaning of $X \hookrightarrow Y$ is given by its inferential connections, it might be that some claims that are connected to $X \hookrightarrow Y$ should be rejected, or that some of the inferential connections one is actually disposed to make are not justified. We already saw arguments of the first type being used by the OECD to reject the three alternative principal-cause generalizations. For example, the policy recommendation to reduce the intensity of competition was rejected, which in turn contributed to the rejection of the claim 'the principal cause of high unemployment is the high intensity of competition'. Arguments of the second type—about rejecting a connection—are also widespread. They have, for instance, been used *against* the OECD's generalizations. It has been argued that the cross-country correlations that the OECD was reporting in 1994 are so weak and unstable that they can hardly be linked to the OECD's generalizations by an evidential-base connection (Baker et al., 2005). Furthermore, the OECD recognized later that its inference from its main causal generalization to policies emphasizing deregulation of the labor market was not sound, since 'flexibility' might also be achieved through wise regulation instead of deregulation (OECD,

wrote the *Jobs Study*. It is not certain that the exact same network is shared by the audience of the report. But full meaning preservation is not necessary for the practice of demanding and supplying generalizations to be reasonable. One would expect that, given the policy-oriented nature of the exercise, what must be transmitted with little distortion are the policy implications of the generalizations, and that an accurate transmission of the evidential-base connections is less important. The numerous summaries of the *Jobs Study*—for the press, for policy makers, and so on—are exactly emphasizing the policy implications of the principal-cause claim.

2006a, p. 19).

We can join scientists in evaluating generalizations on these grounds. There is no reason to think that an inferentialist semantics will restrain us in this task. It is even to be expected that such a semantics will contribute to a more principled assessment of generalizations. By turning the spotlight to the set of inferential connections, an inferentialist semantics should help us spot the weak points in the network.

1.5 Conclusion

The main lesson that we draw from this chapter is that the semantic analysis of causal generalizations in policy-oriented social sciences must recognize that the meaning of these statements is not a property of each of them taken in isolation, but a property of them as units in an inferential practice. The inadequacy of a referentialist semantics stems from its focus on a single word-world connection. An inferentialist semantics, in contrast, opens our eyes to the central roles played by causal generalizations in our cognition. These roles are what make these utterances meaningful, and what makes them valuable.

In the case of the *OECD Jobs Study*, the inflexibility claim (re)structured the thinking patterns of communities dedicated to studying and intervening on labor markets. To the extent that we judge such a structuring role to be valuable—not necessarily in the OECD case but in general—we should worry about a too strict devotion to a referentialist semantics. By trying to speak more meaningfully, we might end up impoverishing our language and, in consequence, our cognition.

Part II

Epistemology

Preliminaries

We, humans, are terrific hypothesis generators. Some maintain that business cycles are caused by fractional reserve banking; others that the protestant ethic caused the rise of capitalism; and yet others that the inflexibility of labor markets is the principal cause of high unemployment. Scientific methodology promises to provide ways to select among this ocean of hypotheses, to guide us in accepting and rejecting claims.

Scientific methodology is partly about evidence, about its assessment and its generation. The notion of evidence is polysemic.[48] For my purposes, evidence can be broadly defined as a reason to believe or disbelieve a hypothesis. Evidence is thus always evidence *for* or *against* a hypothesis. It improves or degrades the reasonableness of accepting a hypothesis.

A great part of our social life is about providing or requesting reasons for the claims we, or others, put forward. A condition of possibility of this practice is the conviction that not anything can count as a reason. The fact that Yvan Dupin is mad about football is no evidence for the hypothesis that my computer will survive a few more hours; there is no way that Dupin's taste will make beliefs about my machine's longevity more or less reasonable. If everything could count as evidence for a hypothesis, providing or requesting reasons would be a strange ritual indeed. This practice would not have the meaning that it has in our social life.

Although we are convinced that not anything *can* count as evidence, it is often not transparent what *does* count as evidence for a given claim. It is even harder to judge the *strength* of a reason. Take the weather forecast in 3 days from now: one website announces 5 millimeters of rain on Rotterdam. Should I have already canceled my barbecue? It is extremely hard to tell from my non-expert position whether the forecast is a strong reason to believe that it will actually rain on that day. As a consumer of weather reports, I have been totally unimpressed by the performance

[48] For general discussions of the notion, see Achinstein (2005); Kelly (2008); Reiss (2011a).

of 3-day-ahead forecasting in Holland, but others have different opinions. I have even been declared unreasonably pessimistic about the reliability of weather forecasting. In other words, I might have no strong *reason* to consider 3-day-ahead forecasting to be an extremely weak *reason* for believing anything about the actual weather in 3 days. This colloquial example shows that, in the practice of providing and requesting reasons, there is also an exchange of reasons for our claims about (strength of) evidential connection—reasons for reason, or second-order reasons. If pressed, for instance, to provide reasons for my limited trust in weather reports, I would say that it often happened that 3-day-ahead forecasts were radically off target. Someone might reply that my memory is selective, that I only remember the failures. And the exchange of reasons could go on.

The literature concerned with scientific methodology contributes to the exchange of second-order reasons. For instance, someone versed in the methodology of weather forecasting could no doubt inform me about the past reliability of 3-day-ahead forecasting. She might even give me a confidence (or credible) interval for the amount of rain and provide the assumptions used in the construction of this interval. More generally, scientific methodology attempts to systematize our assessment of evidence and evidential strength. It gives us guidelines to make our conversation on second-order reasons more systematic and less prejudiced. A corollary of this quest for a more principled assessment of evidence is that the literature on scientific methodology has developed new methods to *generate* evidence, from controlled experiments to econometrics. Once we have a better idea of the properties needed for evidence to be a *strong* reason for or against a hypothesis, we can design methods likely to generate evidence with these properties.

In this dual task of assessing evidence and generating evidential methods, the literature on scientific methodology has typically focused on single methods. We thus have a good understanding of, e.g., the conditions for a randomized controlled trial to give an accurate estimate of the average causal effect of a specific treatment for a given population, and the conditions for the parameter estimates of an ordinary least squares (OLS) regression to be unbiased. There has been considerably less attention given to the issue of evidential strength when multiple evidential elements are combined.[49] In this part of my thesis, I contribute to redress-

[49] One notable exception is the literature on meta-analysis, but this method is still limited to evidential elements which are much alike, i.e., outcomes of statistical research typically based on RCTs (Stegenga, 2011).

ing this imbalance. I make a few steps toward a better methodological understanding of evidential variety.

Chapters 2 and 3 are attempts to put evidential variety on the methodological agenda. Chapter 2 is primarily addressed to economists and economic methodologists. It argues that, although the methodological discussion is focused on single-method assessment, practicing scientists sometimes—and perhaps most of the time—have to jointly draw on multiple methods to gather strong evidence for or against a hypothesis. The necessity of multiplying the sources of evidence comes from the fact that the reliability of each available source is open to doubt. Relying on a single source would thus not supply compelling evidential support for any hypothesis. The reliance on multiple sources raises an epistemic question: How ought we to assess the evidential strength of a diverse body of evidence?

Chapter 3 is targeted at philosophers of causality. It claims that the quest for evidential variety has been misinterpreted as having implications for the semantics (and the metaphysics) of causality. I rather argue that this quest is a consequence of the epistemic situation in which many scientists find themselves: collecting evidence from multiple sources is a reasonable strategy when we face uncertainty regarding the reliability of each source. There is no reason to interpret this strategy as having implications for the meaning of causal claims.

The reader will note that there are significant overlaps between chapters 2 and 3. They however focus on different elements, in part because they have different target audiences and are published in different venues.

Chapter 4 analyzes the notion of evidential variety in a Bayesian framework. This framework attempts to capture in a mathematically tractable way our intuitive ideas about evidence and hypotheses. It postulates that degrees of belief in a hypothesis are probabilities. If we denote a hypothesis by h, the degree of belief in this hypothesis is thus $P(h)$. The evidential relation is captured by conditional probabilities: evidential element e is evidence *for* hypothesis h if, and only if,

$$P(h|e) > P(h). \tag{1.7}$$

In words, the degree of belief in the hypothesis is higher if the evidential element is known. There are multiple proposals to capture the notion of evidential *strength* (Hartmann and Sprenger, 2011). For instance, the 'difference measure' takes the difference between the two probabilities used in formula (1.7):

$$d(h,e) = P(h|e) - P(h) \tag{1.8}$$

The bigger this difference, the higher the evidential strength of e for h. It is with simple formulas like (1.7) and (1.8) that Bayesian epistemology specifies key ideas like being a *reason* for a hypothesis, and the *strength* of this reason.

I said earlier that our social practice of providing and requesting reasons is founded on the belief that not anything can count as a reason for a hypothesis. If we were free to postulate any value (between 0 and 1) for $P(h|e)$ and $P(h)$, Bayesian epistemology would not be compelling. This framework does however constrain the assignment of values, and thus it coheres with our belief that not anything can count as a reason. To be sure, most variants of Bayesian epistemology do not go as far as fully determining these values.[50] They only require that probability assignments are *consistent* in that they must respect the axioms of probability theory (Howson, 1997). One consequence of these axioms is Bayes' theorem

$$P(h|e) = \frac{P(h)}{P(h) + P(\neg h)\frac{P(e|\neg h)}{P(e|h)}} \tag{1.9}$$

which is here presented in the likelihood-ratio form. Note that a corollary of the axioms of probability theory is that $P(\neg h)$—i.e. the prior belief in the *negation* of the hypothesis—is equal to $1 - P(h)$.

Bayes' theorem connects the prior belief in a hypothesis, $P(h)$, to the belief in the hypothesis given the evidential element, $P(h|e)$. The theorem relates these two probabilities through the likelihood ratio $P(e|\neg h)/P(e|h)$; the higher this ratio, the lower the strength of e as a reason for h. The likelihood ratio captures the relation between two beliefs: (i) how likely the evidence is given that the hypothesis is not the case, (ii) how likely the evidence is given that the hypothesis is the case. Equation (1.9) says that, for a fixed prior belief in the hypothesis, the evidential strength of e for this hypothesis increases when (i) e is *less* likely when the hypothesis is not the case, and (ii) e is *more* likely when the hypothesis is the case.

Bayesian epistemology promises to help us better understand evidential variety. When one relies on multiple sources of evidence, the strength of the resulting body of evidence is hard to assess. When it is hard to analyze a situation because of its complexity, scientists typically construct models. It is exactly what I propose to do for epistemological questions concerning evidential variety.[51] Indeed, the probability calculus can help

[50] *Objective* Bayesianism can go as far as giving rules that fully determine probabilities, but I will not go down this road.

[51] I see my project as following Paul Horwich's (1998) program of "therapeutic

us track the conditions on which evidential strength hinges. In chapter 4, I focus on the claim that a condition of independence must hold among the sources of evidence. Relying on and extending the work of Luc Bovens and Stephan Hartmann (2003), I give an interpretation of independence as *reliability independence*, and show that the evidential implications of independence are less straightforward than is typically believed.

This part is a contribution to the methodological understanding of evidential variety. I have to say that this contribution is limited in comparison with the richness of the topic. Since it seems obvious to me that evidential variety and its challenges are central to the practice of science, I do think that my small steps are steps in the right direction.

Bayesianism" and as answering Stephan Hartmann's (2008) call for more modeling in philosophy of science.

Chapter 2

Evidential Variety as a Source of Credibility for Causal Inference: Beyond Sharp Designs and Structural Models

2.1 Introduction

Economists do not agree on the best approach to causal inference. The main divide is between the 'experimentalist school'—what will henceforth be labeled the 'design-based approach' (Angrist and Pischke, 2010; Imbens, 2010)—and the structural approach (with James Heckman (2005, 2008) as its main contemporary proponent). The tone of the debate rose in 2010 with the publication of special issues in the *Journal of Economic Perspectives* and the *Journal of Economic Literature*,[1] one participant going as far as writing that "[f]ew topics in economics evoke more passion than discussions about the correct way to do empirical policy analysis" (Heckman, 2010, p. 356).

*This chapter was originally published as François Claveau (2011), "Evidential Variety as a Source of Credibility for Causal Inference: Beyond Sharp Designs and Structural Models," *Journal of Economic Methodology* 18 (3): 233-253. The present version includes some modifications from the published article.

[1] See the Symposium 'Con out of Economics' in the *Journal of Economic Perspectives* (Vol. 24, No. 2) and the 'Forum on the Estimation of Treatment Effects' in the *Journal of Economic Literature* (Vol. 48, No. 2).

Although the two approaches are not mutually exclusive—indeed, most if not all protagonists in the debate argue for a possible reconciliation (see, especially, Heckman, 2010; Imbens, 2010; Nevo and Whinston, 2010; Stock, 2010)—they offer different answers to the question 'What makes a causal inference credible?' The design-based approach points to sharp study designs—randomized controlled trials (RCTs) or natural experiments—while the structural approach argues that one should rely on structural models informed by economic theory.

Are these two approaches exhausting the possibilities? Certainly not. I will argue that evidential variety must also be recognized as being part of the toolbox of the applied economist.[2] I understand evidential variety narrowly as the outcome of using multiple means of determination to estimate a property of interest. If the different means of determination give similar estimates—i.e. if we have concordant evidence—the result is said to exhibit 'measurement robustness', which is one type of robustness in the typology offered by James Woodward (2006). Evidential variety as discussed here is thus a specific subset of what William Wimsatt (2007b) called 'robustness analysis'. Most importantly, it should not be conflated with sensitivity analysis, a procedure often invoked in economics (Leamer, 1983) but of dubious epistemic value (see Hoover and Perez, 2004; Aldrich, 2006; Woodward, 2006).[3]

Relying on evidential variety is compatible with employing the other tools—sharp designs or structural models—but it is particularly relevant when the other tools are not directly applicable to the causal question

[2] There are certainly other elements in the toolbox which will not be discussed here. To start with, we must recognize that approaches to causal inference are not exhausted by the design-based and the structural approaches even though the recent exchanges in the *Journal of Economic Perspectives* and the *Journal of Economic Literature* might give this impression. For a more inclusive typology, see Hoover (2008).

[3] "[I]n sensitivity analysis a single fixed body of data D is employed and then varying assumptions are considered which are inconsistent with each other to see what follows about some result of interest under each of the assumptions." (Woodward, 2006, pp. 234-5) In contrast, measurement robustness comes from varying the measurement procedures—for instance, using Brownian motion, alpha radiation and helium production in the case of the measurement of Avogadro's number discussed later. Each measurement procedure draws on its own observations and produces its own body of data. Furthermore, the assumptions used to derive results from these measurement procedures are typically consistent with each other, at least we hope so. In Woodward's typology, measurement robustness is also distinguished from causal robustness and derivational robustness. The latter has been used recently to analyze theoretical models in economics (Kuorikoski et al., 2010; it is related to a wider literature including Levins, 1966; Orzack and Sober, 1993; Weisberg, 2006).

under study. My argument is not only that evidential variety might be used but also that it is already in use. To make this point, the chapter draws on a concrete example introduced in the next section: research on the institutional causes of the aggregate unemployment rate. Section 2.3 presents the design-based and the structural approaches together with their recognized limitations. I then raise some doubts about the applicability of these approaches to a class of macro-level causal questions (section 2.4). It appears that their limitations might often be binding for this class of questions. In section 2.5, the literature on the institutional determinants of aggregate unemployment is analyzed as a case of evidential variety. In the conclusion, I come back to the main lesson: sharp designs, structural models, evidential variety, these are elements in our toolbox which might be conducive or not to credible inference—alone or in conjunction—depending on the epistemic context.

2.2 A concrete inferential problem: institutional causes of unemployment

The long-run unemployment performance of countries varies dramatically: over the last 10 years for instance, the mean unemployment rate in the Netherlands has been 3.5% while it has been 8.9% in France.[4] Policy-makers have control over many dimensions of the labor market—e.g. minimum wage rates, unemployment insurance, and layoff restrictions. A natural question is thus: What is the effect of playing with these different levers (henceforth *Inst*) on the aggregate unemployment performance (U)? Let me only focus on the qualitative effect. For example, does the level of unemployment benefits have a positive, negative or null causal effect on the aggregate unemployment rate?

It turns out that economists specializing on the causes of aggregate unemployment agree on the answer to many of these qualitative causal questions. In other words, the evidence amassed lends *credibility* to one answer according to the specialists. How was this credibility achieved? Was it in line with the methodological prescriptions of the design-based

[4] The ‘long-run’ qualifier is added to distinguish between fluctuations of aggregate unemployment with the business cycle and the general level of aggregate unemployment through the cycle. It is an established fact that unemployment increases in economic downturns but some countries have systematically lower unemployment rates than others whenever one makes the comparison. Another label for long-run unemployment is structural unemployment which is contrasted to cyclical unemployment.

approach or the structural approach? I will argue that, to make sense of cases like this one, we have to make room for a third source of credibility, namely evidential variety.

A few comments are in order before turning to the sources of credibility, especially to disambiguate the causal questions under consideration. But first, a comment about credibility itself. The fact of consensus in the relevant scientific community is not taken to imply the truth of the causal claim. The history of science is replete with examples of false propositions, which appeared to be credible to all specialists at the time. Fortunately, the debate between the design-based and the structural approaches is over the sources of credibility. Consensus in the unemployment example can thus be interpreted only as a *credibility* achievement without committing oneself to assert the truth of the consensual answers.

Second, about the level of the causal claims in the unemployment example. They relate macro-level properties. Unemployment can be a property of an individual—i.e. a member of the labor force is either employed or unemployed at any point.[5] I look instead at the macro-level property: for each economy, a fraction of the labor force is unemployed at a given time. The same distinction applies to the policy levers: I focus on *Inst* as being rules of an entire economy. The fact that both causal relata are at the macro-level is important to the present argument. Indeed, it will be argued that one should not assume that the design-based and the structural approaches are applicable to all macro-level causal questions.

Third, about the meaning of causation. As I argue in chapter 1, devising an appropriate semantic analysis of the causal claims in the literature on aggregate unemployment is no trivial task. It seems, in particular, that at least some causal claims require an *inferentialist* analysis. In contrast, the main proponents of the design-based and the structural approaches explicitly endorse a *referentialist* semantics. More specifically, they put forward variants of the manipulationist-counterfactual account of causality, which was already discussed in section 1.2 (see also p. 11 of

[5] A subset of the employed can also be labeled as underemployed—workers having an involuntary part-time job or being overeducated for their current job. Similarly, a subset of individuals officially out of the labor force are closer to potential workers than the rest of the inactive (e.g. they are willing to work but not currently searching). While I stick in the body of the text to the main division between employed and unemployed, one might think that a more comprehensive typology including the underemployed and the discouraged job-seekers is preferable (for well-informed policy decisions for instance). I don't think that my methodological point hinges on this choice. For a discussion of the complex definition of 'unemployed', the reader is referred to subsection 1.2.1.

the introduction).[6] Since my methodological point in this chapter does not require an inferentialist semantics, I follow the design-based and structural approaches in their manipulationist-counterfactual analysis. For our example of policy claims, saying that *Inst* has a positive effect on U is thus interpreted to mean that if *Inst* was higher due to some manipulation *just in the right way*, U would be higher than it would otherwise have been.[7]

My last comment regards the strictness of the causal claim associating *Inst* to U (under a referentialist analysis). Should we understand it as an assertion that, in each and every country in the population of interest (e.g. high-income economies), the causal relationship holds as stated, or is it the weaker assertion that, because of causal heterogeneity, the stated causal relationship holds only *on average* in this population? Causal heterogeneity is, for instance, a plausible assumption for claims about the causal effect of a drug on individuals—e.g. by saying that a certain dose of aspirin reduces headache for humans, one can leave open the possibility that some individuals do not respond to the drug. Similarly, and as argued above in section 1.2.3, causal heterogeneity is most plausible for claims about the institutional determinants of the unemployment rate. Consequently, the causal claims about unemployment will be interpreted here as 'average effect' claims that are obviously strictly weaker than 'homogeneous effect' claims.[8]

[6] Advocates of sharp designs endorse the potential outcome framework (Holland, 1986; also called the Rubin Causal Model) and Heckman opts for a semantics of manipulation of external inputs to a causal structure (Heckman and Vytlacil, 2007, sec. 4).

[7] Variants of the manipulationist-counterfactual account spell out differently the conditions for the manipulation to be just in the right way. These details are not relevant to the present argument.

[8] The reader should note that the set of inferences one can draw from average claims is quite smaller than that for homogeneous ones—e.g. you are not entitled to assert that an intervention *in a specific country* will have this effect. This point is important for not overstating the degree of consensus among specialists. Even if many qualitative, average causal effects were consensual, there might well be a lot of disagreement over policy claims for a given country. Note also that if we take the averaging seriously, the truth of the causal claim does not even ensure that we will *most of the time* accurately capture the direction of the effect of intervening in a country picked at random. Knowing the median causal effect will do that but not the average effect if we cannot rule out that the distribution is skewed.

2.3 Two conflicting approaches

Say we want to investigate whether some policy variables—level and duration of unemployment benefits, strictness of employment protection legislation and so on—have a qualitative causal effect on the aggregate unemployment rate. In search of some methodological guidance, we might turn to the current debate in the econometrics of causal estimation.[9] There we find two rather disjoint sets of recommendations: one from the design-based approach and the other from the structural approach.[10] The debate between these approaches is framed around the question of the sources of *credibility*.[11]

2.3.1 The design-based approach

This approach is tailored to deal with a fundamental challenge for causal inference: the problem of confounding. Figure 2.1 uses a causal graph to present one typical case of confounding.[12] In a causal graph like this, all nodes are variables; a solid circle (for X and Y) represents an observed variable while a hollow circle (for C) means that the variable is unobserved. The directed edges—for instance, X to Y—mean that the variable at the origin of the arrow causes the terminal variable—X causes Y. Say now that, in our data, we observe an association between variables X and Y. We would like to interpret this association as capturing the causal effect from X to Y but, given the causal graph of Figure 2.1,

[9] See footnote 1.

[10] Other labels are available but the distinction between design-based and structural has the advantages of (i) capturing the main divide and (ii) not insinuating that one of the two is better than the other. Other distinctions which do not have these advantages include Heckman's (2005) statistical versus scientific, Heckman's (2008) statistical versus econometric, Imbens' (2010) causal versus structural.

[11] To be precise, proponents of the design-based approach make a disproportionate use of the term credibility. It is far from a new trend to criticize the structural approach for its allegedly "incredible identification" restrictions (Sims, 1980). Contemporary advocates of the design-based approach draw again on this criticism by announcing a "credibility revolution" (Angrist and Pischke, 2010). Defenders of the structural approach will not capitulate easily and their reply, even if expressed in different ways, can be reconstructed in terms of credibility: while the design-based approach does *credibly* identify *specific* causal effects, *credible* policy analysis will typically require an explicit reliance on economic theory (see, e.g. Keane, 2010; Nevo and Whinston, 2010; Heckman, 2010).

[12] See Morgan and Winship (2007, chap. 3) for an introduction to this tool applied to the social sciences; the two standard references on causal graphs are Spirtes et al. (2000) and Pearl (2009).

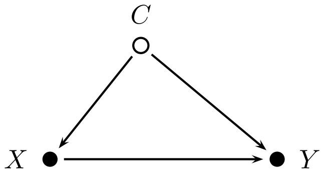


Figure 2.1: *The unobserved C confounds the causal effect of X on Y*

the observed association is produced both by the causal effect of interest and by the causal relationships that C has with X and Y. The challenge for causal inference is thus to find credible ways to rule out that the association we would like to interpret as causal is not produced by such an unobserved C.

The design-based approach locates the main source of credibility of a causal inference in the process that generated the data. It starts from the following observation: “The most credible and influential research designs use random assignment” (Angrist and Pischke, 2008, p. 11). A causal claim made on the basis of a randomized controlled trial (RCT) is credible because of the fact that effective randomization takes care of systematic sources of confounding. If one has a binary variable $X = \{x_0, x_1\}$, assigning units randomly to x_0 or x_1 ensures that the units treated with x_1 will not systematically—they might do so by chance—share prior characteristics that units treated with x_0 lack. The units treated with x_1 are not caused to have this value by confounding factors like C, but by the experimenter setting them to this value. The average difference in outcome Y between the group treated with x_1 and the one treated with x_0 can thus be imputed to the causal effect of X—again with sampling error.

It is important to be precise about the causal effect which is directly supported by a RCT; it is an average causal effect for a specific population of units resulting from changing the cause variable from one value to the other. If units are causally heterogeneous—if they respond differently to treatments—the average claim is not transferable to the unit-level causal effect (see the aspirin example above). Similarly, claiming that the average causal effect is similar in other populations or for different values of the causal variable requires further assumptions. These caveats regarding the precise interpretation of the causal claim directly supported by a RCT are important for the discussion in the following sections.

The observation that random assignment lends a high degree of credibility to causal inference—inference to a specific but well-defined causal effect—leads proponents of the design-based approach to their first methodological recommendation: one should attempt to answer causal questions by designing, if feasible, RCTs. And indeed, the rise to prominence of the design-based approach in the last twenty years is associated with a significant increase of randomized experiments in some field of economics—especially in development economics but also to some extent in labor economics (Imbens and Wooldridge, 2009, sec. 4).

The design-based approach also has something to say when randomized experiments are not feasible:

> But experiments are time consuming, expensive, and may not always be practical. It's difficult to imagine a randomized trial to evaluate the effect of immigrants on the economy of the host country. However, human institutions or the forces of nature can step into the breach with informative natural or quasi-experiments. (Angrist and Pischke, 2010, p. 4)

The idea of natural experiments is that the researcher can have access to data generated by a process akin to random assignment without being the one performing the assignment; randomization is done by institutions or by Nature—e.g. families randomly receive a voucher in the mail to partly pay for private school (see Morgan and Winship, 2007, chap. 7). The econometrics associated to natural experiments are instrumental variables, regression discontinuity and difference-in-difference methods (Angrist and Pischke, 2008; Imbens and Wooldridge, 2009, sec. 6.3-6.5).[13] In studies relying on these methods, the case for the validity of the design hinges on a "credibly exogenous source of variation" (Angrist and Pischke, 2010, p. 16). The second methodological recommendation of the design-based approach is thus to focus energy on finding study designs for which the 'as-good-as-randomly-assigned' assumption (Angrist and Pischke, 2010, p. 12) plausibly holds.

Again, one has to carefully specify the causal effect identified by a natural experiment—the local average treatment effect (LATE)[14] in the

[13] These techniques are not solely associated to the design-based approach. Instrumental variable, for instance, is also extensively used in the estimation of structural models.

[14] LATE is 'local' in the sense that it identifies the average treatment effect for a specific subpopulation, i.e. the units induced by the instrument to change their treatment status. One can equate LATE to the population average treatment effect only under the rather restrictive assumption that the subpopulation is a representative sample of the general population. Note that, on top of the usual exclusion restrictions for instrumental variables, LATE requires the monotonicity assumption (Imbens and Angrist, 1994, p. 469): the instrument must not have a positive effect on the probability

case of instrumental variable (see Imbens and Angrist, 1994 for the original derivation; for a discussion, see Morgan and Winship, 2007, chap. 7). But the main point is that the credibility of the causal inference comes directly from properties of the study design. Persuasive arguments about the design are the key to credible inference.

The major limitation of the design-based approach is that actual and natural experiments are not forthcoming for all causal questions. Both because of cost and feasibility, only a small subset of causal questions of interest will be answered by RCTs in the foreseeable future. The scope of natural experiments is also limited. Even the most optimistic will not dare to claim that the exclusion restrictions necessary for valid instruments are easily achieved.

2.3.2 The structural approach

The critics of the design-based approach emphasize the limitation just outlined: "I am not entirely certain what credibility means, but it is surely undermined if the parameter being estimated is not what we want to know" (Deaton, 2010, p. 430). Critics are willing to concede that controlled or natural experiments lead to credible inference for *some* causal questions but they add that these study designs will not necessarily (and indeed usually) be available to answer the causal question of interest. For the proponents of the structural approach, structural models have to be brought in to widen the scope of answerable causal questions.

The structural approach—championed by James Heckman (2000, 2005, 2008) among others—has a long history.[15] It is a direct heritage of the econometrics of the Cowles Commission (Koopmans, 1950; Hood and Koopmans, 1953). The structural approach proposes its own solution to the problem of confounding: using economic theory to model the causal structure relevant to our question of interest. The estimation procedure is meant to recover the structural parameters of the model, which, in turn, can serve to answer a multitude of causal questions.

In recent times—and especially in the work of Heckman—the structural approach has put special emphasis on modeling the potential source of confounding due to *self-selection* of units in different treatments. For

of being treated for a portion of the population and a negative effect on another portion. It must either affect positively all units (i.e. no defier) or affects them negatively (no complier).

[15] The design-based approach has an even longer history (outside economics) since it can be traced back to the work of statisticians like R.A. Fisher (1935).

instance, if one wants to learn the average causal effect of private versus public schooling on test scores of students in a given population, the average difference in test scores between the two types of school will typically be a bad indicator of this causal effect. The reason is that students (or most often their parents) have *chosen* their school and the reasons of their choice are plausibly not independent of the reasons why they perform better or worse with respect to test scores—e.g. more educated parents might be more likely to both send their children to private school and help them with their studies. It is thus likely that there are some confounding factors like C in figure 2.1 which cause both school type and test scores. Since confounding is likely to come from the fact that individuals *choose*, the proponents of the structural approach argue that the help of economic theory—which is in big part a formal apparatus to model choice—is to be welcomed.

A good economic model of a situation together with unbiased estimates of the parameters of this model—an outcome which is clearly difficult to achieve—amounts to an extremely powerful tool for causal reasoning. It is not limited to the small set of causal questions that can be directly answered by the design-based approach. The methodological recommendation of the structural approach is thus to base causal analysis on explicit economic models.

The limitation of the structural approach comes from its reliance on a set of identifying restrictions, which are said to be legitimated by economic theory.[16] A structural model which is deemed flawed will lead only to incredible causal inference. In fact, the rising popularity of the design-based approach in the last 20 years is in great part due to a suspicion regarding structural models. The design-based approach promises credible causal inference with as little as possible reliance on doubtful theory: “In a design-based framework, economic theory helps us understand the picture that emerges from a constellation of empirical findings, but does not help us paint that picture” (Angrist and Pischke, 2010, p. 23). In turn, the defenders of the structural approach claim that advances in theory and in its empirical implementation make the structural approach more credible than perhaps before.

Before assessing whether the methodological recommendations of these

[16] One can doubt that ‘economic theory’ is (principally) supplying the assumptions. It seems that *a priori* guessing and statistical conventions are playing a large role. Since my argument does not hinge on the relative importance of these other sources of restrictions, I stick to the main narrative of the structural approach which is ‘we need the support of economic theory to be successful in our causal inquiries’.

two approaches apply to the research on the causal effects of specific labor market institutions on the aggregate unemployment rate, I must note that one should not overstress the divide between the two approaches. In fact, the participants in the debate point to a potential reconciliation. The simple idea expressed by them is that the expertise of the design-based approach in finding exogenous sources of variation can be used to identify structural parameters, which can then be of help in answering a diversity of causal questions.[17] The extent to which this avenue can lead to a fusion of the two approaches does not matter to my current argument.

2.4 Initial doubts about the applicability to macro-level questions

The methodological recommendations of the design-based and structural approaches are meant to apply both to microeconomics and to macroeconomics. The empirical successes of the design-based approach are by and large microeconomic but Angrist and Pischke (2010, pp. 18-20) explain this imbalance by the failure of 'theory-centric' macroeconomists to embrace their approach. Similarly, the econometric work of Heckman in the last decade has been mainly microeconomic but he continues to press that "[t]he agent can be a household, a patient, a firm, *or a country*" (Heckman, 2008, p. 5, emphasis added).[18]

There are reasons to doubt that these methodological recommendations are easily transferable to (at least) a subset of macro-level causal questions. The doubts arise because the epistemic context of these questions is plausibly such that the two general limitations of the approaches—no access to sharp designs and no well-established theory—are binding.

Before explaining why these limitations will more often be binding, let me note that I have in mind a specific subset of macro-level causal

[17] This union comes from the widespread need in structural econometrics (acknowledged above in footnote 13) to use instrumental variables to identify the key parameters.

[18] Heckman (2010, p. 358fn) writes: "For brevity, in this paper my emphasis is on microeconometric approaches. There are parallel developments and dichotomies in the macroeconomic time series and policy evaluation literatures. See Heckman (2000) for a discussion of that literature." At play here is the assumption that the lesson learnt from microeconometrics can easily be transferred to the macro-level. The approaches discussed in Heckman (2000) are vector autoregression (VAR), structural estimation using dynamic stochastic general equilibrium (DSGE) models, calibration, sensitivity analysis and natural experiments.

questions which is a transposition from the individual level to the country level of the types of questions considered by the design-based and the structural approaches. These approaches typically ask 'What is, for an agent, the effect on outcome Y of doing x_1 instead of x_0.' Since an agent is usually observed in only one state, either x_1 or x_0—i.e. the famous evaluation problem—the question is shifted from the individual effect to a population analogue—e.g. the average treatment effect. As the quotation from Heckman in the beginning of this section suggests, one can take the agent as referring to a human individual but the agent can also be a country. I am thus considering a class of macro-level questions about causal relationships *for a population of countries*—e.g. developed economies.[19]

While this class of questions is part of economics—it is especially dear to international research institutions like the OECD—macroeconomics is more commonly about *country-specific* causal relations. This research typically relies on time series analysis and, when its goal is causal inference, its causal claims are, strictly speaking, only relevant to the country under study. Obviously, one can wish to *extrapolate* country-specific results to the whole population but let us leave this idea aside until we get to evidential variety. For now, the reader should keep in mind the peculiarity of the class of macro-level questions on which I want to focus.

2.4.1 Sharp designs?

Most discussants of the Angrist-Pischke paper express scepticism regarding the applicability of the design-based approach to macroeconomics: "What the essay says about macroeconomics is mainly nonsense" (Sims, 2010, p. 59; see also Leamer, 2010, p. 44; Stock, 2010, pp. 89-92). Guido Imbens (2010, p. 401), a proponent of the design-based approach, also doubts that macroeconomics could greatly benefit from turning to experiments. These claims are made for the whole of macroeconomics but they seem to apply a fortiori to my class of questions.

Widespread use of randomized controlled experiments at the macro-level is implausible. To establish an average causal effect at the country level would require taking a representative sample of countries and to randomly allocate these countries to either treatment or control group. For instance, the contrast could be a system of generous unemployment

[19] Another example in this class is the famous question of the institutional causes of long-run growth (for a methodological discussion of the economic literature on this question, see Kincaid, 2009). Other examples abound.

benefits versus a system of low benefits. Needless to say, sovereignty concerns are likely to block such proposals. Nevertheless, RCTs might still be run for some macro-level causal issues belonging to my class of questions. For instance, since donor countries often control international aid, the average effectiveness of aid could be assessed by randomizing the funds across recipient countries. Obviously, humanitarian concerns might replace sovereignty issues here. In any case, the point is that extensive use of macro-level RCTs is implausible.[20] 'Run RCTs' is thus not so helpful as a methodological recommendation when dealing with macro-level causal questions.

The situation is similar for natural experiments. It is surely not altogether impossible to find valid instrumental variables to answer macro-level questions but one should not have too high expectations. If one focuses first on the country-specific causal effect of some institutional change,[21] the core problem is the endogeneity of policy choices:

> It is hard to imagine a real-world quasi-experiment in which a central bank changed its monetary policy rule for reasons not rooted in prior macroeconomic conditions and expectations of macroeconomic benefits that would flow from the change. (Stock, 2010, p. 91)

For a natural experiment, one would need to find some source of variation in the policy variable, which is unrelated to potential benefits. Again, the possibility of finding such exogenous source cannot be ruled out, but there are reasons to be skeptical of the wide applicability of the method.

The problem of endogeneity of policy choices morphs into a self-selection problem at the macro-level when the causal claim shifts from being country specific to being about the population of countries. I will say more on that in the next subsection. But the same point holds: credible instrumental variables are likely scarce.

[20] Macro-level RCTs should not be conflated with 'Thatcher's experiment' and the like. The latter do not involve the design of treatment versus control groups and the causal claim that they can support is, strictly speaking, country specific. They might still be one source of information among many when one relies on evidential variety.

[21] James Stock (2010, pp. 89-91) makes a distinction between three types of macro-level questions. The two types which matter for us here are shocks (e.g. technology shock) and institutional change. Stock claims that the design-based studies hold more promise for the first sort than the second. The causal question under study here is of the second sort. Note that Stock focuses implicitly on country-specific causal questions.

In short, the general limitation of the design-based approach might be more often binding when considering macro-level questions and especially the subclass on which I focus.

2.4.2 Credible theory?

In the structural approach, the credibility of causal inference is affected by the credibility of the theory used to impose identifying restrictions. I do not aim here to assess the overall credibility of 'economic theory' in all its various forms. I simply want to point out that there are reasons to doubt that the structural approach is currently able to fulfill its promise for a vast range of macro-level causal questions.

Before focusing on my class of questions, I want to emphasize that the outcome might be different if I were to focus on country-specific causal questions. After all, some prominent macroeconomists were confident enough to say, at the onset of the recent economic crisis, that "[t]he state of macro is good" (Blanchard, 2008) and that we have experienced a "convergence in macroeconomics" (Woodford, 2009). The convergence is mainly accounted for by the extensive use of dynamic stochastic general equilibrium (DSGE) models. These models are structural models which have now reached such a state of development that they can be estimated with data and then used for policy analysis. In other words, estimated DSGE models are clear instantiations of the structural approach applied to macro-level causal questions.[22]

The reader can make her own assessment of the credibility of DSGE models. They at least supply evidence that the macro-level is not totally incompatible with the structural approach.

When one shifts the focus to causal claims about the population of countries, the central challenge faced by the structural approach is that of modeling the policy choices themselves. Why is it that the Netherlands has more generous unemployment benefits than the U.S.? Policy-makers do not select policies on a whim. The reasons why the Dutch and the U.S. policy-makers made these specific institutional choices might well be related to characteristics of their countries which are causally connected to our outcome variable, the aggregate unemployment rate. For instance, Dutch policy-makers might expect that, given the specific features of the Dutch economy, having generous unemployment benefits would not detri-

[22] And DSGE models are only the most recent incarnation of macro-level structural econometrics. One can think, for instance, of the giant Keynesian models from the previous era.

mentally affect their aggregate unemployment. We thus face at the macro-level the problem of self-selection of units into different treatments—now the policy-makers selecting for their country. As before with the example of school choice, controlling for self-selection is needed to avoid reaching the wrong causal conclusions.

Modeling self-selection at the macro-level is a rather different business than modeling self-selection for individual choices. What needs to be captured here are the causes of policy choices. It is not that economists do not try to model the messy world of politics; there is indeed a field labeling itself 'political economy' (Weingast and Wittman, 2008) which does exactly that—a somewhat modernized version of public choice. Perhaps this field will succeed one day in establishing some benchmark model—a model that could be used to control for self-selection by researchers who are not primarily interested in the causes of policy choices but rather in, say, the causal effect of employment protection. This, however, remains an unfulfilled promise.

In sum, there are reasons to doubt that researchers concerned with macro-level causal questions—and especially questions about causal effects for the population of countries—can exclusively rely on the design-based approach or the structural approach to increase the credibility of their inference. It is not that using these approaches is hopeless for any macro-level question but rather that their applicability cannot be assumed whatever the question. It is plausible that sharp designs and credible theories will not be forthcoming for the causal question that one happens to be interested in.

2.5 A third source of credibility: evidential variety

When both the general limitation of the design-based approach and the one of the structural approach are binding, are researchers doomed to make implausible inference? One source of optimism comes from the possibility that sharp designs and solid theories are not the only two sources of credible causal inference. By looking at the practice of economics, one sees that at least another source is relied on: evidential variety.

Relying on a variety of evidence is a common practice in science—as well as elsewhere (e.g. in law; Haack, 2008). Two examples outside economics might help see the point. The paradigmatic example of evidential variety is Jean Perrin's measurement of Avogadro's number by 13 different methods (Perrin, 1913; Nye, 1972; Salmon, 1984, pp. 213-27; 1998, p. 87; Weber, 2005, p. 281; Woodward, 2006). Perrin did not only use this impressive concordance of results as evidence for the correct value of Avogadro's number but also as supporting the proposition that atoms exist. A contemporary example comes from climate science. As evidence for an increase in global mean surface temperature, climate scientists use conventional thermometers scattered around the globe as well as other, more indirect, sources—for example, average glacier length, ocean heat content, tree rings, ice cores and satellite data (Dessler and Parson, 2010, pp. 61-81). Since these different means of measurement give approximately the same picture for the twentieth century, the proposition that the earth climate has been warming is said to be strongly supported (Oreskes, 2007; Dessler and Parson, 2010, p. 81). In short, we have measurement robustness (Woodward, 2006, sec. 6) for both Avogadro's number and global mean surface temperature.[23]

While there are many ways to think about evidence as being diverse, I want to use 'evidential variety' in a narrow sense. First, I take evi-

[23] The intuitive appeal of evidential variety is such that it has received a wide range of labels in the methodological and philosophical literature. The term 'evidential variety' is found in the Bayesian literature in the philosophy of science, in which the 'variety-of-evidence thesis' is debated (e.g. Earman, 1992, pp. 77-9; Wayne, 1995; Bovens and Hartmann, 2003, chap. 4; Novack, 2007). The term 'robustness' stems from William Wimsatt (2007b) and has been used by other scholars like Sylvia Culp (1994, 1995) and Jacob Stegenga (2009). While Wimsatt used the term broadly, Culp and Stegenga understand it in the more restrictive sense that I use. To avoid confusion, I thus use 'measurement robustness' (Woodward, 2006) to refer to this restrictive meaning (see footnote 3 above for the other concepts of robustness in Woodward, 2006). Another term close to 'evidential variety' is 'independent determinations' (e.g. Weber, 2005, pp. 281-7). The problem with this term is that it gives too much importance to the problematic concept of 'independence' (see footnote 26 below and the accompanying text). Yet another term is 'consilience of evidence', which is used exactly in line with my analysis by Naomi Oreskes (2007, pp. 89-91) in her discussion of the results of climate science. However, Oreskes' only reference to the contemporary literature using the term is Edward Wilson's (1998) book *Consilience: The Unity of Knowledge*. Wilson uses "consilience of evidence" interchangeably with "unification" and he is specifically arguing that the humanities should be integrated with the sciences. This project is quite far from mine. Finally, let me note that Nancy Cartwright's (2007, p. 25) discussion of methods "that merely vouch for the conclusion" and especially her presentation of "mixed indirect support" (Cartwright, 2007, pp. 36-37) have much in common with my own project. However, I note that what Cartwright sees as methods that 'clinch' their results will usually only vouch them because of uncertainty regarding the underlying assumptions (Hoover, 2009, p. 494); in this case, evidential variety will still be worth seeking.

dence to be always evidence *for* a specific proposition. Such a proposition can be quantitative—e.g. 'Avogadro's number is approximately equal to 6×10^{23}'—or qualitative—e.g. 'global mean surface temperature has been increasing in the last century'. The propositions in my case study below will be about qualitative causal effects. The general message though is that evidential variety is drawn on for the purpose of confirming specific propositions. In contrast, one can find scholars arguing that we should seek a diversity of evidence for the purpose of better understanding a complex object of study. Some of the scholars using the term 'triangulation'[24] voice this proposal: "Inquirers are ... using triangulation ... as a means of enlarging the landscape of their inquiry, offering a deeper and more comprehensive picture" (Tobin and Begley, 2004, p. 393). While I have no principled objection to such a project, it is not the one that I pursue here.

Second, I wish to consider variety with respect to "first-order evidence" for the target proposition, not "second-order evidence" (Staley, 2004, p. 469). The distinction amounts to the following. First-order evidence is what we naturally think of when we talk about a result being evidence for a proposition. For instance, an upward trend in readings of thermometers around the globe is thought to be (first-order) evidence for surface warming. But the claim that such result is evidence of warming is open to doubt. An actual source of doubt in this case is the possibility that the upward trend is due to the urban heat island (UHI) effect—i.e. temperature tends to be higher in cities and an increasing proportion of

[24] In the writings of Donald Campbell and collaborators (e.g. Campbell and Fiske, 1959; Webb et al., 1966), the term triangulation was used in a manner compatible with my own analysis. The term then broadened under the influence of Norman Denzin (1978, chap. 10). Denzin maintained that a study uses triangulation if it varies data, investigators, theories or methods—only method triangulation was in Campbell's writings. Note that the term triangulation is in widespread use in some fields—e.g. in nursing science (Thurmond, 2001; Tobin and Begley, 2004)—and Paul Downward and Andrew Mearman (2007) attempt to bring it in economics. I however refrain from using the term for two reasons. First, as said in the main text, some authors use it to mean approaches quite unlike what I have in mind. Second, the metaphor of 'triangulation' evokes an image which does not fit my use of evidential variety. Triangulation refers to the determination of the location of one point by using information on the location of two other points and the angles of the triangle formed by these three points. The problem with the analogy is that it suggests that using at least two means of determination is *necessary* to estimate the property of interest. But it is not. What is understood here by evidential variety is that each means of determination gives *by itself* an estimate of the property of interest. The combination is worthwhile only because one has doubts about the reliability of any estimate.

thermometer readings comes from urban areas because of a global urbanization trend. Second-order evidence for surface warming (the target proposition) can come at this point: climate scientists provided evidence that the bias due to the UHI effect can only be extremely small (see IPCC, 2007, pp. 243-45). Such second-order evidence lends support to the target proposition because one source of potential unreliability of the first-order evidence is ruled out.

Strictly speaking, second-order evidence is thus also evidence for the target proposition, but there is an asymmetry with respect to first-order evidence. Learning about higher thermometer readings should affect our strength of belief about surface warming even in the absence of evidence that the UHI effect can only be small. In contrast, learning the insignificance of the UHI effect on *potential* thermometer readings should have no effect on the support of the target proposition if there is actually no data of thermometer readings to draw on. In other words, for second-order evidence to be evidence for the target proposition, the first-order element to which it relates should be available but not the other way around.[25]

When second-order evidence is provided, there is no reason to refrain from saying that the variety of evidence increases. But, as I said, use of evidential variety here focusses on first-order evidential variety. As an aside, I can however briefly mention how second-order evidential variety could relate to the design-based or structural approaches. In any application, the reliability of these two approaches hinges on specific assumptions. The credibility of their results can thus be boosted by providing evidence that these assumptions are not violated.

Why would evidential variety—read ‘first-order evidential variety’ for now on—lead to more *credible* inference? The intuitive argument is that it is highly improbable that multiple means of determination would give concordant findings if this result is in fact mistaken. But this argument is unsatisfying if it is not supplemented by an explication of what one understands by the variety of means. Why is it that running the same regression 10 times—same data, same equation(s), same estimator, same statistical package—does not make an inference more credible? Intu-

[25] Note that my discussion of first-order versus second-order evidence departs somewhat from the discussion of Kent Staley (2004, p. 469): “If some fact E constitutes first-order evidence with respect to a hypothesis H, then it provides some reason to believe (or indicates) that H is the case.” According to me, this definition fails to capture the essential distinction because second-order evidence also “provides some reason to believe that H is the case”, given the proviso that the first-order evidence to which it relates is known.

itively, it is because these 10 runs are not varied enough.

In the philosophical literature on evidential variety, a condition of 'error independence' is generally taken to be necessary for measurement robustness to make an inference more credible.[26] Since the 10 runs in the regression example share almost all potential sources of error—e.g. measurement error in the data, missing confounders in the equation, undesirable small-sample properties of our estimator, mistakes in the code of the statistical package—we are not tempted to give any epistemic weight to the robustness of the result. In other words, if any of these errors influenced the result of the first regression, it will also influence subsequent results. In contrast, using means of determination with independent sources of potential errors increases the credibility of inference: "if all of the measurement procedures produce nearly the same result, it seems unlikely that the result is fundamentally wrong, since this would require an implausible alignment of all of these independent errors." (Woodward, 2006, p. 234)

But the error-independence condition should not be taken too strictly. It is hard to conceive of any two means of determination, which would be fully error independent. It seems, for instance, that the reliability of all (empirical) means of determination rely on some fundamental ontological beliefs about the structure of space and time and so forth. It is also misguided to think that the degree of error independence is assessed by comparing the lists of *all* assumptions on which the reliability of different means hinges. We have no access to such exhaustive lists. The notion of error independence, like the notion of credibility, should rather be understood from the perspective of the epistemic agent. For a given means, the agent holds beliefs about some possible sources of errors and typically has in mind a short list of the most likely culprits. It seems highly plausible that the credibility boost from measurement robustness is higher when the degree of overlap in these short lists of suspects is smaller.

2.5.1 Evidential variety in the macroeconomics of unemployment

The cases of evidential variety in economics are not as impressive as Perrin's multiple measurements of Avogadro's number, but they are nonethe-

[26] Error independence might, in fact, not be a necessary condition for measurement robustness to lead to higher credibility of inference. See chapter 4 for my attempt to use a Bayesian framework to shed some light on this issue.

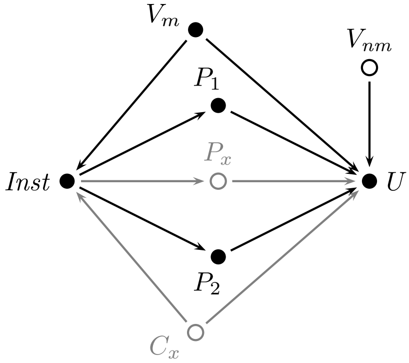


Figure 2.2: *Two identification strategies with potential sources of error*

less genuine. They point to a third source of credibility for inference to be distinguished from sharp designs and structural models. While these three sources should not be taken as mutually exclusive, it is relevant to distinguish evidential variety from the other two sources because it can be relied on in epistemic contexts in which sharp designs and credible structural models are absent—i.e. when the two established approaches face their own limitations.

The example used in what follows—relying on evidential variety to infer the causal effect of institutional variables on the aggregate unemployment rate—is a rational reconstruction of the scientific practice. In particular, economists working on this issue do not use the typology that will be introduced here. It is, however, clear from reading the literature on the topic—especially, survey papers (e.g. Blanchard, 2006) and advanced textbooks (Cahuc and Zylberberg, 2004; Boeri and van Ours, 2008, e.g.) which present how the various research efforts fit together—that they distinguish between different means of determination along similar lines.

In the macroeconomics of unemployment, evidence comes basically from two means of determination. Borrowing the terminology of Morgan and Winship (2007), the two strategies at work can be called the conditioning and the mechanistic strategies. Each strategy is meant to identify the causal effect of interest but there are reasons to doubt their reliability. Since the most likely source of error of the conditioning strategy is not shared by the mechanistic strategy (and vice versa), one can make a robustness claim when results of the two means of determination concur. The difference between the two means of determination can be sharply seen from Figure 2.2.

The causal effect of interest is from the variable *Inst*—which is a generic variable standing for different policy levers like the level or duration of unemployment benefits, the minimum wage, the strictness of employment legislation—to the variable U—i.e. the aggregate unemployment rate. Given the causal graph (abstracting from the gray objects for now), two identification strategies are open to us. The first is the conditioning strategy, namely measuring the probabilistic dependence between *Inst* and U conditional on V_m—the measured control variables. The second strategy, the mechanistic strategy, instead uses the intermediate nodes P_1 and P_2 to measure the causal effect as it propagates through the different paths connecting the cause to its effect.[27]

The graph also shows in gray a potential source of error for each means of determination. The conditioning strategy might be unreliable if some unmeasured confounder C_x exists.[28] In the case of the mechanistic strategy, error can occur if the measured paths do not exhaust the paths connecting *Inst* to U—i.e. if one fails to acknowledge the existence of some P_x. A few more words on each strategy are offered in the next two subsections.

The conditioning strategy

Concretely, this strategy uses cross-country regressions to study the relationship between the aggregate unemployment rate and diverse variables capturing various dimensions of labor market institutions (e.g. Nickell, 1997; Blanchard and Wolfers, 2000; Belot and van Ours, 2001; Nickell et al., 2005; Bassanini and Duval, 2006). The OECD supplies most data here—a measure of unemployment rate harmonized across countries and diverse measures attempting to capture institutional dimensions like the strictness of employment protection and so on. Researchers using this strategy specify a regression equation in which they include other regressors beside the one(s) of interest in the hope that, by so doing, they eliminate confounding.

We see here an important similarity with the structural approach: researchers try to control for confounding in observational data, not cir-

[27] Those familiar with the work and terminology of Pearl (2009) will recognize that the conditioning strategy is the application of the back-door criterion while the mechanistic strategy uses the front-door criterion.

[28] The strategy *might* but *need not* be unreliable. Especially when the causal effect of interest is qualitative, a real common cause might have too weak a contribution to result in erroneous inferences. In general, one might even think that causal models of social systems always leave some confounders out, and that the credibility of the inference thus hinges on a belief that these left-out confounders make negligible causal contributions, not that there is no confounder.

cumvent it by finding an instance of Nature's randomization. There is however a crucial difference between such a strategy and the Heckman's structural approach: there is no attempt to model how different countries come to have different institutional arrangements in the first place—i.e. how self-selection by policy-makers occurs. In commenting on the loose specification used in a related literature—the growth regressions—Deaton (2010, p. 433), a sympathizer of the structural approach, concludes that "the analysis ... is not a Cowles model at all". The same holds here: the selected equation is 'theoretically informed' in the sense that one has to choose which regressors to include and some theoretical propositions might affect these choices, but the equation is not anywhere close to a full-fledged economic model.

Researchers using this strategy do not stick to loose specifications because they are unaware of the potential for self-selection bias. They uniformly recognize the problem. They are especially worried that policies are influenced by the unemployment figures themselves—i.e. a reverse causality. A typical story is that policy-makers face public pressure to offer a stronger safety net when the unemployment rate is high.

One sees from this story what would need to be modeled: the factors influencing the degree of public pressure for a given policy and the factors making this public pressure more or less effective at shaping the policy decision itself. Having a *credible* model of this whole process might improve the credibility of the conditioning strategy, but the great challenge is to come up with such a credible model.

Thus far, economists using the conditioning strategy have remained content with their loose specifications. They do not base their investigation on a structural model which would probably be deemed implausible by the community. Shying away from structural models does not make the problem of confounding vanish. Reverse causality remains a possibility, as well as the existence of unmeasured common causes like the C_x in Figure 2.2. Because of these possible sources of confounding, those running the regressions warn us not to put too much epistemic weight on them. In short, if the literature on the institutional determinants of aggregate unemployment was limited to the conditioning strategy, causal inference would likely remain implausible.

The mechanistic strategy

Researchers can fortunately rely on a different means of determination, a strategy quite unlike the conditioning strategy. To make sense of this mechanistic strategy, a few comments on the meaning of 'mechanistic' are in order—unfortunately, Morgan and Winship (2007, especially chap. 8) are highly ambiguous in their use of the term (Weber and Leuridan, 2008, pp. 199-200).

The adjective 'mechanistic' is sometimes used to mean nothing more than a causal chain. The abstract representation of the mechanistic strategy in Figure 2.2 – i.e. a path from X to Y intersecting some P_i – encourages this interpretation. This reading would be compatible with relating only macro-level variables—what Harold Kincaid (1996, p. 179) calls "horizontal mechanisms". For instance, one could think that the effect of some *Inst* on U is mediated through the average wage rate W. The 'mechanistic' strategy would thus amount to estimating a path model using aggregate data for *Inst*, W and U.

While one could, in principle, attempt to measure the causal effect of *Inst* on U by drawing on such macro-level causal chains, economists working on aggregate unemployment proceed to a decomposition of the labor market into component parts—mainly, firms, workers and jobless. They thus investigate how modifying some institutional variable, say the strictness of employment protection, affects the behavior of labor market participants and how this change in behavior affects the aggregate unemployment rate. Since each institutional variable plausibly affects labor market participants in multiple ways—or affects multiple labor market participants—economists divide the total effect into different potential causal paths, each path generating a *partial* causal effect of the institutional variable on the unemployment rate. If one manages to take into account *all* the paths, the total causal effect can be recovered.

In sum, the 'mechanistic' element in this strategy is not only referring to the causal chain but also to the notion of decomposing a system into component parts—such decomposition procedure is central to the discussion of mechanisms in contemporary philosophy of science (e.g. Machamer et al., 2000; Bechtel and Abrahamsen, 2005; Kuorikoski, 2009; Hedström and Ylikoski, 2010)

The traditional way in economics to investigate lower-level mechanisms is with rational choice models. For the economics of unemployment, the benchmark is now the Diamond-Mortensen-Pissarides (DMP) model which presents the labor market as a search, matching and bargaining process (Pissarides, 2000). From the model, economists distinguish between different 'effects' of a given institutional variable on aggregate unemployment; each one of these effects captures what is meant here by the partial causal effect of a path P_i. For instance, researchers distinguish between the layoff effect of employment protection—i.e. firms would tend to fire workers at a lower rate when protection is stricter—and the hiring effect—i.e. hiring rates would also decrease because firms take into account the future firing costs (e.g. Boeri and van Ours, 2008, p. 17).[29]

On top of what one might call 'model evidence', the mechanistic strategy draws on micro-data studies—i.e. studies using firms or workers as units to investigate how their behavior and labor market outcomes are affected by a given institutional variable. These studies provide evidence for the causal paths present in the rational choice models.

Interestingly, many of these micro-data studies are pure examples of the design-based approach either of the RCT or the 'natural experiment' variants. Sharp designs are indeed available at this level because either governments are willing to run randomized controlled experiments on a segment of their population—for instance, treating a sample of job seekers with different schemes of unemployment benefits (see Meyer, 1995)—or have enacted a policy that is, by design, not applied uniformly to the population—e.g. the 1989 Austrian reform of the unemployment benefit system (Lalive et al., 2006), or the 1990 Italian reform of employment protection (Kugler and Pica, 2008).

While these micro-data studies can count on sharp designs, they do not offer a direct answer to the macro-level causal question. It is when extrapolating from the causal claims established at the micro-level to a claim about the average causal effect across countries that the mechanistic strategy can get into problems—what is usually labeled the problem of external validity. One major source of worry is that the effects measured with micro-level data might miss some causal paths that become only significant when the policy is scaled up to an entire economy.[30] For instance, a change in employment protection applied to a fraction of the

[29] According to the models, there would also be two more intricate effects of employment protection on the wages, one pushing it up while the other down. Higher wages in turn translate, in the model, into higher unemployment.

[30] In economics, these effects are often labeled general-equilibrium effects. Outside economics, they are typically referred to as failure of the Stable Unit Treatment Value Assumption (SUTVA; see Morgan and Winship, 2007, pp. 37-40; Imbens and Wooldridge, 2009, pp. 13-14). Note also that, to get from the micro-level causal effect to the average macro-level effect, the extrapolation can be broken into two: first the micro-level claim must hold at the country level, second the claim must hold for the population of countries. The failure of SUTVA applies to the first step; the second step is problematic because of causal heterogeneity across countries. The discussion in the main text focuses on the first problem which is enough to establish the main point: it is likely that the inferential base provided by the mechanistic strategy alone is not credible enough.

firms might not have substantial wage effects compared to when it is scaled up to the whole economy because the policy will be internalized differently in the wage bargaining process. This problem is represented in Figure 2.2 by the overlooked P_x—here the process of wage setting.

The main point is not whether this particular example of scaling-up effect is plausible. It is rather that the credibility of the inference relying on the mechanistic strategy—even though this strategy can build on sharp designs at the micro-level—is compromised by the gap between what is established in the studies and what one wants to establish. In consequence, the situation parallels the one of the conditioning strategy: if the mechanistic strategy was the only evidential base for the macro-level causal question, the implausibility of the inference might follow.

Combining the strategies

The end of the story is that, luckily, the evidential elements from the conditioning strategy and from the mechanistic strategy sometimes concord. This convergence can be illustrated with the case of the causal effect of the strictness of employment protection on U, which is null on average according to the experts on the question. On the side of the conditioning strategy, the parameter for the relevant variable is typically found to be not significantly different from zero. When one turns to mechanistic evidence, the finding is that the relevant mechanisms come in pairs with mutually canceling effects—e.g. hiring versus layoff effects. The general picture is thus that employment protection induces quite some change in the labor market but that the net effect for aggregate unemployment is null.[31]

The two strategies taken together offer a stronger inferential base than the two strategies taken on their own. This boost in credibility comes from the fact that the main reason why one strategy would be unreliable is believed not to be shared by the other strategy. Economists can thus make a robustness claim on the ground that their two means of determination are fairly error independent. And, indeed, one finds in the literature that there is a broad consensus around causal claims jointly supported by the

[31] Strictness of employment protection is believed to have significant effects on other relevant variables. For instance, an increase in strictness is thought to increase average unemployment duration and to decrease flows in and out of unemployment. These two results are tightly linked with the main result of zero net effect on aggregate unemployment: if a policy decreases flows in and out in the same proportion at any unemployment rate, the unemployment rate will stay put but the average duration of a spell will increase.

two strategies. Economists seem to find these inferences credible.

A caveat must follow this result. We have here a case of measurement robustness which is obviously far weaker than Perrin's measurement of Avogadro's number. One ought to remember that we have only two means of determination in this case.[32] On top of that, one must also take into account the degree of doubt about the reliability of each means. If the two means are akin to two coin tosses—i.e. processes not at all related to the property of interest—the fact that we get two heads should in no way be relevant to our belief in a causal claim. Similarly, the fact that the two identification strategies agree can be imputed either to the fact that they both more or less track the property of interest or to pure chance.

While this caveat is important, the methodological results of this section are in no way affected by it. First, evidential variety is an appropriate tool to increase the credibility of an inference. If we are unsure whether each means of determination is reliable or not, and if some of the main reasons of unreliably are means-specific (i.e. some error independence), having concordant measurements from different means is conducive to more credible inferences. Second, evidential variety is used by scholars working on a specific set of macro-level causal questions and, in this particular epistemic community, agreement between means of determination seems to be interpreted as lending more credibility to inference.

2.6 Conclusion

There is a debate in economics over the most promising approach to causal inference. One camp claims that the locus of credibility is the study design—randomized controlled trials being the gold standard. The other camp maintains that structural models drawing on economic theory should continue to play the central role of controlling for confounders in observational data—especially the self-selection problem. Good enough, but what if, for some causal questions, neither sharp designs nor established theories can be relied on? I have argued that there was no need to despair, that evidential variety might provide the credibility that

[32] Note that the individuation of means of determination is arbitrary to some degree. For instance, the mechanistic strategy could have been divided into model evidence and micro-data evidence. Individuation must certainly be guided by the error-independence criterion—in this respect, the existence of a P_x is a potential source of error for both the model and micro-data evidential elements—but this guidance still leaves us leeway in our individuation choices.

economists strive for.[33] The literature on the institutional determinants of the aggregate unemployment rate is an *in vivo* example of evidential variety.

One thing must be clear. While evidential variety can be a source of credibility independently of the availability of sharp designs or solid theories, seeking varied evidence is compatible with the design-based and the structural approaches. When sharp designs are available for the causal question of interest, it would be foolish to overlook them. When strongly confirmed theories are on offer, the causal inference should take profit of them. If we are lucky, we might have evidential elements from a randomized controlled trial directly applicable to our causal question and an estimated structural model relying on a widely acclaimed theory. If we are so lucky, nothing should stop us from combining these two means of determination in a robustness argument.[34] Credibility can only be served.

Highlighting the possibility and the actual practice of evidential variety is nonetheless important because it gives us a sense of the richness of our epistemic toolbox. We get a messier picture of causal inference but a more encouraging picture too.

[33] We are left with epistemic contexts for which we have poor theories, fuzzy designs and, on top, discordant evidence. There might not be another strategy to employ in such contexts beyond the simple recommendation to search harder for good theories, sharp designs and reasons to rule out some evidential elements.

[34] Imbens (2010, pp. 418-9), a proponent of the design-based approach, has a nice discussion at the end of his recent methodological article asking what one should do when in possession of evidential elements from a RCT and from an observational study using a structural model. His proposal is directly interpretable in terms of evidential variety. He even goes beyond measurement robustness to assess what one should do in the case of *discordant* evidential elements.

Chapter 3

The Russo-Williamson Theses in the Social Sciences: Causal Inference Drawing on Two Types of Evidence

3.1 Introduction

In macroeconomic research, a widespread strategy is to investigate how macroeconomic phenomena can result from the behavior of agents. Taking the case of aggregate unemployment, economists do not restrict themselves to measure the cross-country dependencies between the unemployment rate and diverse measures of labor market institutions—e.g., minimum wage, stringency of employment protection and generosity of unemployment benefits—they put a lot of energy into modeling and empirically assessing how changes in these institutions affect the behavior of individuals and firms. Their analysis is thus filled with statements about the mechanism through which a change in an institutional variable would affect the decisions of agents, in turn leading to a modification of the aggregate unemployment rate.

The strategy of using mechanistic evidence in addition to evidence about higher-level dependencies is not peculiar to economics. Indeed, Federica Russo and Jon Williamson (2007, p. 159) argue, based on their

*This chapter is now published online as François Claveau (2012), "The Russo–Williamson Theses in the Social Sciences: Causal Inference Drawing on Two Types of Evidence," *Studies in History and Philosophy of Biological and Biomedical Sciences*. The present version includes slight modifications from the published article.

study of inferential practices in the health sciences, that to "establish causal claims, scientists need the mutual support of mechanisms and dependencies." This claim has been dubbed the Russo-Williamson Thesis (e.g., Gillies, 2010; Illari, 2011; Reiss, 2011b). Since Russo and Williamson defend more than one thesis in their 2007 article, I label this one the *first* Russo-Williamson Thesis: In science, a causal claim can be established only if it is jointly supported by difference-making and mechanistic evidence.

Russo and Williamson further maintain that the truth of their first thesis is bad news for extant theories of causality. More specifically, what I call the *second* Russo-Williamson Thesis is that the two great families of theories of causality—the difference-making and the mechanistic theories—cannot "adequately account for the need for two types of evidence—mechanistic and [difference-making]—for a single causal claim." (Russo and Williamson, 2007, p. 164)

Russo and Williamson move on to defend other claims—i.e. pluralism about causality will not do and the 'epistemic theory of causality' is a promising alternative—but this chapter focuses only on the first two theses. I consider them in the context of the social sciences. Russo and Williamson (2007, p. 169) themselves suggest this turn to the social sciences in the conclusion of their article:

> Although in this paper we restrict our scope to the health sciences, our point about interpreting causality can be generalised to other domains. For instance, evidence in the social sciences is very diverse too.

By looking specifically at the economics of unemployment, I argue in favor of two main conclusions. First, even though the joint use of difference-making and mechanistic evidence is extremely widespread in this narrow subset of the social sciences, the first Russo-Williamson Thesis is not supported because one established causal claim is only supported by mechanistic evidence. This point hinges on the meaning of key terms such as 'difference-making evidence', 'mechanistic evidence', 'causal claim' and 'to establish', which will be clarified below.

Why do economists typically attempt to provide both difference-making and mechanistic evidence? My second point is that this practice is not at all mysterious if one thinks about causality in terms of a counterfactual-manipulationist account—an account which is popular among philosophers *and* social scientists—and is attentive to the epistemic challenges faced by social scientists. Note that this point is not a direct rebuttal of the second Russo-Williamson Thesis since it only rationalizes a *typical* research strategy in the social sciences, not a *necessary* research strategy—my first point being that this strategy does not seem necessary. It is however an important point: the social scientists endorsing the counterfactual-manipulationist account do not have to worry that their favorite (monist) account of causality is incompatible with their dualist causal epistemology. There is a compelling alternative explanation.

There is already a literature discussing these two Russo-Williamson Theses—e.g., Donald Gillies (2010) criticizes the first Thesis, Erik Weber (2009) submits that Giere's probabilistic theory of causality constitutes a counter-example to the second Thesis, and Phyllis McKay Illari (2011) sorts out ambiguities in the first thesis. The present chapter adds to this literature in three ways: it changes the focus to the social sciences; it uses a prominent account of causality which so far has not received enough attention in this literature; and it presents the issues differently by connecting the debate to the common notion of evidential variety. Beyond the direct assessment of the Russo-Williamson Theses, the chapter also contributes to the general project of better understanding how different types of evidence combine for causal inference.

3.2 The counterfactual-manipulationist account

In this section, I put forward an account of causality—the counterfactual-manipulationist account—which falls neatly into the category labeled 'difference-making accounts' by Russo and Williamson. In the following sections, I will use examples from the economics of unemployment to show that the joint use of mechanistic and difference-making evidence makes perfect sense from the standpoint of this account. If my story is accepted, it offers an alternative explanation for the fact that both health and social scientists endeavor to provide mechanistic and difference-making evidence for their causal claims. While the second Russo-Williamson Thesis suggests that one must depart from extant monist accounts of causality to explain this fact, my explanation stays closer to home. Although the reason given by Russo and Williamson is not compelling, there might well be other good reasons to reject the conterfactual-manipulationist account. Indeed, I argue in chapter 1 that this account fares badly for the semantic analysis of some causal claims. My defense of the account here should thus not be read as a full endorsement of it; I simply defend it against what I take to be a misguided criticism.

In philosophy, the best-known version of the counterfactual-manipulationist account of causality is the one of James Woodward (2003), but there are many variants of this account in the social sciences (and beyond). Many philosophers also know the work of Judea Pearl (2009), whom Woodward follows closely. The potential outcome framework (Rubin, 1974, 1990; Holland, 1986)—"now standard in both the statistics and econometrics literature" (Imbens and Wooldridge, 2009, p. 7)—is also a version of the counterfactual-manipulationist account. One should also add to the list the versions of James Heckman (Heckman, 2005; Heckman and Vytlacil, 2007) and Kevin Hoover (1990, 2001, 2011).

The shared core of these versions is the notion of a manipulationist counterfactual. All these versions agree that, for two disjoint variables X and Y, X is a (total)[1] cause of Y if and only if there is an ideal manipulation on X that changes the value of Y or its probability distribution. The worlds in which X takes a value different from its actual value are counterfactual worlds, and the notion of an ideal manipulation specifies the set of counterfactual worlds to be considered in evaluating the truth value of a causal claim.

The versions of the account listed above differ in their characterization of an ideal manipulation.[2] The underlying idea is however the same: the manipulation, which does not need to be humanly possible, should wiggle X as surgically as possible. We especially want to avoid that the manipulation is itself caused by other variables causally relevant to Y or that the manipulation causes Y by some path not passing through X.

The counterfactual-manipulationist account is rich in distinctions but I only need some basic ingredients (inspired by the potential outcome framework) for the purpose of this chapter. Define a population of units $N = \{1, ..., n\}$. For simplicity, let the cause variable be binary, i.e., for each unit $i \in N$, the cause variable can take two values $X_i = \{x_i^1, x_i^2\}$. One of these values is the actual value of X_i, the other one is the counterfactual value (the instantiation of which is determined by the particular counterfactual semantics of each version). Let Y_i be a real-valued vari-

[1] X could fail to be a total cause but still be a direct or contributing cause of Y due to canceling or to overdetermination. We don't need these other notions of cause for the purpose of this chapter.

[2] Woodward specifies an intervention variable with precise properties, Judea Pearl uses his *do* semantics, the potential outcome framework relies on the idea of a hypothetical experiment, Heckman opts for a semantics about external inputs to a causal structure, and Hoover similarly defines variation-free parameters as the locus of manipulation.

able standing for the purported effect for i. This variable is understood as taking well-defined values when X_i is in either of its two states—i.e., $Y_i = y_i^1$ when $X_i = x_i^1$ and $Y_i = y_i^2$ when $X_i = x_i^2$. The causal effect of X_i being in state 2 in contrast to state 1 for unit i is thus simply defined as $y_i^2 - y_i^1$; the average effect in the population is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i^2 - y_i^1).$$

Note that this averaging method assumes no interaction effects among units; I will come back to the relevance of interaction effects in discussing mechanistic evidence.

3.3 Establishing causal claims

In order to assess the first Russo-Williamson Thesis—i.e., both difference-making and mechanistic evidence are required to establish a causal claim—I need to clarify the meaning of a few key terms. I start in this section with what is meant by the goal of 'establishing a causal claim'.

Russo and Williamson (2007, p. 163) write that they "are not concerned, here, with how scientists came up with (controversial) causal hypotheses ... but rather with how those hypotheses have become accepted by the medical community." Likewise I focus on causal claims meeting general agreement in the relevant scientific community—the only difference being that the relevant community is now comprised of economists working on the determinants of unemployment.[3] An established claim should not be understood as one that is established *without any remaining doubt*. Science is fallible; scientists remain open to revise their strongly-held beliefs in light of new evidence. Requiring full certainty to count an empirical claim as established amounts to ruling out all claims. Like Russo and Williamson, I take a claim to be established if it is accepted as true in the relevant community.

Establishing a claim must be distinguished from the formulation of it—the action of 'coming up' with the hypothesis. I look at the evidential elements sufficient for some claims to become established, not for them to be initially formulated. Likewise, Russo and Williamson (2007) select claims such as 'smoking causes lung cancer' and '*Helicobacter pylori* causes gastric ulcers' and argue that a probabilistic association between

[3] It goes without saying that consensual causal claims make only a tiny fraction of all the causal claims that one can find in the social sciences.

cause and effect was not enough to conclude the scientific debate, it was only brought to an end when the evidential set was supplemented by mechanistic evidence.

Establishing a claim in a community is also not the same as justifying the claim. Depending on how demanding the philosopher's theory of justification is, members of a community could well agree on a claim while it is considered unjustified by the philosopher (or the other way around). In focusing on the establishment of claims, this chapter is less of a normative exercise than would be one assessing the degree of justification of causal claims. I do however believe that studying what scientists deem sufficient evidence should inform the philosophical accounts of justification.

Establishing a claim should finally be distinguished from other scientific goals. One could want to explain an already established causal claim—e.g., why is it that smoking causes lung cancer? One could also want to extrapolate a causal claim established in one population to another population—e.g., *Helicobacter pylori* causes gastric ulcers to humans but does it do the same to chimpanzees? Another possibility is to turn a causal claim into usable knowledge for policy—smoking causes lung cancer but is banning cigarettes an effective strategy to reduce the incidence of lung cancer? It is plausible that the evidence required for these goals—explaining, extrapolating, guiding policy—differs from the ones needed to establish a causal claim (Reiss, 2009, 2011b, 2012).[4]

3.4 Being precise about the claims

I introduce now three causal claims relating policy variables to the aggregate unemployment rate. The effect variable is thus the same for the three claims: the country-level unemployment rate (U) which is the ratio of unemployed workers to the total labor force. Regarding the cause, the first two claims focus on aspects of the unemployment benefit (UB) system while the third focuses instead on the employment protection legislation (EPL). Let me describe briefly these two institutions.

[4] From the standpoint of the counterfactual-manipulationist account, Russo and Williamson are not careful enough in distinguishing the different goals. For instance, they give different reasons why the "mechanistic aspect" is crucial including (1) "mechanisms explain the dependencies" and (2) they "allow us to *generalise* a causal relation" (Russo and Williamson, 2007, p. 159). The problem is obviously that even though mechanisms could be crucial for these goals, they might not be crucial for the goal of establishing a causal claim.

The UB system provides replacement income for the jobless. Economists distinguish four dimensions of the UB system: "the level of benefits, the duration of entitlement, the coverage of the system and the strictness with which the system is operated." (Nickell et al., 2005, p. 4) Only the first dimension and one aspect of the two last dimensions are relevant for the causal claims that I consider. I limit myself to discussing those aspects. The first dimension, the level of benefits, is typically reported as the 'benefit replacement ratio', i.e. the ratio of unemployment benefits to previous employment earnings. For a given country, the level of benefits varies from one job seeker to the other according to parameters such as previous earnings, family situation and length of the unemployment spell. Nevertheless, economists talk about the level of benefits of each country and do so by aggregating the benefit levels across categories of job seekers. The Organisation for Economic Co-operation and Development (OECD) is the major provider of such aggregate measures. Let me thus define a variable B_l standing for such an aggregate measure of benefit level.

Only a fraction of job seekers receive unemployment benefits because eligibility to benefits is conditional on many elements. One such element is that many countries require a minimum amount of search intensity. In other words, job seekers may have their transfers discontinued (or significantly reduced) if they fail to demonstrate that they are intensely searching for a job and that they are ready to accept job offers. I define another variable B_s standing for the strictness of the monitoring and sanctions associated to "job search and acceptance" (Blanchard, 2007, p. 415).

Now I turn to the employment protection legislation (EPL) which consists, simply put, in "regulations which make it difficult to dismiss workers." (Holland et al., 2009, p. 37) With such regulations, the employer can be required to notify in advance the worker about to be dismissed, and/or to financially compensate her. The EPL is, like the UB system, multidimensional and, like for the levels of unemployment benefits, economists attempt to reduce this complexity to a single number for each country. The OECD thus provides an overall measure of strictness of EPL. I refer to this overall strictness of employment protection by P.

We now have three institutional variables—$Inst = \{B_l, B_s, P\}$— and our effect variable U. In the relevant literature, three *qualitative* causal claims appear to be established:[5] (i) $B_l \overset{+}{\hookrightarrow} U$, (ii) $B_s \overset{-}{\hookrightarrow} U$, and (iii)

[5] For evidence that these claims are widely accepted by economists, see inter alia Blanchard (2006; 2007, p. 415) and Boeri and van Ours (2008).

$P \overset{0}{\hookrightarrow} U$. In words, (i) the level of unemployment benefits is a positive cause of U, (ii) the strictness of the UB eligibility conditions associated to "job search and acceptance" is a negative cause of U, and (iii) the strictness of EPL has no net effect on U. These claims are analogous to the claims used by Russo and Williamson such as 'smoking is a positive cause of lung cancer'—i.e. $S \overset{+}{\hookrightarrow} C$.

There is a major ambiguity in these claims which has to be resolved before moving forward.[6] Take claim (i). Using the manipulationist-counterfactual account of causality, my first approximation to the meaning of this claim is the following. Let b_l^a be the actual value of the variable and b_l^c be a counterfactual value (specified by the manipulationist counterfactual). Define also the difference in unemployment between the counterfactual state and the actual state to be $\Delta U_l = u_l^c - u_l^a$. The claim means that $b_l^c > b_l^a \Rightarrow \Delta U_l > 0$, and $b_l^c < b_l^a \Rightarrow \Delta U_l < 0$. In words, a counterfactual increase in the benefit level is associated to an increase in unemployment, and a decrease in the benefit level to a decrease in unemployment.

To which country's benefit level is the claim referring to? Is it that whatever country you take in the population of interest (say the OECD countries) the above interpretation will hold as stated? Or is it rather the weaker claim that, while some countries' unemployment may in fact be inversely affected by their level of benefits, the *average* effect holds as stated? Formally, the 'homogeneous effect' interpretation can be ex-

[6] Far from me the idea that what follows deals with the only remaining source of ambiguity. As I maintain in chapter 1, a referentialist semantics leads one to militate for the disambiguation of many aspects of such claims. Furthermore, I argue in the same chapter that a referentialist semantics might be misguided for some types of claims. It seems however that, from my temporary standpoint as a defender of the counterfactual-manipulationist account and for the purpose of this chapter, dealing with this ambiguity is sufficient. My move might perplex many readers: Is it not inconsistent to argue against referentialism in chapter 1 and to rely on it here? There is no inconsistency when one understands correctly the project of the current chapter: I want to defend the counterfactual-manipulationist account against what I see as a misguided criticism. I do think, however, that the semantic considerations of chapter 1 amount to a strong criticism of this account. A rejoinder might be that success at disambiguating the causal claims in the current section undercuts my argument of chapter 1 that a referentialist semantics cannot put the finger on their actual meaning. This objection would be sound if I were to engage in a semantic analysis which approaches the level of detail of what is attempted in chapter 1, but the reader will soon realize that, in contrasting only homogeneous and average effects, I keep the box closed. From chapter 1, we know that this box is Pandora's.

pressed

$$\forall i \in N, \quad \Delta U_{li} \begin{cases} > 0 & \text{if } b^c_{li} > b^a_{li} \\ < 0 & \text{if } b^c_{li} < b^a_{li} \end{cases} \tag{3.1}$$

while the ‘average effect’ interpretation can be expressed by

$$\frac{1}{n}\sum_{i=1}^{n} \Delta U_{li} \begin{cases} > 0 & \text{if } b^c_{li} > b^a_{li} \text{ for all } i \in N \\ < 0 & \text{if } b^c_{li} < b^a_{li} \text{ for all } i \in N. \end{cases} \tag{3.2}$$

Expression 3.1 entails expression 3.2, which makes the latter strictly weaker.

I will use the ‘average effect’ interpretation for the three causal claims. The reason why I favor the average interpretation is the same reason why causal claims about the effects of toxins (like cigarette tar) or drugs (like aspirin) on humans are most plausibly also average claims: unit heterogeneity (Dupré, 1984; Hitchcock, 2001a; Weber, 2009; Hausman, 2010). In the simple framework used here, unit heterogeneity for the effect of benefit levels means that there are at least two units i, j in N such that $\Delta U_{li} \neq \Delta U_{lj}$. In the health sciences, we find the widespread belief (supported by ample evidence) that each person reacts in her own way to a product; what saves someone could well be detrimental to another. Similarly, countries have their own cultural and institutional characteristics which could well mean that a beneficial policy for one will turn to be bad for another. In the economics of unemployment, there is thus a literature investigating institutional interactions (e.g., Boeri and van Ours 2008, ch. 13; OECD 2006b, ch. 6). Given this belief that the units of the relevant population are heterogeneous, it is implausible that an established causal claim is best interpreted as being about a homogeneous effect across units.

In sum, we have three causal claims—$B_l \overset{+}{\hookrightarrow} U$, $B_s \overset{-}{\hookrightarrow} U$, and $P \overset{0}{\hookrightarrow} U$—which I interpret as average claims applicable to a population of countries.

3.5 Two types of evidence

On what types of evidence can economists draw to support the causal claims just highlighted? The first Russo-Williamson Thesis tells us that a claim cannot be established unless it is supported by both difference-making and mechanistic evidence. What they mean by these two types of evidence is notoriously ambiguous (Illari, 2011). In this section, I propose one plausible interpretation of the distinction informed by the counterfactual-manipulationist account and show what it means concretely in my case study.[7]

The central thing to note is that the distinction makes sense only once we fixed the causal claim to be established—this is exactly why the previous section took the pain of disambiguating the causal claims under consideration. In other words, the same statement—e.g., 'W and Y are correlated'— could count as difference-making evidence for one claim and as mechanistic evidence for another.

3.5.1 Difference-making evidence

Say we want to establish that $X \overset{+}{\hookrightarrow} Y$ in a given population N, in which X and Y are (conceptually distinct) variables and the claim is meant as an average effect as explicated in the previous section. Difference-making evidence for this claim amounts to the statistical association between the two relata. There are however different qualities of difference-making evidence for the same claim. The crudest difference-making evidence would be that X and Y are positively correlated in a sample of N. Observing only this correlation would however hardly count as *strong* difference-making evidence for our causal claim because of the standard problems of confounders—other variables may be causing both X and Y—causal direction—it could well be Y that is causing X—and unrepresentative sample—there could be a correlation in the sample but not in the population.[8]

There are multiple ways to improve on difference-making evidence if one starts from this crudest correlation. At the other end of the spectrum, the counterfactual-manipulationist account tells us what would be the best difference-making evidence for a given causal claim: actually implementing the ideal manipulation. For an average causal claim, this

[7] A note on terminology. I tag the two types of evidence 'difference-making' and 'mechanistic' but, in the original article of Russo and Williamson (2007), difference-making evidence was instead called probabilistic evidence. That was an unfortunate choice of word. Russo and Williamson have now switched to 'difference-making evidence' (Russo and Williamson, 2010, 2011) and this relabeling is adopted by Illari (2011). In my interpretation of the terms, the distinction is analogous to the one of Stephen L. Morgan and Christopher Winship (2007) between 'conditioning' and 'mechanistic', which is based on Judea Pearl's (2009) distinction between the 'back-door' and the 'front-door' criteria. I indeed use in chapter 2 the label 'conditioning' for what I call here 'difference-making'. Daniel Steel's (2011) distinction between direct and indirect causal inference is also similar if not identical.

[8] There are even more reasons for correlation to be present while the causal claim is false, like non-causal association due to, for instance, non-stationarity of the variables.

means that each unit in the population would be surgically administered each value of X and the resultant outcomes Y would be recorded. There are multiple practical obstacles in our way toward this optimal piece of evidence. First, it is usually impossible to administer different values of X to the same unit. The standard response to this problem is to randomly allocate units to different values of X. This is the key idea behind randomized controlled trials. Note that the evidence produced is already departing from the optimal: in a finite sample, we will always face the possibility that the estimated causal effect is confounded. Second, it is usually impossible to be sure that the actual manipulation is indeed ideal: it can be correlated with another cause of the outcome, it can disrupt the causal structure and so on. As is well-known, even randomized controlled trials are often criticized because the actual manipulation seems to fail to be ideal (for an example of such criticism in economics, see Heckman, 1992, sec. 4). Third, the manipulation cannot always be done on the full population which brings with it problems of sample representativity. Finally, manipulating the cause is often not even an option—e.g., we cannot force people to start smoking and we cannot force states to select specific values of B_l, B_s and P. This last problem brings us back to observational data for which multiple methods have been designed to improve on simple correlation.

A brief comment before turning to the difference-making evidence available in my case study. When explicating their second thesis, Russo and Williamson (2007, p. 164) maintain that the problem with difference-making theories of causality is that they cannot "account for the fact that mechanisms are required even when *appropriate* probabilistic associations are well established." (my emphasis) The above discussion makes clear that, from the standpoint of a counterfactual-manipulationist account, the requirements for *fully appropriate* difference-making evidence are extremely demanding. To the best of my knowledge, there is no social science example for which these requirements are actually met. Furthermore, most (if not all) causal claims from the health sciences discussed in Russo and Williamson (2007) have also to rely on difference-making evidence which is far from perfect according to the standards of the counterfactual-manipulationist account—'smoking causes lung cancer' is a case in point. With the second Russo-Williamson Thesis in mind, I grant that it would be an anomaly for the counterfactual-manipulationist account if other (mechanistic) evidence for a specific claim was demanded even though *this same claim*[9] was already backed by the best difference-

[9] Note the centrality of this identity. If, for instance, one had fully appropriate

making evidence. This proposition has however little practical relevance since difference-making evidence is perhaps never of the ‘perfectly adequate’ type.

The difference-making evidence for claims relating institutional variables to the aggregate unemployment rate is far from the ideal sketched above. A central obstacle is the impossibility of overriding national sovereignty in order to freely manipulate labor market institutions. There is however a vast literature trying to get as much as possible out of the data available (e.g., Nickell, 1997; Blanchard and Wolfers, 2000; Belot and van Ours, 2001; Nickell et al., 2005; Bassanini and Duval, 2006). The common strategy is to take data from a sample of countries—their unnemployment rate (U), many of their institutional characteristics and other potentially relevant variables (e.g., GDP growth rate, inflation)—and to attempt more-or-less sophisticated regression analyses. At its simplest, the regression equation looks like:

$$U = \alpha \mathit{Inst} + \mathbf{X_m}\boldsymbol{\delta} + \varepsilon \tag{3.3}$$

where α is the parameter associated to *Inst*, the institutional variable of interest (e.g., B_l);[10] $\boldsymbol{\delta}$ is a vector of parameters for the matrix $\mathbf{X_m}$ of measured variables (including other institutional variables) meant to control for confounding; and ε is the error term which is, in part, meant to capture the effect of the *non-measured* causes of U (hopefully not correlated with the included variables).

One would like to interpret the estimate of α as the average causal effect of *Inst* on U. This interpretation could however go wrong for multiple reasons which are extensively discussed in the literature. To name just two, it is first far from guaranteed that the empirical specification deals adequately with the problem of confounding; it might still be the case that high values of *Inst* are associated to high values of U because of non-measured common causes (e.g., cultural characteristics) or because the estimated equation assumes an inappropriate (typically linear) form for the measured common causes (Belot and van Ours, 2001; Freeman, 2005, pp. 139-41). Secondly, the estimated coefficient may be biased

difference-making evidence for a claim $X \overset{+}{\hookrightarrow} Y$ for a subpopulation N_1, this evidence should be sufficient, according to the counterfactual-manipulationist account, to establish the claim for N_1. But, for obvious reasons of representativity, the same difference-making evidence would not be fully appropriate for the claim about the broader population N.

[10] Note that B_s is not included in such exercises due to lack of appropriate data. More on this below.

due to reverse causality. This problem is simple enough. While we want to interpret α as capturing *Inst* $\hookrightarrow U$, it could well be that U is the cause of *Inst*. One story making the latter case plausible for B_l and P is that policy makers react to public pressure: when unemployment is high, they increase the generosity of unemployment benefits or they tighten employment protection to soothe the electorate. If reverse causality affects substantially our estimate of α, our inference regarding *Inst* $\hookrightarrow U$ can go seriously wrong.

3.5.2 Mechanistic evidence

Now I come to my interpretation of the second type of evidence described by Russo and Williamson. Remember that I take some evidence to be mechanistic only in relation to a given claim. A causal claim is composed of two relata—cause and effect—which are described at one level. For instance, we have ‘smoking’ as an activity of a person related to ‘lung cancer’ as an illness diagnosed again for a person. As an example from my case study, we have a country’s level of unemployment benefits and a country’s aggregate unemployment rate. The first step in getting to mechanistic evidence is to redescribe the two relata at a lower level. What does it mean for a person to smoke or to have lung cancer? It means (roughly) inhaling a quantity of cigarette smoke above normal and it means having cluster(s) of cancerous cells in the lungs. Similarly (and straightforwardly), a higher B_l means that some job seekers in the country have a higher income while unemployed and a higher U means that more potential workers are out of job.

Another way of putting this is to think about causal claims as relating properties of a system—the system being a human individual in the case of $S \overset{+}{\hookrightarrow} C$ and a nationwide labor market (or economy) in the case of *Inst* $\hookrightarrow U$.[11] The first step toward mechanistic evidence is thus to pin down lower-level properties of this system which are constitutively related to the causal relata at a higher level of description (Craver and Bechtel, 2007). This is a constitutive, not a causal relation—e.g., the unemployment rate

[11] To avoid confusion, I am not maintaining that the claim $S \overset{+}{\hookrightarrow} C$ applies to a single individual. Like *Inst* $\hookrightarrow U$, it is best interpreted as an average claim in an implicit population. The truth of such claims does not imply the truth of unit-level claims such as ‘for this specific individual in the population, smoking causes lung cancer’. Nevertheless, the average causal claim is still relating properties of individuals as causal systems, namely how smoking relates to cancer *on average* in the population of such systems.

does not cause Bob and Jane to be unemployed (or the other way around), it *is* for them to be out of work.

Once redescribed, the cause and effect can be related in a different fashion than they were at the upper level. Most importantly, one can investigate how the parts of the system interact.[12] What is an effect of inhaling tar? It is to destroy "hair-like cilia in the lungs"; this, in turn, results in "[c]ancer-producing agents in cigarette smoke [being] trapped in the mucus"; the presence of these agents increase the probability of altered cells; and, finally, we get a greater chance of having clusters of cancerous cells (Russo and Williamson, 2007, p. 162). An important point for us is that the counterfactual-manipulationist account has the resources to make sense of this causal story: at any point in the chain, the causal claim can be analyzed with a manipulationist counterfactual.[13]

In the economics of unemployment the current way to mechanistically understand the causes of unemployment is strongly influenced by the Diamond-Mortensen-Pissarides (DMP) benchmark model of the labor market (see Pissarides, 2000). This benchmark model conceptualizes the unemployment rate as a resultant of the matching and bargaining dynamics between workers and employers. Apart from the decomposition between different types of agents—e.g. job seekers supplying and firms demanding labor—the model presents the unemployment rate as a stock variable—i.e., the ratio of unemployed workers to the total labor force—affected by the flows in and out of employment, i.e., by job creation and destruction.[14] In an economy with some amount of job destruction, there will inevitably be some unemployed workers at any given point because time is needed to find a new job. Let us assume, for ease of exposition, that it takes an average of one month to find a new position and that the flows in and out of employment cancel out—i.e., the unemployment rate is in steady state. The observed unemployment rate can thus be directly

[12] See the mechanistic literature in the philosophy of science as applied to biology (e.g., Machamer et al., 2000; Bechtel and Abrahamsen, 2005), to the social sciences (e.g., Kuorikoski, 2009; Hedström and Ylikoski, 2010) or to both (Steel, 2008). Steel's (2008, p. 48) definition of social mechanisms seems to fit particularly well the case of the economics of unemployment: "Social mechanisms are complexes of interacting agents—usually classified into specific social categories—that produce regularities among macrolevel variables."

[13] For recent defenses of such a counterfactual analysis of mechanisms, see Glennan (2011) and Woodward (2011). This interpretation departs from an 'actualist' interpretation of mechanisms like the one of Machamer (2004). Russo and Williamson (2007, p. 162) do not commit themselves to a particular understanding of mechanisms and clearly leave open the view endorsed here.

[14] I am abstracting here from the flows in and out of the *labor force*.

derived from the flow out of (or in) employment. If 5% of individuals in the labor force lose their job each month, the unemployment rate will be steady at 5%. From this perspective, interventions can affect the unemployment rate by changing the flows. But these paths are only half of the story; the unemployment rate is also affected by the average duration of an unemployment spell. Keeping the flows in our previous example, we can double the observed unemployment rate (from 5 to 10%) by doubling the average time it takes for an individual to find a new job (from 1 to 2 months). In light of the stock/flow distinction, institutional variables such as B_l, B_s or P can be said to be causally relevant to U in the model if a path affecting either the flows or the average duration of the spell can be uncovered.

The empirical element in mechanistic evidence comes from micro-data studies which have steadily grown in popularity in the last two decades. Many researchers now see them as the avenue to secure consensus (e.g. Freeman, 2005). Here are two typical examples. One can look at the average duration of the unemployment spell in groups of workers with different requirements for job search (linking to B_s) or different profiles of unemployment benefits (linking to B_l). Another example is to look at whether firms subject to a stricter employment protection legislation have different firing and hiring behaviors than firms subject to a weaker legislation. These studies provide direct evidence for claims such as 'in this sample, average duration increases under a given UB system' or 'in this sample, firing rates drop with stricter employment protection'. The reader should note that the direct evidence is difference-making evidence for the type of claims just stated. It is however *mechanistic* evidence for the claims that I initially posited ($\mathit{Inst} \hookrightarrow U$).

Why are economists putting so much faith in micro-data studies? Their clear advantage over the cross-country regressions discussed above is that they allow for stricter research designs (sometimes based on randomized controlled trials). With such designs, the possibility of erroneously attributing, in the sample under study, a causal role to some variable—e.g., search requirements causing average duration—is reduced. This advantage is why many labor economists seem to agree with Blanchard that "much has been learned about the effects of the various pieces" (Blanchard, 2006, p. 45) from micro studies.

The drawbacks of micro-data studies are apparent from Blanchard's quotation. They are about 'various pieces' but they do not supply the full picture. Firstly, each micro-data study typically looks at only one potential path from *Inst* to U. For example, in investigating whether higher benefits cause longer unemployment spells, a study could allocate randomly half of a sample of job seekers to a more generous level of benefits and compute the difference in duration between the two halves of the sample. By doing so, the study might miss the effect of higher benefits on the search duration of another group of job seekers, namely those unemployed but uncovered by the UB system.[15] Theoretical work has indeed highlighted the possibility of an 'entitlement effect' for those who are jobless but are not covered by the UB system (Boeri and van Ours, 2008, sec. 11.2.2). The story goes like this: these jobseekers will have a stronger incentive to find an employment quickly if the unemployment benefits are generous. This incentive comes from the fact that getting a job will plausibly make them eligible to generous unemployment transfers in the future. Thus incentivized by more generous benefits, the non-covered individuals will have shorter unemployment duration on average, in turn leading to a lower U. This is a case of path missed by a specific study.

The second drawback of micro-data studies when our target claims are average causal effects for country-level relata comes from the possibility of interaction or general-equilibrium effects (see Rubin, 1986; Morgan and Winship, 2007, pp. 37-40; Imbens and Wooldridge, 2009, pp. 13-4). The typical strategy in using evidence from micro-data studies to inform aggregate level questions is to simply blow up the sample result to the whole population—e.g. an increase in average duration of x following a localized reform is taken to be evidence that average duration would have increased by x if the reform had been national. Such an extrapolation is only warranted in cases in which there is no interaction or general-equilibrium effects—i.e. when "treatments received by one unit do not affect outcomes for another unit" (Imbens and Wooldridge, 2009, p. 13). Here's an illustration:

> It is clear that a labor market program that affects the labor market outcomes for one individual potentially has an effect on the labor market outcomes for others. In a world with a fixed number of jobs, a training program could only redistribute the jobs, and ignoring this constraint on the number of jobs by using a partial, instead of a general, equilibrium

[15] The proportion of non-covered individuals in the total unemployed labor force is larger than one might expect. For the twelve European countries listed in Boeri and van Ours (2008, Table 11.2, pp. 234-8), the coverage of the UB system—i.e. "the fraction of LFS unemployed declaring that they were receiving unemployment benefits"—is between 13% (Greece) and 53% (Netherlands).

> analysis could lead one to erroneously conclude that extending the program to the entire population would raise aggregate employment. (Imbens and Wooldridge, 2009, pp. 13-4)

The assumption of 'a world with a fixed number of jobs' is certainly far-fetched—it is, after all, called the 'lump-of-labor fallacy'—but the point remains: there is an inferential leap from the micro-data study to a claim about the effect of a national reform *for a specific country.*

Whereas the second drawback just discussed concerns the challenge of moving from the result of a sub-country study to an aggregate claim about this same country, the final drawback that I want to discuss arises in attempting to move further away, from a claim in one country to the target claim about an average effect *across countries.* As discussed in section 3.4, the effect of *Inst* on U is most likely to vary from one country to the other. Consequently, even though we might be right in extrapolating from a micro-data result in country i to a claim about the aggregate effect in the same country i, we might still be mistaken to hold that this result reflects the average effect across countries.

Summing up this section on the two types of evidence, we saw that, once a target causal claim is determined, one can distinguish between difference-making and mechanistic evidence. The counterfactual-manipulationist account of causality can obviously rationalize why difference-making evidence is evidence for the target claim, and it can (perhaps less obviously) do the same for mechanistic evidence. In the latter case, the causal relations in the mechanism are analyzed in terms of manipulationist counterfactuals, in a way exactly paralleling how the target causal claim is analyzed. This section also established a point which will be crucial in the next section in which I look directly at the two Russo-Williamson Theses: each type of evidence has potential sources of bias—mainly confounding and reverse causality for the difference-making evidence, and missing paths, general-equilibrium effects, and country heterogeneity for the mechanistic evidence.

3.6 Evidential variety and the RW Theses

What is the rationale behind the common practice of exhibiting both types of evidence for a given causal claim? In attempting to explain this practice, Russo and Williamson are led to argue that all monist accounts of causality are misguided (i.e., the second Russo-Williamson Thesis).

There is however an alternative explanation which does not threaten the counterfactual-manipulationist account of causality.

One general message of the previous section is that the available difference-making and mechanistic evidence might not be reliable. Even though these two types of evidence are far from ideal, it should also be clear that the reasons why they would err are somewhat orthogonal to each other. In other words, if one has good reasons to believe that one type of evidence is unreliable—e.g., the aggregate data are indeed driven by reverse causality—it is hard to see why our belief in the reliability of the other type should be affected. When an evidential set has this property, we can say that its elements are error independent.

The notion of error independence brings us directly to why economists (and perhaps also health scientists) try to exhibit both mechanistic and difference-making evidence for a given causal claim. This rationale is *evidential variety* or what Jacob Stegenga (2009, p. 650) recently called "the common platitude" of robustness: "hypotheses are better supported with evidence generated by multiple techniques that rely on different background assumptions".[16] Each type of evidence can be doubted because all evidence-generating methods are fallible. This doubt should however weaken if multiple, (more or less) error-independent, evidential elements support the same claim.

Let me illustrate the functioning of evidential variety by looking briefly at the evidence for my target causal claims. I start with $B_l \overset{+}{\hookrightarrow} U$. In cross-country regressions—i.e., the method generating difference-making evidence—the association of B_l with U is almost always found to be significantly positive. Moving to mechanistic evidence, most studies find that job seekers receiving higher levels of benefits have, on average, longer unemployment spells (OECD, 2006b, p. 59; Boeri and van Ours, 2008, pp. 239-40). While the quantitative effect of B_l on U extrapolated from micro-data studies is typically smaller than what the difference-making evidence suggests,[17] the two types of evidence agree qualitatively. We thus have two types of evidence with (partially) independent sources of error which agree on the qualitative result.

[16] Following Woodward (2006), it is perhaps better to call it "measurement robustness" to avoid confusion with other types of robustness. Like most platitudes, measurement robustness has received numerous labels; see the previous chapter (especially section 2.5) for a more detailed analysis and more references to the literature. For a Bayesian analysis of the concept of error independence, see chapter 4.

[17] This discrepancy is perhaps explainable by a bias in cross-country regressions due to reverse causality (Boeri and van Ours, 2008, pp. 243-4).

The evidence for $P \overset{0}{\hookrightarrow} U$ is even more interesting because it is a case of mutually canceling causal paths. Regarding difference-making evidence, the relevant coefficient is rarely found to be significantly different from zero (OECD, 2006b, p. 96; Boeri and van Ours, 2008, pp. 211-2). This regression result does not mean that employment protection has no effect on the behavior of agents. Indeed, micro-data studies tell the following story (e.g., Kugler and Pica, 2008). Stricter employment protection reduces aggregate unemployment through one path: dismissal rates of firms subject to stricter rules are lower. Theoretically, this effect is explained by the increased costs associated with layoffs. But there is a counteracting path: with a stricter legislation, firms are less willing to employ new workers, they open fewer vacancies, thus it takes longer for the average unemployed to find a job. We thus have a diminution in the unemployment inflow but also an increase in the average duration of an unemployment spell; the generic result being that the two paths cancel out leaving the unemployment rate as it is.

These two first examples should be enough to illustrate the logic of evidential variety. It is also easy to see from them that evidential variety does not clinch (Cartwright, 2007) the target causal claims. Indeed, it might well be that the two types of evidence brought in support of, say, $B_l \overset{+}{\hookrightarrow} U$ are both strongly biased but, by chance, in the same direction. The possibility of 'chance agreement' would persist even if we found other error-independent evidential elements; the probability of chance agreement would just diminish as evidential elements accumulate. So, when saying that my two causal claims are established, I do not mean that they are established beyond doubt, that no more research is required to investigate whether they might, in fact, be wrong. These claims are established in the weaker sense that, after due consideration of the evidence, no specialist rejects them. See how the prominent macroeconomist Olivier Blanchard (2006, p. 45) puts it: "From both the macro evidence and this body of microeconomic work, a large consensus – right or wrong – has emerged." Specialists still envisage the possibility that they might be wrong.

What about the claim $B_s \overset{-}{\hookrightarrow} U$? How was it established? It was established *without the aid of difference-making evidence.* This causal claim is thus my counter-example to the first Russo-Williamson Thesis. The reason why difference-making evidence was not relied on is rather trivial: there was no measure of B_s comparable across countries (and there is still none as far as I am aware). Economists drew on the available evidence, i.e., mechanistic evidence. And it seems that this evidence—

including clear model predictions and micro-data evidence in line with them (see Fredriksson and Holmlund, 2006, sec. 4)—was sufficient to gather general support to the claim.

To be more precise, the OECD provides data for the overall cost of 'active labor market policies' (ALMPs) for each country and these data are routinely used in cross-country regressions. Since monitoring and sanctions are included under the umbrella term ALMPs, it is initially plausible to say that difference-making evidence for $B_s \overset{-}{\hookrightarrow} U$ comes from the coefficient associated to the cost of ALMPs in cross-country regressions. There are however two major problems with this argument. Firstly, ALMPs include many things beside what one wants to capture in B_s—e.g., placement services, subsidized training, and subsidized employment. Given the nature of these additional components, the overall cost of ALMPs is not likely to be driven by B_s. In other words, the ALMP variable is a really bad proxy for B_s. Secondly, I know of one study (Bassanini and Duval, 2006, pp. 81-3) which disaggregates ALMP expenditures into five components, one of which—'Public Employment Services'[18]—is arguably a better (though still poor) proxy for B_s. The problem is that the coefficient on the relevant component is unstable in the three specifications used: negative and statistically significant in one, non-significant in another, and positive and significant in the last. This can hardly be taken as evidence for $B_s \overset{-}{\hookrightarrow} U$.

$B_s \overset{-}{\hookrightarrow} U$ was thus established without difference-making evidence. Two comments come to mind in connection with the idea of evidential variety. First, since the causal claim was established solely with mechanistic evidence, it must mean that the specialists are quite confident in the reliability of mechanistic evidence for this specific claim—i.e., they don't expect missing paths, general-equilibrium effects and country heterogeneity to be major problems. Second, one should not be surprised to see that, if a measure of B_s comparable across countries becomes available, economists will use it to test $B_s \overset{-}{\hookrightarrow} U$. Evidential variety will just make economists more comfortable in holding this claim.

3.7 Conclusion

Russo and Williamson (2007) report that it seems necessary, in order to

[18] To get expenditures on PES, one removes from total ALMP expenditures the cost of subsidized training, subsidized employment, youth measures and measures for the disabled.

establish a causal claim in the health sciences, that the claim be supported by both difference-making and mechanistic evidence. They maintain that the extant monist accounts of causality cannot rationalize this observation.

My conclusion, based on the study of a specific case in the social sciences, is different. Firstly, it is in fact *not necessary* in order to establish a causal claim that it be supported by both difference-making and mechanistic evidence. Among the three established causal claims analyzed here, the last one could only count on the support of mechanistic evidence. Second, the counterfactual-manipulationist account—a monist account popular in the social sciences—can perfectly make sense of the practice of jointly supplying difference-making and mechanistic evidence for a single claim. From the standpoint of this account, this practice is an instance of the common strategy of increasing evidential variety. Since we (extremely) rarely have access to fully appropriate evidence of any of the two types, it is not mysterious that scientists endeavor to collect both.

To argue for these two points, I had to offer interpretations of quite a few concepts relevant to the two Russo-Williamson Theses. Firstly, the scientific moment analyzed here is properly the establishment of a causal claim, which should be distinguished from its formulation, its justification, its explanation, its extrapolation, and its practical use. Secondly, one must be clear about the causal claim under consideration and, specifically, about what the units are and whether the causal relation is meant to apply on average to these units or rather homogeneously to each of them. Finally, only once the claim is disambiguated can we tell what counts as difference-making and mechanistic evidence *for* this claim.

Why are scientific communities often producing both difference-making and mechanistic evidence for the same causal claim? The fundamental reason does *not* seem to be that we specifically need ‘mechanisms’ and ‘dependencies’ for a claim to be well supported but that agreement among “multiple means of determination” (Wimsatt, 2007b, p. 43) increases our confidence in the claim. Reporting the two types of evidence discussed here is simply one way among others to generate evidential variety.

Chapter 4

The Independence Condition in the Variety-of-Evidence Thesis

4.1 Introduction

Seeking a variety of evidence for a hypothesis is standard practice in science, as well as in normal life. The members of the OPERA Collaboration, for instance, appealed to the value of evidential variety when they disclosed their measurement of neutrinos apparently traveling faster than light: "While OPERA researchers will continue their studies, we are also looking forward to independent measurements to fully assess the nature of this observation." (Istituto Nazionale di Fisica Nucleare, 2011)

Evidential variety is also prized in economics. For example, it is commonplace in labor economics for a causal hypothesis to be seen as more strongly supported if it can rely, not only on macro-data evidence but also on micro-data evidence. If the hypothesis under consideration is 'the relatively long duration of unemployment benefits in France is a cause of its relatively high unemployment rate', the proposition 'there is a positive statistical association between average duration of benefits and unemployment rates among industrial countries' (macro-data evidence) will be interpreted as supporting the hypothesis, but the support will be even higher if the evidential elements also include the propositions 'the average length of an unemployment spell increased in Austria for the category of

*This chapter (with a shortened appendix) is forthcoming in *Philosophy of Science*.

job seekers affected by the 1989 reform of benefits duration' (micro-data evidence).[1]

The widespread quest for evidential variety can be justified by what Bayesians call the variety-of-evidence thesis.

Variety-of-evidence thesis. *Ceteris paribus*, the strength of confirmation of a hypothesis by an evidential set increases with the diversity of the evidential elements in that set.

Some Bayesians maintain that this thesis could be given a formal proof once its key terms—the *ceteris paribus* clause, confirmation, variety—are properly defined. In seeking this proof, the most popular interpretation of variety has been to equate it to a measure of independence among evidential elements.[2] The intuitive idea behind the proposals of Earman (1992) and Howson and Urbach (1993) is that an evidential set is varied to the extent that each element e_i is *not* made significantly more likely by learning other elements in the set—the extreme case being full probabilistic independence between e_i and any conjunct of the other elements.

It turns out that one runs into problems in trying to prove the variety-of-evidence thesis using such a measure of independence. It is indeed clear from a measure introduced by Myrvold (1996) and recently labeled "focused correlation" by Wheeler (2009) that, in order to prove the variety-of-evidence thesis using the most popular interpretation of variety, one must either assume that the hypothesis entails the evidence, which would be forgetting the role of auxiliary hypotheses; or one must smuggle into the *ceteris paribus* clause the measure of independence *conditional on* the hypothesis, which seems unwarranted.

In parallel to these developments, Bovens and Hartmann introduced another characterization of variety as *reliability* independence. Using this notion of independence, they challenged the belief that Bayesianism can prove the variety-of-evidence thesis. According to their model, "less varied evidence may indeed provide more confirmation to the hypothesis" (Bovens and Hartmann 2002, 47; 2003, 106).

Bovens and Hartmann use what seems to be a plausible understanding of variety: evidential elements for a given hypothesis are varied to the extent that they do not share potential reasons for being unreliable.

[1] These evidential propositions come respectively from OECD (2006b, table 3.3) and from Lalive et al. (2006).

[2] In contrast, Horwich (1982, 1998) connected variety with the capacity of an evidential set to disconfirm alternative hypotheses. For discussions and criticisms, see Wayne (1995), Fitelson (1996) and Bovens and Hartmann (2003, 107).

For example, the ICARUS Collaboration (2012) was in a position to produce evidence for or against the hypothesis of faster-than-light neutrinos, evidence which was partially independent of the OPERA experiment. The independence is partial here because, while on the one hand, the two groups shared the same neutrino beams from CERN—making them share some potential reasons to be unreliable— on the other hand, they used different detectors—opening up the possibility that the ICARUS measurement is unbiased while the OPERA measurement is systematically biased due to a defect in the OPERA detector.[3]

This chapter takes a second look at Bovens and Hartmann's result. My primary concern is to assess whether their result should affect the status of the variety-of-evidence thesis as a guide to scientific practice. Endorsing their (plausible) interpretation of variety as reliability independence, I argue that two aspects of their model shed doubt on the relevance of their result for actual science. Firstly, the *unreliable* sources in their model are not like unreliable sources in actual science—i.e., their unreliable sources are randomly biased while systematic bias is far more likely to be the issue. I show, in section 4.4, that the variety-of-evidence thesis is rescued when the model is slightly modified to capture unreliability as systematic bias. Secondly, their model, and my first modification to it, contrast full independence to full dependence while variety in the variety-of-evidence thesis is more a question of degrees of independence. In section 4.5, I extend the model to consider degrees of independence. I then show that the variety-of-evidence thesis, as Bovens and Hartmann initially claimed, is false.

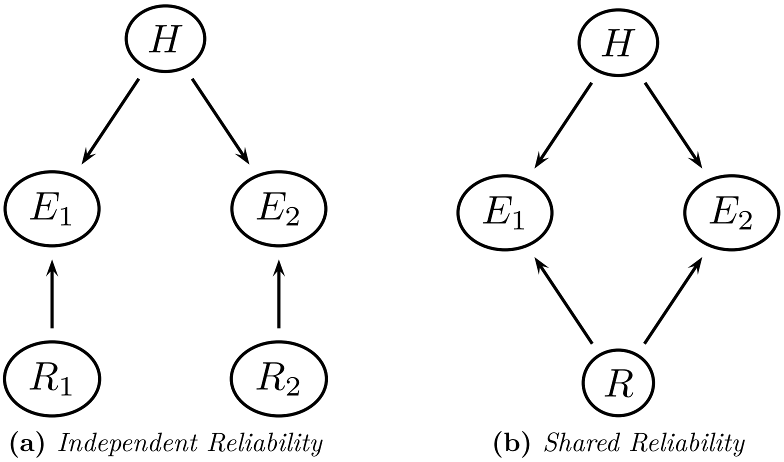
4.2 Bovens and Hartmann's result

In this section, I present a simplified version of Bovens and Hartmann's model and reproduce one of their results against the variety-of-evidence thesis.[4] The model uses three types of propositional variables:

- The hypothesis variable $H = \{h, \neg h\}$, where h stands for the proposition that the hypothesis of interest is true (e.g., 'some neutrinos can travel faster than light') and $\neg h$ stands for its negation.

[3] This is now the official explanation of the OPERA anomaly (CERN, 2012).

[4] Their model is slightly more complex because it adds another propositional variable to the three that I consider—i.e., the 'testable consequence' C (Bovens and Hartmann, 2003, 89-90). Since I focus on the issue of independent reliability—not on the issue of independent testable consequences—this addition is superfluous.



(a) *Independent Reliability* **(b)** *Shared Reliability*

Figure 4.1: *Two cases of partially reliable evidential sources*

- The evidential variable $E_i = \{e_i, \neg e_i\}$, where e_i stands for a positive report regarding h, that is, a report to the effect that a testable consequence of h holds (e.g., 'the *measured* velocity of the neutrinos in this experiment is higher than the speed of light') and $\neg e_i$ stands for a negative report.

- The reliability variable $R_i = \{r_i, \neg r_i\}$, where r_i stands for the proposition that the evidential source i (the one having as output E_i) is reliable and $\neg r_i$ stands for the proposition that the source is unreliable.

Two joint probability distributions are constructed over the set of variables $\{H, E_1, E_2, R_1, R_2\}$. The assumed probabilistic independencies among the variables can be read off the Bayesian networks in Figure 4.1 by using the d-separation criterion (Pearl, 1988, 117-18).[5] The probability

[5] The two distributions share the condition $H \perp\!\!\!\perp R_1, R_2, R$—i.e., before learning the evidential report E_i, learning that the associated evidential source is reliable or not has no bearing on the strength of belief in the hypothesis (and vice versa). For panel (a), we also have $E_i \perp\!\!\!\perp E_j, R_j | H$ for $i \neq j$, which means that once the realization of H is known, learning the realization of E_j or R_j for $j \neq i$ is not relevant to the probability distribution of E_i. A similar condition for panel (b) is $E_i \perp\!\!\!\perp E_j | H, R$ for $i \neq j$, which means that, in this case, one needs to condition on the reliability variable too in order for the two evidential reports to be irrelevant to each other. These conditions do not universally apply to what can be considered evidential elements for a hypothesis (e.g., Wheeler and Scheines, 2011, fig. 3). This fact must be kept in mind in interpreting the result of Bovens and Hartmann as well as my results.

distribution $P_I(\cdot)$ associated with the network in panel (a) is meant to capture the idea that two sources are *reliability independent*—i.e., $R_1 \perp\!\!\!\perp R_2$. The probability distribution $P_S(\cdot)$ associated with panel (b) captures the other extreme when evidential sources have *fully-shared reliabilities*—i.e., $R_1 = R_2 = R$. The notion of variety modeled here is thus:

Reliability independence. Two evidential elements are independent if their reliabilities are independent.

The joint probability distributions are further specified. The root variables H and R_i are given prior probabilities:

$$P(h) = h_0 \quad \text{and} \quad P(r_i) = \rho_i \tag{4.1}$$

where h_0 and ρ_i are parameters strictly between 0 and 1. h_0 is thus the prior degree of belief that the hypothesis is true, and ρ_i is the prior degree of belief that the evidential source i is reliable. For compactness, I will write $\bar{h}_0$ for the prior probability that the hypothesis is false $(1 - h_0)$, and $\bar{\rho}_i$ for the prior probability that source i is *un*reliable $(1 - \rho_i)$.

What remains to be spelled out is how the evidential variable E_i varies with its parents. It is assumed that, when a source is reliable, the evidential report is a perfect truth tracker:

$$P(e_i|h, r_i) = 1 \quad \text{and} \quad P(e_i|\neg h, r_i) = 0 \quad \text{for } i = \{1, 2\} \tag{4.2}$$

That is, when the hypothesis is true, a *reliable* evidential source will give a positive report; when the hypothesis is false, such a source will give a negative report.

What happens when the source is unreliable? To specify this case, Bovens and Hartmann rely on the following intuition:

Irrelevance of an unreliable source. If one knows for sure that a given source is unreliable ($R_i = \neg r_i$), the report coming from this source (e_i or $\neg e_i$) should not have any effect on the degree of belief in the hypothesis h.

This can be written

$$P(h|e_i, \neg r_i) = P(h|\neg e_i, \neg r_i) = P(h|\neg r_i). \tag{4.3}$$

In other words, an unreliable source gives garbage information regarding the truth of the hypothesis. Upon learning the information from an unreliable source, the agent makes no updating to the subjective probability of the hypothesis.

Table 4.1: *Probability of a positive report given the values of H and R_i*

$P(e_i \mid H, R_i)$	r_i	$\neg r_i$
h	1	α_i
$\neg h$	0	α_i

Note that this is not the only plausible interpretation of 'unreliable source'. The interpretation clashes, in particular, with the idea that unreliability might be due to calibration issues. Taking again the OPERA experiment as an example, some early critics claimed that the anomalous result might be due to a problem with clock synchronization (Contaldi, 2011). The estimated time of travel would be systematically below the actual time because the clock at the end of the tube clicked slightly later then the clock at the beginning.[6] With this type of unreliability, it is possible to undo the bias: given the estimated value and given the bias, one can retrieve the actual travel time. Knowing, for example, that the experimental setup is biased ($\neg r_i$) in such a way that the estimated time is systematically underestimated by a factor b, obtaining a positive report e_i that the time of travel is t should matter for one's belief in the hypothesis because an unbiased estimate of the travel time can be retrieved by computing t/b. For such calibration problems, we thus have that $P(h|e_i, \neg r_i) \neq P(h|\neg r_i)$, which contradicts condition (4.3). In this chapter, I stick to interpretations of unreliability compatible with condition (4.3), and keep the calibration interpretation for future work. I will later give examples of reasons for unreliability which are compatible with condition (4.3).

Condition (4.3) implies the following (proof in Bovens and Hartmann, 2003, Appendix C.1):

$$P(e_i|h, \neg r_i) = P(e_i|\neg h, \neg r_i) =: \alpha_i \tag{4.4}$$

Another way to express this condition is $E_i \perp\!\!\!\perp H|\neg r_i$.

Parameter α_i is the last parameter of the model; it is the probability that the evidential report is positive given that the source is unreliable. The probability that the evidential report is *negative* given that the source is unreliable is simply $1-\alpha_i =: \bar{\alpha}_i$. Table 4.1 sums up how the realizations of E_i depend on the values taken by H and R_i.

Bovens and Hartmann offer a specific interpretation of α_i in terms of a randomizing evidential source. I will later offer an alternative interpretation of this parameter; but let me first reproduce their result relative to

[6] The OPERA researchers have indeed identified similar biases by now.

the variety-of-evidence thesis. The probability of interest is the posterior belief in the hypothesis given two positive reports: $P(h|e_1, e_2) = P^*(h)$. For the two joint probability distributions $P_I(\cdot)$ and $P_S(\cdot)$—i.e., the distributions associated with the reliability-*I*ndependent version (Figure 4.1a) and the *S*hared-reliability version (Figure 4.1b)—this posterior can be written (see Appendix A.1.1)

$$P^*(h) = \frac{h_0}{h_0 + \bar{h}_0 L}, \quad \text{where } L = \frac{P(e_1, e_2|\neg h)}{P(e_1, e_2|h)}. \tag{4.5}$$

Which posterior is higher, $P_I^*(h)$ or $P_S^*(h)$? To turn this comparison into an assessment of the variety-of-evidence thesis, we need a plausible interpretation of the *ceteris paribus* condition of this thesis. Bovens and Hartmann impose restrictions that seem sufficient to meet the condition. First, we want to rule out comparing hypotheses starting with unequal degrees of confirmation. We thus impose $P_I(h) = P_S(h)$. Second, we want the evidential sets to potentially differ in confirmatory strengths for no other reason than their relative variety. A *sufficient* condition for this goal is to require that all positive reports e_i in the independent-reliability and the shared-reliability versions have the same confirmatory strength for h—i.e., $P_I(h|e_i) = P_S(h|e_j)$ for $i, j = \{1, 2\}$. This condition holds if the different α_i and ρ_i are reduced to only a single α and a single ρ across the two models (proof in Appendix A.1.3).

Given this interpretation of the *ceteris paribus* condition, the likelihood ratios associated with the two posteriors $P_I^*(h)$ and $P_S^*(h)$ are (proof in Appendix A.1.2):

$$L_I = \frac{(\alpha\bar{\rho})^2}{(\rho + \alpha\bar{\rho})^2} \tag{4.6}$$

$$L_S = \frac{\alpha^2\bar{\rho}}{\rho + \alpha^2\bar{\rho}} \tag{4.7}$$

Since we assume that the prior probability of the hypothesis is the same in the two models, the variety-of-evidence thesis implies that $P_I^*(h) > P_S^*(h)$ for all admissible parameter values—i.e., we should have a higher confidence in the hypothesis if our two positive reports come from reliability-independent sources as compared to reliability-shared sources. This is equivalent to $L_S > L_I$. It turns out, however, that the inequality is reversed for some combinations of values of the parameters α and ρ (proof in Appendix A.1.4):

$$P_I^*(h) > P_S^*(h) \quad \text{if and only if} \quad .5 > \bar{\alpha}\bar{\rho} \tag{4.8}$$

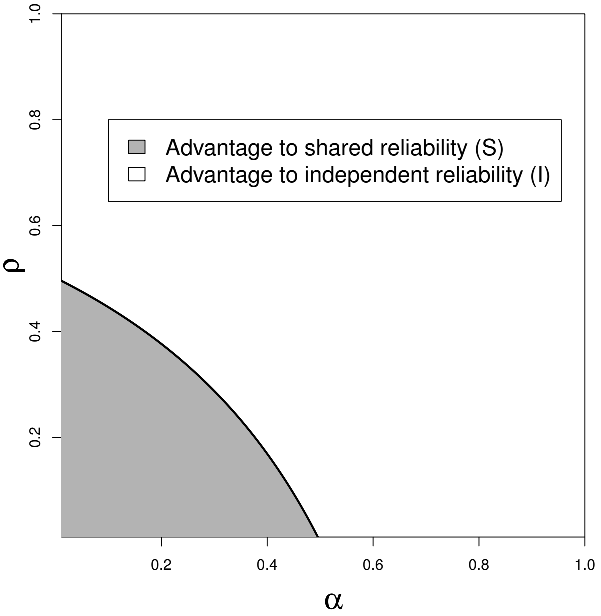


Figure 4.2: *Parameter space showing when shared reliability is more confirmatory than independent reliability*

Figure 4.2 divides the parameter space in two regions: the bigger white region where independence is more confirmatory, and the gray region where shared reliability is more confirmatory.

What happens? Why is it sometimes better for confirmation to have no reliability independence rather than full independence? To understand this, it is crucial to see what *shared* reliability entails in the second version of the model: namely that E_1 is a truth teller if and only if E_2 is a truth teller. As truth tellers, E_1 and E_2 will always give concordant reports. But when none of the evidential variables is a truth teller (i.e., when $\neg r$), then each evidential variable has a probability α of producing a positive report. It is crucial to recognize that this probability is not affected by the value the other evidential variable is realizing. It implies that when, *and only when*, they are unreliable, E_1 and E_2 might realize discordant reports. A second concordant report in the shared-reliability model thus contributes to confirmation in the following way: "we feel more confident that the instrument is not a randomizer and this increase in confidence in the reliability of the instrument benefits the confirmation of the hypothesis." (Bovens and Hartmann, 2003, 98) The region of the parameter space where shared-reliability is better is the one where this channel of 'higher confirmation of h because higher confidence in the reliability of the source' is the most effective. With low values of α, it is unlikely that an *un*reliable source would output two *positive* reports; it is thus likely that the two positive reports come from a *reliable* source. With low values of ρ, the agent starts with little confidence in the source; there is both great room for improving confidence and little to be gained for the belief in the hypothesis from the direct effect of a positive report since such a report is not likely to be truth tracking.

4.3 Questioning Bovens and Hartmann's result

While the logic of this result is simple, its implications for the variety-of-evidence thesis are less clear. As Bovens and Hartmann (2003, 95fn) recognize, the result of their model "does not apply to unreliable instruments that do not randomize". It thus seems that the champions of the variety-of-evidence thesis would have little to worry about if evidential sources were rarely as Bovens and Hartmann depict them to be. And it indeed seems to be the case that scientists do not think of their evidential sources in the manner depicted by the shared-reliability model.

For example, consider the macro-data evidence for the hypothesis that one cause of the relatively high French unemployment rate is its relatively long duration of unemployment benefits. To simplify, let us imagine that the macro-data evidence is Pearson's correlation coefficient between the legislated duration of unemployment benefits and the unemployment rate for a sample of industrial countries. There are many potential reasons why this coefficient would not be truth tracking (i.e., would be unreliable) with respect to the hypothesis of interest—e.g., the correlation might not be a sign of causation from benefits to the unemployment rate because of the presence of a common cause, or perhaps the causal structure in France deviates substantially from the ones in the sampled countries.

Now imagine that I decide to compute the correlation coefficient twice. That is, I have the data for the two variables in my computer and I use my favorite statistical software to compute the correlation twice (e.g., send the command `cor(Bd,U)` to R twice). It is reasonable to say that the two results would share a single reliability state: my second correlation coefficient would be reliable evidence for the hypothesis if and only if my first coefficient is also reliable evidence for the hypothesis. Does Bovens and Hartmann's shared-reliability situation come anywhere close to capturing how we think about these two results? Obviously not.

According to the model, the two results will always be in concordance if the procedure is reliable. This implication seems fine. But something strange must happen with my statistical software when the procedure is unreliable. In this case, the two results might be at odds. Furthermore, if I were to compute the coefficient again and again, I would necessarily get discordant results (provided α is strictly between 0 and 1). To be sure, it is possible that my procedure is unreliable *and* is randomizing in this way. My correlation command might have been redefined such that it randomly outputs a number between -1 and 1. However, this is not what would normally be of concern. The reasons given above for why a correlation might be unreliable evidence for a specific causal claim will not bring about such randomness. If, for example, there is a confounding common cause, one would expect both coefficients to be identically affected.

Another way to see the problem with the model is to imagine that the data used to compute the correlation coefficients are known to be totally unrelated to French unemployment. Let us say that the two variables are the respective prices of two types of fish at the Grote Markt in Rotterdam. Since the correlation between these two variables is not tracking the truth of the hypothesis about French unemployment, the source is unreliable for this hypothesis ($R = \neg r$). Now the model tells us that, since the source is unreliable, the two computed correlations must be probabilistically independent. The agent thus starts with some strength of belief α that the correlation between the fish prices is positive. She sends the correlation command and reads a *first* positive coefficient. Strangely, learning this first coefficient will not make her revise her belief about the probable value of the *second* coefficient; just before pressing `Enter` again on her computer, she will still believe to strength α that the computer output will be positive.

The counter-intuitiveness of Bovens and Hartmann's model is not an artifact of my specific choice of example. Take the complex experimental setup of the OPERA Collaboration. Imagine that the research team—before announcing that it had located biases in its experimental procedure—had rerun the experiment *with the exact same setup* (imagine this to be the case even though the exact same setup is a physical impossibility) and that the results had *corroborated* the initial measurement. What would have been the reaction of the scientific community? The model tells us that the new results should have been taken as evidence that the setup is truth tracking. But it seems more intuitive that these results would have been met with indifference. Scientists distrusted the first measurement because they believed that the experiment suffered from a *systematic* (yet unknown) bias. Concordant results from a second identical experiment could thus be explained away by saying that the systematic bias was again operating (as it should if experimenters were careful enough in reproducing the setup). To make some progress in the debate, OPERA researchers needed to find a way to decrease the strength of the belief in the existence of a systematic bias. Rerunning the exact same experiment over and over would not have achieved that.

The upshot of this discussion is that the result of Bovens and Hartmann, as it stands, should not worry scientists and philosophers very much, if at all. There is still room for an unqualified variety-of-evidence thesis when the sources of evidence resemble the ones in science, rather than the ones in the model.[7]

4.4 First modification to the model

Doubt has crept in: Can it not be shown that a more appropriate modeling of the evidential sources still results in a qualified variety-of-evidence thesis? In this section, I offer a negative answer to this question by making a single modification to the model of Bovens and Hartmann.

Bovens and Hartmann's modeling choices are guided by a specific interpretation of the parameter α: for them it means that an unreliable evidential source acts like a randomizer. This becomes clear in their discussion of witnesses as a special case of an evidential source:

> So, we assume that if witnesses are not reliable, then they are like randomizers. It is as if they do not even look at the state of the world to determine whether the hypothesis is true, but rather flip a coin or cast a die to determine whether they will provide a report to the effect that the hypothesis is true. (Bovens and Hartmann, 2003, 57)

While they interpret the parameter as capturing a property of the evidential source, I would rather interpret it from the point of view of the agent: just knowing that the source i is unreliable, α_i is the agent's degree of belief that the report of this source will be positive.

It turns out that Bovens and Hartmann could have modeled this epistemic fact in a different way. Such an alternative way specifies that an

[7] Hartmann (2008, 108) later wrote the following about his assumption regarding unreliability: "This way of modeling a partially reliable instrument is clearly a strong idealization, which will not hold in many cases. "

unreliable evidential source is *systematically biased*, not randomizing. I want to emphasize at the outset that systematically-biased sources of the kind I will model do not cover all the potential kinds of unreliability in science. Obviously, they do not cover randomizing sources (if such sources exist). More importantly, they fail to encompass the miscalibrated sources previously mentioned in section 4.2.

I still think that what I model as ‘systematically-biased sources’ capture important reasons why one can judge a source to be unreliable. Here are a few hints at these reasons without any claim to be comprehensive. One general class of cases comprises the diverse ways in which an evidential report can be affected by a preconceived view of what is the ‘good’ answer. That might come from researchers performing data mining until they get the answer they want or from them simply falsifying their results because of their sponsor’s interests. It can also come from institutional pressures in science: peer review systematically favoring some sort of result, or deeply-rooted hypotheses making scientists revise their experimental procedure until the output fits ‘what is known’.

Another class of cases has to do with the risks of using something as a stand-in (as a model) in order to learn about something else. If one uses, for instance, an animal to learn about the potential side effects of a drug on humans, the extrapolation might go wrong because there is some biological mechanism in the model not present in the target (or the other way around) which makes the drug have some effect in one group of subjects but not in the other. The result from the model subjects will thus be systematically biased when used as a report for the target subjects.

To model sources that are potentially systematically biased, I redefine the reliability variable:

- the new reliability variable $R_i = \{r_i, b_i^h, b_i^{\neg h}\}$, where r_i stands, as before, for the proposition that the source is reliable, b_i^h stands for the proposition that the source is biased toward a positive report for the hypothesis regardless of its truth, and $b_i^{\neg h}$ stands for the proposition that the source is biased toward a negative report.

This ternary variable (all the previous variables were binary) is arrived at by giving a more finely-grained specification of the proposition $\neg r_i$. It is now decomposed into two disjoint propositions—i.e., we now have $\neg r_i = b_i^h \cup b_i^{\neg h}$.

My first modification inserts this new variable into the previous model. The probabilistic independencies that can be read off the Bayesian net-

Table 4.2: *Probability of a positive report given the values of H and R_i*

$P(e_i \vert H, R_i)$	r_i	b_i^h	$b_i^{\neg h}$
h	1	1	0
$\neg h$	0	1	0

works in Figure 4.1 still hold. Furthermore, the specifications of the prior probabilities h_0 and ρ_i in condition (4.1) are retained. We need however to specify more probabilities for R_i:

$$P(b_i^h|\neg r_i) = \alpha_i \quad \text{and} \quad P(b_i^{\neg h}|\neg r_i) = \bar{\alpha}_i. \tag{4.9}$$

This condition assigns prior probabilities to the propositions about positive and negative biases given that the source is already known to be unreliable. Note that what was interpreted as a 'randomization parameter' by Bovens and Hartmann is used explicitly as a strength of belief here. Combining condition (4.9) with condition (4.1), we have the following: the prior probability of a positive bias $P(b_i^h)$ is $\alpha_i\bar{\rho}_i$ and the prior probability of a negative bias $P(b_i^{\neg h})$ is $\bar{\alpha}_i\bar{\rho}_i$.

Finally, we need to expand Table 4.1 by stating explicitly how likely e_i is conditional on b_i^h and $b_i^{\neg h}$. This expansion gives us Table 4.2. Since the uncertainty which figured initially in $P(e_i|H, R_i)$ has been shifted to R_i, the evidential variable E_i is now a deterministic function of H and R_i. This deterministic relation might come as a surprise to some, but it shouldn't be surprising. If we remain committed to the 'irrelevance of an unreliable source' (IUS, see p. 127), the columns for b_i^h and for $b_i^{\neg h}$ in Table 4.2 must each contain the same value twice. If instead of having 1 and 0 for these values, we opt for values strictly between these two, we reintroduce into the model Bovens and Hartmann's unreliability as randomizing. The counterexamples used in section 4.3 would thus apply. As long as we remain committed to the IUS condition, the notion of a systematic bias must be captured by a deterministic function. In future work, I will drop the IUS condition in the context of calibration issues, but I keep it here since it seems pertinent for some sources of unreliability.

The probabilities in Table 4.2 and in equations (4.9) give us two new versions of the model. Define $P_{I'}(\cdot)$ as the joint probability distribution associated with the independent-reliability situation (i.e., the distribution associated with subfigure 4.1a), and $P_{S'}(\cdot)$ as the distribution associated with the shared-reliability situation (i.e., the distribution associated with subfigure 4.1b). We can now assess the variety-of-evidence thesis: Is it always the case that $P^*_{I'}(h) > P^*_{S'}(h)$?

In fact, $P^*_{I'}(h)$ is no different from $P^*_I(h)$—i.e., for the case of sources with independent reliabilities, Bovens and Hartmann's version and my version give the same result (proof in Appendix A.2.1). This is welcome news given that Bovens and Hartmann's version seems to capture what one means in saying that two evidential reports are fully independent regarding a hypothesis—i.e., it concurs with what Shogenji (2005, 308) presents as "a general consensus among probability theorists on how to formalize the condition that two pieces of evidence E_1 and E_2 are independent of each other with respect to proposition A."[8]

Things are different when we turn to the new version with shared reliability. The posterior probability of the hypothesis is now (proof in Appendix A.2.2)

$$P^*_{S'}(h) = \frac{h_0}{h_0 + \bar{h}_0 L_{S'}} \quad \text{where } L_{S'} = \frac{\alpha\bar{\rho}}{\rho + \alpha\bar{\rho}}. \tag{4.10}$$

The likelihood ratio $L_{S'}$ is identical to the one resulting from an evidential set with only a single element instead of two (see equation A.3 in Appendix A.1.3). In other words, adding a second positive report in this new shared-reliability model has no effect on the degree of confirmation of the hypothesis. The reason for this result is simple: the second report cannot be anything but consistent with the first report in this model. The two evidential variables not only share reliability, they also share the direction of the bias if they are indeed biased. There is no longer the possibility of detecting that a source is unreliable by finding discordant reports coming from this source. Since this possibility no longer obtains, multiplying the reports from the same source becomes useless.

Is it still possible that the reports in the shared-reliability situation are more confirmatory than the ones in the independent-reliability situation? No. The posterior probability of h is strictly higher in the independency case if the probability that the sources are reliable is higher than 0 (see Appendix A.2.3 for the proof). There is thus no combination of admissible parameter values for which having shared reliability is, *ceteris paribus*, better. Using the same source again is not conducive to confirmation because it no longer holds the promise of detecting the potential unreliability of the source. A second independent report is thus necessarily more confirmatory. The variety-of-evidence thesis holds without qualification in this version of the models.

[8] It is a case of Sober's conjunctive fork (Sober, 1989; Fitelson, 2001)

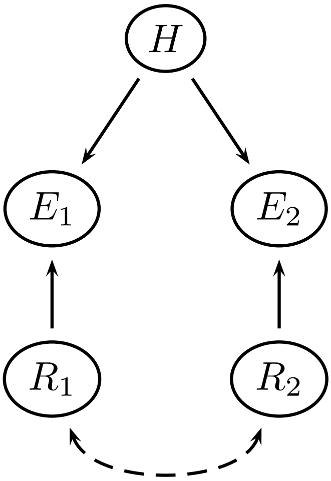


Figure 4.3: *Extended model with degrees of independence*

4.5 Degrees of independence

The result supporting the variety-of-evidence thesis in the previous section suffers from a major limitation. While the variety-of-evidence thesis explicitly compares *more* independent to *less* independent evidential elements, the comparison made with our two models is between *fully* independent and *fully* dependent evidential elements. Our comparison of confirmation was restricted to the two ends of a spectrum whereas the variety-of-evidence thesis deals with how confirmation changes with changes in the degree of independence.

There is a simple way to model degrees of independence by extending the setup of the previous section. The graphical representation of this extended model is in Figure 4.3 and its associated probability distribution will be labeled $P_F(\cdot)$. The modification here adds a probabilistic association between the two reliability variables R_1 and R_2.[9] The rest of the model remains intact.

The association between the reliability variables is fully captured by specifying the probabilities for the nine possible combinations of their values. Panel (a) of Table 4.3 offers a general notation for these nine possibilities. For instance, ω_{rr} is the probability that both sources are

[9] Bovens and Hartmann (2003, 75-77) offer a model with a super-reliability variable which is specified as a common cause of the R_i's. They however do not use it to discuss the variety-of-evidence thesis. Since this super-reliability variable is difficult to interpret and since only modeling a probabilistic association between R_1 and R_2 is sufficient for my goal here, I opt for the second option.

Table 4.3: *Joint probabilities for the reliability variables (assuming symmetry)*

(a) *General case*

$P(R_1,R_2)$	r_2	b_2^h	$b_2^{\neg h}$
r_1	ω_{rr}	ω_{rh}	$\omega_{r\neg h}$
b_1^h	ω_{rh}	ω_{hh}	$\omega_{h\neg h}$
$b_1^{\neg h}$	$\omega_{r\neg h}$	$\omega_{h\neg h}$	$\omega_{\neg h\neg h}$

(b) *Fully-shared reliability*

$P(R_1,R_2)$	r_2	b_2^h	$b_2^{\neg h}$
r_1	ρ	0	0
b_1^h	0	$\bar{\rho}\alpha$	0
$b_1^{\neg h}$	0	0	$\bar{\rho}\bar{\alpha}$

(c) *Fully-independent reliability*

$P(R_1,R_2)$	r_2	b_2^h	$b_2^{\neg h}$
r_1	ρ^2	$\rho\bar{\rho}\alpha$	$\rho\bar{\rho}\bar{\alpha}$
b_1^h	$\rho\bar{\rho}\alpha$	$(\bar{\rho}\alpha)^2$	$\bar{\rho}^2\alpha\bar{\alpha}$
$b_1^{\neg h}$	$\rho\bar{\rho}\bar{\alpha}$	$\bar{\rho}^2\alpha\bar{\alpha}$	$(\bar{\rho}\bar{\alpha})^2$

reliable. The elements on the main diagonal $(\omega_{rr}, \omega_{hh}, \omega_{\neg h\neg h})$ are the probabilities associated with the proposition that the two sources are in the same reliability state. Note that the Table already assumes symmetry between the two sources—i.e., $P(r_1, b_2^h) = P(b_1^h, r_2) = \omega_{rh}$ and so forth. This assumption was also used in the previous sections as part of the assumptions sufficient to meet the *ceteris paribus* condition of the variety-of-evidence thesis.

In this new model, the posterior belief in the hypothesis given two positive reports is (proof in Appendix A.3.1)

$$P_F^*(h) = \frac{h_0}{h_0 + \bar{h}_0 L_F} \quad \text{where } L_F = \frac{\omega_{hh}}{\omega_{rr} + 2\omega_{rh} + \omega_{hh}}. \tag{4.11}$$

Tables 4.3b and 4.3c give the specific values taken by the ω's for the two extreme cases on which the previous sections focused. It can be easily verified using these Tables that the expression in (4.11) reduces to (4.10) or to (4.6) for each of these extreme cases—i.e., the model of the previous section is fully embedded into this one (including its result for the variety-of-evidence thesis).

Is there a ready measure of degrees of independence? My proposal is based on the following consideration. Compare Tables 4.3b and 4.3c. The probability mass is all on the main diagonal in the first case. In other words, it never happens that the two sources are in different reliability states. In the case of full independence, the probability mass is more spread out since the joint probability $P(R_1, R_2)$ is simply the product of the marginal probabilities, i.e., $P(R_1)P(R_2)$. In fact, each element on the main diagonal in Table 4.3c is exactly the square of the same element in Table 4.3b. This fact suggests a specific metric to characterize degrees of independence.

Define a variable $\delta \in [0,1]$ which is interpreted as measuring the distance of the evidential set from fully-shared reliability—i.e., when $\delta = 0$ we have no independence, when $\delta = 1$ we have full independence, and when δ is strictly between 0 and 1, we have only partial independence. Given values for ρ, α and δ, the elements on the main diagonal are:

$$\omega_{rr} = \rho^{1+\delta} \qquad \omega_{hh} = (\bar{\rho}\alpha)^{1+\delta} \qquad \omega_{\neg h \neg h} = (\bar{\rho}\bar{\alpha})^{1+\delta}. \tag{4.12}$$

These relations entail that the probability mass is shifted away from the elements on the main diagonal as the degree of independence increases. In other words, it becomes less likely that the two sources share the same reliability state.

With this variable δ, the variety-of-evidence thesis can be restated.

Variety-of-evidence thesis. *Ceteris paribus*, $\partial P^*_F(h)/\partial \delta > 0$, for all admissible values of ρ, α and δ.

The restatement of the thesis is thus that the posterior degree of belief in the hypothesis invariably increases as we marginally increase the degree of independence of the evidential sources.

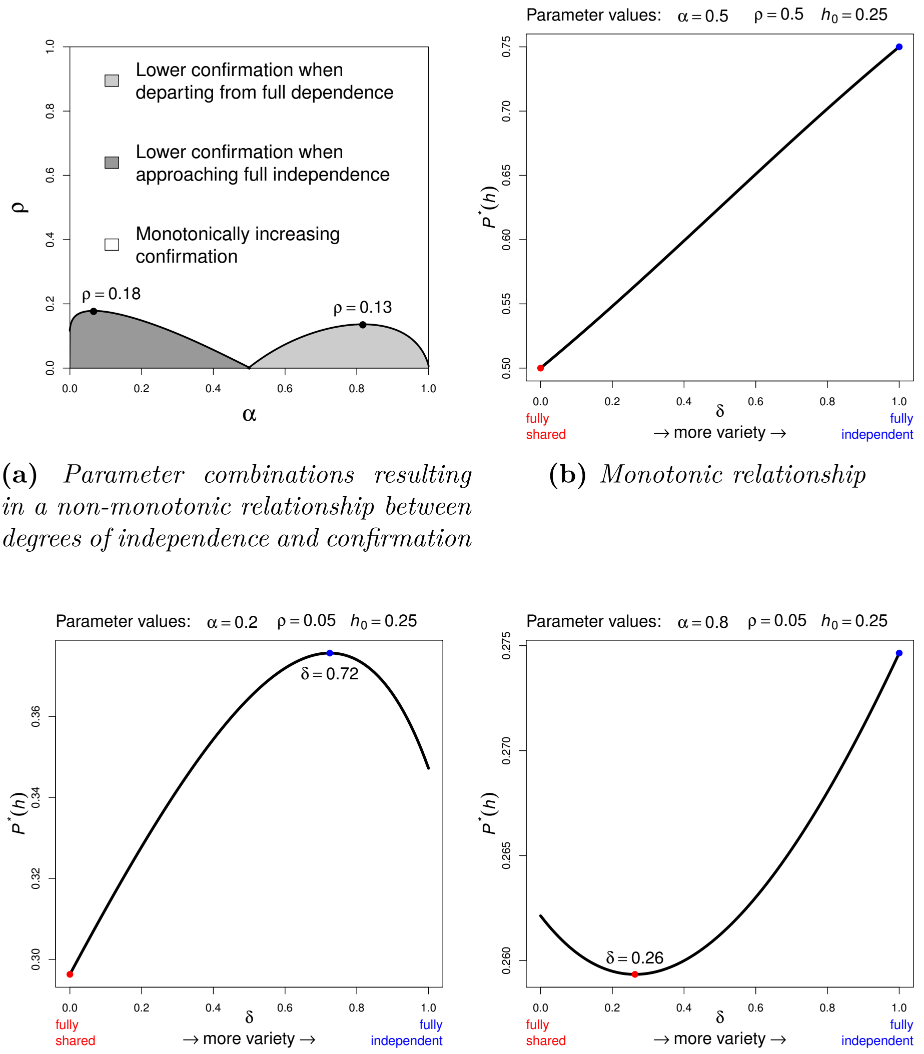
Before we assess this thesis, we need to specify how the off-diagonal elements in Table 4.3a change as δ is modified. One obvious restriction is that the sum of all the elements (the 9 ω's) must be 1. The interpretation of the *ceteris paribus* condition previously used also restricts the values of the off-diagonal elements but not enough to ensure uniqueness. In addition to these restrictions, I thus also stipulate that the marginal probabilities of R_1 and R_2 are not a function of δ—i.e., $P(r_i) = \rho$, $P(b_i^h) = \bar{\rho}\alpha$, and $P(b_i^{\neg h}) = \bar{\rho}\bar{\alpha}$, for $i = \{1,2\}$ and for all $\delta \in [0,1]$ (see Appendix A.3.2).

This model leads to a qualification of the variety-of-evidence thesis (proof in Appendix A.3.3). Increasing the degree of independence leads to more confirmation if and only if the following condition holds:

$$(1 - 2\bar{\rho}\bar{\alpha})\ln(\bar{\rho}\alpha) + (\bar{\rho}\bar{\alpha})^{1+\delta}\ln(\alpha/\bar{\alpha}) < 0 \tag{4.13}$$

But there are combinations of admissible values for ρ, α and δ which violate this condition.

Figure 4.4 presents graphically the different possibilities. For most combinations of ρ and α, the relationship between degree of independence and confirmation is as stated by the variety-of-evidence thesis (panel (b)



(a) *Parameter combinations resulting in a non-monotonic relationship between degrees of independence and confirmation*

(b) *Monotonic relationship*

(c) *Low α*

(d) *High α*

Figure 4.4: *Non-monotonicity is possible*

presents such a case).[10] Panel (a) shows that there are in fact two distinct regions of the parameter space where the relationship between independence and confirmation is non-monotonic. These two possibilities share (extremely) low values of ρ. In other words, these are situations where prior information leads the agent to believe that it is highly unlikely that a given source is truth tracking. The relationship is indeed always monotonic when the trust in the source is above .18 (in Bovens and Hartmann's model this was .5).

There are two features distinguishing the two non-monotonic situations from each other. First, as shown in panel (a), they differ in their values for α—i.e., the probability that the report is positive given that the source is unreliable. Second, as is evident by comparing panels (c) and (d), the two regions are associated with different shapes of non-monotonicity (concave versus convex functions).

What is going on? As in Bovens and Hartmann's model, what happens is that, upon learning e_1 and e_2, the agent reassesses the probability that the sources are reliable. The region of the space $\alpha \times \rho \times \delta$ where confirmation *decreases* with independence, is exactly the region where *trust in the reliability of the sources decreases with independence* (proof in Appendix A.3.4). In other words, to compare two evidential sets (for h)—say $\mathbf{E} = \{e_1, e_2\}$ and $\mathbf{E}' = \{e'_1, e'_2\}$—which differ only with respect to their degree of reliability independence (δ)—the elements of $\mathbf{E}$ being more independent than the elements of $\mathbf{E}'$—one simply needs to assess the following ratio for each set:

$$\frac{\omega_{rr} + 2\omega_{rh}}{\omega_{hh}} \tag{4.14}$$

The set with the higher ratio is more confirmatory than the other. The numerator of this ratio captures the probability of realizations of R_1 and R_2 that generate two evidential elements (e_1 and e_2) which are indeed truth revealing for h. The denominator is the probability that the two sources are producing positive-but-garbage reports for h. The denominator of $\mathbf{E}$ will always be smaller than that of $\mathbf{E}'$. This fact might capture the intuitive appeal of the variety-of-evidence thesis: it is less likely to

[10] The proportion of the parameter space $\alpha \times \rho$ where the relationship between independence and confirmation is not monotonically increasing—i.e., the area of the two gray regions in panel (a)—is 10.3%. As a point of comparison, this proportion is 15.3% in Bovens and Hartmann's model (as depicted in Figure 4.2). If one considers instead the three dimensional space $\alpha \times \rho \times \delta$, only 2% of it gives $\partial P^*_F(h)/\partial \delta < 0$. These proportions should not be interpreted as probabilities.

get two garbage reports from sources that are (more) independent. But the full ratio is what ultimately decides between $\mathbf{E}$ and $\mathbf{E}'$.

Let me briefly discuss the only two situations in which the variety-of-evidence thesis is turned upside down. First, for low values of α combined with extremely low values of ρ (as in Figure 4.4c), getting two positive reports comes as a surprise— it was judged far more likely to receive at least one negative report because of the realization of $b_i^{\neg h}$. In this case, moving toward independence is initially beneficial, but more independence becomes detrimental to confirmation as one approaches the extreme of full independence. This result is interesting because it means that slightly departing from full independence sometimes increases confirmation.

Second, for high values of α combined with extremely low values for ρ (as in Figure 4.4d), getting two positive reports is not surprising; however, the agent judges it highly likely that the information is worthless (because b_1^h and b_2^h are likely to be realized). In this case, a departure from full dependence adversely affects confirmation. An implication of this non-monotonicity is that the second positive report can be disconfirming h—i.e., $P_F(h|e_1, e_2) < P_F(h|e_1)$. Remember from section 4.4 that the posterior belief in h after two *fully-dependent* reports (i.e., $\delta = 0$) is identical to the posterior belief after a single report. Both are represented by the point at the extreme left of the curve in Figure 4.4d. All the points lying below this point are thus cases in which the second report is disconfirming h. The agent puts so little trust in the evidential sources that a second report is interpreted as a sign that both sources are positively biased, and the initial (slight) increase in the belief for h is cut back.

4.6 Conclusion

The variety-of-evidence thesis seems to be a widespread implicit guideline in scientific practice. This thesis says that, *ceteris paribus*, the confirmatory power of an evidential set for a given hypothesis increases with the diversity (i.e., the independence) of the evidential elements in the set. Thus one should praise ‘independent evidence’ and be suspicious of the rest.

Bovens and Hartmann (2002, 2003) cast doubt on the universal applicability of this thesis by showing with a simple model that, in some peculiar epistemic situations, it is sometimes a disadvantage for confirmation to have independent evidential elements, *ceteris paribus.* I have argued that the relevance of this result is diminished by two characteristics of their model.

First, their idea that unreliable sources are randomizers leads them to model *fully-dependent* sources in a way which is unlikely to reflect how scientists think about their sources of evidence. The problem is that Bovens and Hartmann assume that two *fully-dependent* sources still produce two *independent* reports when they are unreliable—i.e., $E_1 \perp\!\!\!\perp E_2 | \neg r$. Instead, in actual scientific settings it seems to be the case that two reports coming from fully-dependent sources will always coincide even when the sources are unreliable. In section 4.4, I showed that the variety-of-evidence thesis is rehabilitated once the independent-randomizer assumption is dropped and replaced with the assumption that an unreliable source is *systematically* biased. This modification is compatible with the key intuition behind the notion of reliability in Bovens and Hartmann's model—i.e., the irrelevance of an unreliable source (see p. 127).

Second, there is another serious limitation in Bovens and Hartmann's model, a limitation that my first modification of their model shares. The comparison made to assess the variety-of-evidence thesis is between extremes; it is between fully-independent and fully-dependent evidential elements. The most relevant comparison is rather one of degree: *less* versus *more* independence of the sources. In section 4.5, I showed that when the model is modified to enable comparisons of degrees of independence, the variety-of-evidence thesis needs to be qualified. There are special epistemic situations wherein more independence does not give more confirmation. This qualification only applies to a subinterval of the spectrum from full dependence to full independence. Indeed, the two extremes of the spectrum always stand in the confirmatory relationship depicted by the variety-of-evidence thesis.

Where do my modeling efforts leave us? First, the usual caveat about idealization applies: it might well be that the way in which epistemic situations have been modeled here does not capture what is pertinent for the variety-of-evidence thesis. It is certain that my model does not encompass all the ways in which an evidential source can be unreliable (e.g., the calibration problems mentioned in section 4.2).

Even if one accepts the idealizations, the conclusion to draw about the variety-of-evidence thesis is not straightforward. One plausible reaction to the result of the last section is as follows. The variety-of-evidence thesis can break down in the extended model only if the agent has enormous doubts about the reliability of the evidential source; she must judge it to be at least 82% likely that the source is unreliable. One could thus read the result as highlighting the danger of using extremely weak evidential sources, rather than as a direct refutation of the variety-of-evidence thesis. This thesis could be interpreted as implicitly assuming that the evidential sources are sufficiently trustworthy to begin with.

The fate of the variety-of-evidence thesis is not yet settled.

Part III

The Dynamics of an Eclectic Science

Chapter 5

Deviant Cases in an Eclectic Science: Considerations from Recent Economics

5.1 Introduction

Post-positivist philosophy of science has given a central role to deviant cases in the scientific process. This role is evident in the work of giants such as Karl Popper, Thomas Kuhn, and Imre Lakatos. Popper's "falsifiers" are statements of deviant cases—i.e. the elements demonstrating that one's bold conjecture was wrong (Popper, [1959] 1992, [1963] 2002). According to Kuhn (1962), deviant cases are the unsolved "puzzles", or "anomalies", which plague any paradigm. These troublesome cases might be ignored by scientists for a while, but they still bear the seeds of crisis. Similarly, deviant cases are the Lakatosian "anomalies" requiring some fiddling with the protective belt. This fiddling is, in turn, what makes the research program either progressive or degenerating (Lakatos, 1978). Walking in the footsteps of these giants, contemporary philosophers still recognize—as they should—the centrality of deviant cases for the dynamics of science.

Most of post-positivist philosophy of science has also been characterized by a view on the nature and centrality of theories in science. Popper can serve here as an extreme example of this view. For him, "[t]he empirical sciences are systems of theories", and "[s]cientific theories are universal statements" which are (ideally) organized in an axiomatized system (Popper, [1959] 1992, pp. 37-48). Adding initial conditions to the set of universal statements, one can deduce empirical propositions.[1]

This understanding of theory is a version of what Nancy Cartwright (1999, p. 184) dubs the "vending machine view":

> The theory is a vending machine: you feed it input in certain prescribed forms for the desired output; it gurgitates for a while; then it drops out the sought-for representation, plonk, on the tray, fully formed, as Athena from the brain of Zeus.

The vending-machine view has been highly influential in philosophy of science. It has, however, been argued—by Cartwright and many others[2] —that philosophical accounts relying on this view leave in the dark much of scientific practice. The construction of alternative accounts is well under way. The present chapter is a modest contribution to this constructive process: I seek an understanding of research on *deviant cases* in what I call 'eclectic science'—a type of science which does not have a theory working like a vending machine, but is yet full of resources.[3]

Looking at a single deviant case in economics, I give tentative answers to three questions. What makes a case deviant (section 5.3)? What is the epistemic goal of deviant-case research (section 5.4)? Given this goal, how should research proceed (section 5.5)? Before developing my own account, I present in the next section how these three questions can be answered from the perspective of the vending-machine view. One goal of the chapter is then to argue that these answers are inappropriate for deviant-case research in eclectic sciences. My overarching goal is however constructive: I aim at a new philosophical account—both descriptive and normative—of deviant-case research.

There are at least three reasons why this new account is worth seeking. First, Popper, Kuhn, and Lakatos were right to think that dealing with deviant cases is a central moment in science. Some attempts to explain deviance turn out to cause major scientific changes. A philosophical account of eclectic sciences would thus be seriously incomplete if it did

[1] Note that my other two giants do not entirely agree with Popper here. For both of them, the sciences are more than sets of universal statements; in particular, sciences are also characterized by sophisticated problem-solving methods. This divergence is why Popper is an *extreme* example of what I call, following Cartwright, the vending-machine view. Because my goal is not exegetical, I will not attempt to specify the exact view of any of my three giants.

[2] These other scholars include Ronald Giere (2006), Paul Teller (2001) and William Wimsatt (2007a).

[3] My reasons for using the term 'eclectic' will become clear as my chapter unfolds—or so I hope.

not include an analysis of deviant-case research. Second, philosophical accounts of eclectic sciences are still fragmentary. It has proven far easier to establish what eclectic sciences are not, than to articulate what their nature actually is. Scrutinizing deviant-case research holds the promise of revealing key characteristics of eclectic sciences, and thus aiding the development of a larger philosophical account. Finally and more speculatively, my analysis of deviant-case research can also be motivated by concerns about the performative effects of the earlier account. My alternative account might help practicing scientists see what their deviant-case research is really about.

5.2 Deviant cases for the vending-machine view

Many philosophers of science have been in the grip of the vending-machine view, and newcomers to the field are inevitably introduced to it—though not under this label. The view straddles the divide between syntactic and semantic views of theories.[4] It is also intimately connected with the hypothetico-deductive model of theory testing.

When a theory is understood as a set of lawlike propositions appropriately related, the two key ingredients of the vending-machine view for me are (i) that the theoretical propositions, when combined with secondary propositions (e.g. auxiliary hypotheses, initial conditions), entail empirical propositions, and (ii) that all these propositions are (provisionally) believed to be true of the world. The first ingredient is what gives the automated character of the vending machine: you plug in the secondary propositions, and empirical predictions come out. The second ingredient indicates how to fix the machine when the output is not to our tastes. There must be (at least) one false proposition—either theoretical or secondary. Find it and fix it.

Note that the notion of entailment in the first ingredient is deductive; if it were not, it would not necessarily be the case that there is at least

[4] More precisely, it encompasses the syntactic view and at least a subset of the semantic view (e.g. van Fraassen, 1980). For a more careful treatment, see Cartwright (1999, pp. 179-86). I will not use the (somewhat awkward) vocabulary of any of these two views in the main text, and will rather stick to "the looser informal construal of theories as collections of lawlike *statements* ... systematically related to one another" (Hausman, 1992, p. 297). I trust that the reader can make the appropriate translation—e.g. from the truth of statements to model-world isomorphism.

one false premise when the output is not to our tastes. In other words, while *inference* can be generally defined as the process of reasoning from premises to conclusion, the vending-machine view concentrates on one type in the rich set of inferences: the inferences for which the conclusion cannot be false when all the premises are true. It should be obvious to the reader that most inferences do not have this property. For instance, I came to believe the proposition 'you [the present reader] think[s] this paragraph is superfluous' because I believe that 'in general, readers of academic philosophy know by heart the definition of deductive validity'. My conclusion is obviously not necessitated by my premise.[5] But if my conclusion is nevertheless correct this time, I now ask you to withhold your judgment until later, since non-deductive inferences will play a major role in this chapter.

The vending-machine view can provide answers to my three questions about deviant-case research. Let me sketch how these answers would look. Note that I do not want to attribute these answers to specific philosophers. I instead take them to be variations on familiar thoughts in the philosophy of science.

First, what makes a case deviant? In the vending-machine view, a proposition D is (a statement of) a deviant case in relation to a theory T and secondary propositions S if the conjunction of T and S entails not-D. How it works can be illustrated by the *mythical* example of the boiling point of water.[6] Our theory here is made of a single law which can be expressed as follows: under normal atmospheric pressure (1 bar), water boils at 100°C. Given the form of this law, the proposition 's is boiling' is entailed by combining the law with three propositions: 'atmospheric pressure is normal', 's is a sample of water', and 's is at 100°C'. We are confronted with a deviant case if s is in fact not boiling. The proposition 's is not boiling' falsifies the set of propositions made of the law and the three assumptions.

Falsification points to the answer to my second question. From the perspective of the vending-machine view, the epistemic goal of deviant-

[5] Don't tell me that my inference is an enthymeme with a deductive structure that would become obvious if the unexpressed premises were stated. My only stated premise is a generic; this type of propositions—widespread and highly useful—cannot support a deduction to a singular instance. More on this later.

[6] Hasok Chang (2007) supplies evidence that this example is a myth. According to his research, the conditions under which water boils are far more complex than what we all believe. Chang's research is one piece of evidence among many that the vending-machine view grossly distorts, not only our understanding of the social sciences, but also what has so often been depicted as the 'harder' sciences.

case research is to remove the inconsistency in the set of propositions. This goal is both what scientists are depicted as striving for, and an implicit prescription. If we believe the empirical proposition D, logic requires us not to accept all the other propositions used to deduce not-D. At least one proposition must go. But which proposition *should* we drop?

Here comes the third question: How should research proceed? If (pace Popper) we allow for a notion of degree of confirmation, a strategy emerges such as “the weak-link principle” (Hausman, 1992, p. 297). Take again the example of the boiling point of water. My observation is unequivocal: s is not boiling. It must then be the case that either my law or one of the three secondary propositions is false. Since I probably do not have full certainty in any of these propositions, I might (and maybe must) start probing the one I have least confidence in—i.e. the weak link. For instance, perhaps the new thermometer that made me believe ‘s is at 100°C’ is rubbish. If this investigation leads me to reject one of the three secondary propositions—e.g. my good, old thermometer gives me 90°C—everything is ‘normal science’ again. Consistency is restored among my beliefs.

If the doubt falls on the law, matters get more complicated. According to many philosophers, there is always the possibility of brushing this evidential element aside. If there is no promising alternative theory on the radar, there is not much to worry about. If we are brave enough (or forced by the circumstances) to address the deviant case as an anomaly for our law, science becomes wild—imagination and bold conjectures come in.

Philosophers have formulated many suggestions on how to police this innovative burst. Lakatos, for instance, worries that consistency of beliefs is restored *ad hoc*—e.g. I could add a clause to my law such that it is made inoperative only in circumstances exactly like the ones of the deviant case. We want to rule that out as bad practice. His solution is to ask that the modified theory makes novel predictions, which are then tested.[7]

The vending-machine view offers an entertaining story of deviant-case research. The problem is that it mischaracterizes much of *actual* research on deviant cases. I mean this claim first descriptively: researchers in eclectic sciences will typically neither experience nor react to deviant cases in the way depicted here. But I also mean it as a normative failure: the above prescriptions about how scientists *should* react to deviant cases are misguided. In the rest of the chapter, I will support these claims about

[7] Other philosophical accounts of science will have alternative methodological prescriptions—e.g. ‘be a good Bayesian’.

the descriptive and normative failures of the vending-machine view by analyzing a specific instance of deviant-case research in economics: the research on the behavior of the German unemployment rate during the 2008-9 economic crisis.

5.3 What makes a case deviant?

Deviance always appears against the backdrop of expectations. In the vending-machine view, these expectations are arrived at by a deduction from believed premises. One might water down this view by saying that the above depiction is a 'rational' reconstruction of 'mature' disciplines, but it remains that the vending-machine view sells a particular representation of science: the expected case is what is derived from the conjunction of lawlike propositions that we believe in, together with assumptions that we also deem likely to hold.

This picture of science does not provide an account of deviance in situations where we either lack a deductive structure or we don't believe in some of our premises. These situations might however be the rule in eclectic sciences like economics. We need an alternative account.

Now to my case. At the onset of the 2008-9 economic crisis, economists expected a particular behavior of the unemployment rate. This expectation was based on a more general belief about how the unemployment rate relates to economic growth: the change in the unemployment rate of one economic unit (e.g. a country) is expected to be inversely and monotonically related to its GDP growth rate. Most importantly, when GDP growth falls, the unemployment rate is expected to increase, and the deeper the drop in the growth rate, the steeper the unemployment hike.

This expectation of economists comes from particular sources. One can divide these sources in two: a clear empirical pattern and results from model worlds. None behaves like the philosopher's vending machine. Let me state two points before discussing each source in turn. First, these sources do not exhaust the resources of economists. They provide expectations, not the last word on any topic. This point is connected with the nature of *eclectic* sciences: they are rich in disparate resources. This richness will only become evident later in this chapter. Second, an economist would not typically think about empirical patterns and model worlds in a perfectly compartmentalized way. In actually generating expectations, both sources are intermingled; I divide them for the sake of analysis. This point is also connected with the nature of eclectic sciences: they are about combination.

5.3.1 The role of empirical patterns

National accounting is an impressive technical achievement of the 20th century (Vanoli, 2008). While early economic thinkers recognized key economic regularities, national accounting now allows the quantification of those regularities. Quantification is crucial because, as we will see in a moment, deviance appears more starkly when one has access to quantified relationships.

When it comes to the relationship between fluctuations in economic growth and changes in the unemployment rate, the quantified relationship takes the name of Okun's law. The original formulation of it is "the approximate 3-to-1 link between output and the unemployment rate" (Okun, 1962, p. 3). What Arthur Okun means by this compact formulation is that a difference of one percentage point of the aggregate growth rate is *generically* associated to a difference of $\frac{1}{3}$ percentage point of the unemployment rate in the opposite direction. In a formula, this reads

$$\dot{U}_t \approx a - \frac{1}{3}\hat{Y}_t \tag{5.1}$$

where $\dot{U}_t$ is the change of the unemployment rate from one period to the next ($U_t - U_{t-1}$), $\hat{Y}_t$ is the growth rate of aggregate output (GNP in Okun, 1962), and a is a constant determining when the growth rate is sufficient to have a steady unemployment rate ($\dot{U}_t = 0$ when $\hat{Y}_t \approx 3 * a$).

What sort of 'law' is Okun's law? I want to argue that it does not square with how laws are depicted in received philosophy of science.[8] This point is not at all meant to belittle Okun's work—he established an extremely important relationship—but simply to be able to argue that Okun's law does not (and cannot) play the role of a law in the vending-machine view. Furthermore, my goal is not to say that Okun's law is no 'true' law; I don't mind the use of this term provided one recognizes that what philosophers refer to in using the term does not include Okun's relationship.

[8] As a historical aside, Arthur Okun was far from presenting this empirical relationship as a law. He was, in fact, explicit about the rule-of-thumb nature of his "3-to-1 link". Nevertheless, the term 'Okun's law' appeared in print the same year in a paper by Robert Solow (1962, pp. 82-3).

First, Okun's law is not a high-level principle like Newton's laws. It does not postulate any additional property like 'force', but rather state a relationship between independently-accessible quantities. Second, taken as an 'empirical' law—e.g. Hooke's law—it is neither deterministic nor probabilistic. As Okun's use of the qualifier 'approximate' makes clear—and as '≈' in formula 5.1 tries to capture—the change in unemployment on a given period is not perfectly predicted by the growth rate. And it cannot be read as a probabilistic law because it fails to provide a probability distribution.

Let me elaborate on this last point to avoid confusion. I do not claim that Okun's law could not be turned into a probabilistic law, but simply that, as it is, it is no such thing. The simplest probabilistic reading of the relationship (5.1) would be to add an additive error term to the right-hand-side to obtain

$$\dot{U}_t = a - \frac{1}{3}\hat{Y}_t + \varepsilon_t \tag{5.2}$$

where ε_t would be attributed a specific probability distribution—e.g. $\text{NID}(0, \sigma^2)$. But economists do not endorse Okun's 3-to-1 link with a specific but implicit probability distribution appended to it. As an illustration, take how Okun himself establishes his law. His 3-to-1 link is arrived at by combining three methods of estimation, the simplest being a least-squares regression. For this last method, he does not inspect the residuals, which show clear signs of serial correlation. In other words, he does not care to specify the probability distribution of his error term. After Okun, many economists have obviously tried to turn his insight into 'proper econometrics' by adding lags and other variables, testing for structural breaks, and so on (see Cuaresma, 2008). But this work is irrelevant to the point here: Okun's simple relationship—not these refinements—is one source of expectations for economists. This relationship cannot and does not pretend to be a probabilistic law.

Finally, could Okun's law be read as a *ceteris paribus* law? I think not. Remember the function attributed to a *ceteris paribus* condition: it is meant to rule out cases in which a lawlike association is disrupted or hidden due to other factors intervening. It is well-known that econometric methods like multiple regressions have been developed with the goal of securing the *ceteris paribus* condition—i.e. one adds control variables in order to keep intervening factors constant (Morgan, 1990). Now, the problem with a *ceteris paribus* reading of Okun's law is exactly that the relationship is retrieved from data *without controlling for anything.* Okun computed an *unconditional* bivariate relationship between output growth rates and changes in unemployment—there is nothing 'kept constant', all other factors are left free to fluctuate.

A reply here is to say that there is another interpretation of *ceteris paribus* under which the Latin locution is not substitutable for 'all things equal' but for 'in general', 'under normal circumstances', or something of this sort. I only have a terminological quibble here: *ceteris paribus* is clearly a misnomer. Beside terminology, my favorite reading of Okun's statement is quite close to this alternative interpretation. I take Okun's statement to be one instantiation of what semanticists of natural languages call 'generics'. These statements are often *imperfectly* reformulated by introducing them with 'in general', 'under normal circumstances' and so on. That the reformulation is imperfect should be clear from reflecting on these different locutions. What is 'normal' is not necessarily 'general' and so on. The fact that we are unsure which locution is appropriate is strong evidence that our reformulations do not perfectly preserve the meaning of the original generic.

Indeed, linguists and philosophers of language agree that generics, even though they are widespread in our linguistic practices, have a meaning which escapes our most careful semantic analysis. Roughly, generics are sentences that "report a kind of general property, that is, report a regularity which summarizes groups of particular episodes or facts." (Krifka et al., 1995, p. 2) Typical examples of generics are 'tigers are striped' and 'cars have radios'. Two key characteristics of generics are that they allow for exceptions, and that they do not include any explicit quantifier—neither the typical existential and universal quantifiers from deductive logic, nor a probability distribution. Recent research suggests that generics are a manifestation of our most basic strategies to orient ourselves in the world, to know what to *expect* and react accordingly (Leslie, 2007, 2008). For us, what matters most is that, by their nature, generics cannot play the role of universal statements in a deduction; since they allow for exceptions, and since we do not have a list of these exceptions, any inference from them to a singular statement will not necessarily be truth preserving.[9]

My suggestion is to take Okun's statement as an instance of generics.[10] Like other generics, it is an expectation-generating statement. And

[9] For a few more words on generics, see subsection 1.4.1 above.

[10] The ones wedded to the language of *ceteris paribus* statements can substitute my 'generics' by 'indefinite *ceteris paribus* statements', where 'indefinite' is important to make clear that the list of factors kept constant, controlled for, or assumed absent

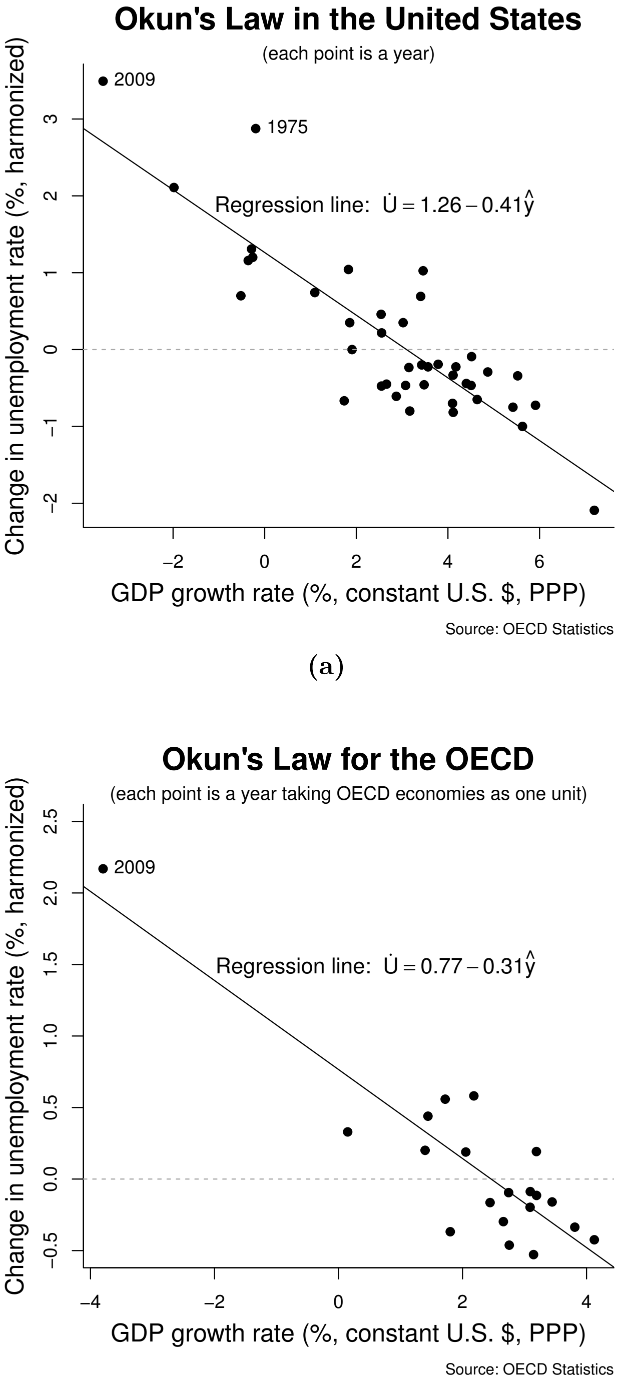
as a generic statement, it is a well-supported one, celebrated by James Tobin (1987) as "one of the most reliable empirical regularities of macroeconomics". Figure 5.1 is a graphical representation of Okun's law for the United States and for all the countries of the OECD taken together. Anyone familiar with social statistics will recognize here an unusually strong bivariate association (for variables that are conceptually distinct).

Why did I plot the relationship for the United States and for the OECD instead of choosing any other country (e.g. Germany) or combination of countries? There are two reasons. The *sociological* reason is the well-known U.S. bias of economic research: the bulk of economists are trained or work in the United States, textbook examples tend to use U.S. data, and so on. The consequence is that the macroeconomic patterns of the United States are the best known among economists.

The second reason has to do with the logic of the social sciences—and other sciences—which force their practitioners to constantly rely on extrapolation to form expectations. By extrapolation, I mean the inference from a proposition believed to be true of one population to the same proposition applied to a different population, i.e. the *target* population (Steel, 2008, p. 3). The target population being Germany in this case, one can either rely on cross-population extrapolation by using the U.S. result in subfigure (a), or downward extrapolation by using instead the OECD result in subfigure (b). Before the crisis, only a small fraction of economists might have known Okun's law for Germany, even though the data to estimate it were readily available. That did not stop the rest of them from having firm expectations, based on extrapolation, about how German unemployment would fluctuate with output growth.

Here is my main message about the role of empirical patterns like Okun's law in expectation formation: by using a *deductive structure* to reconstruct expectation formation based on empirical patterns, one grossly misrepresents what is going on. The problem here is not that the premises are not believed—Okun's law, when it is well understood as a *generic* proposition, is strongly believed in economics. The problem is that the nature of generic propositions does not allow for deductive structures—i.e. you cannot deduce from the proposition 'generically x's are F', that a particular x_i is F, you can produce only informed guesses. That's fine. We often get home safely with informed guesses. And we also know how to argue about whether a particular guess is that well informed.

is unknown (Reutlinger et al., 2011, sec. 3.2). But why use such a misnomer? Furthermore, using 'generics' has the advantage of explicitly connecting the language of science to our rich everyday language.



Okun's Law in the United States

(each point is a year)

Source: OECD Statistics

(a)

Okun's Law for the OECD

(each point is a year taking OECD economies as one unit)

Source: OECD Statistics

(b)

Figure 5.1: *Okun's law is a strong relationship*

5.3.2 The role of model worlds

The empirical pattern in figure 5.1 is not the only source of information which led economists to form firm expectations about the behavior of the German unemployment rate. Economists have stories to tell about why unemployment and output growth are inversely related and, as is well known, economists tell stories through models. In fact, economists have more than one model to account for this relationship—i.e. they have different stories coexisting next to each other. Stories can be hard to individuate. In the present case, there are at least two clearly distinct storylines.

The first model/story is the one to which Okun participated: the macro Keynesian story centered around the concept of potential output.[11] The main idea is that the productive capacities of an economy at a given time are such that it could potentially produce an output Y_t^*. If the potential output Y_t^* were realized, the economy would be at full employment. Potential output grows with time and it happens that *actual* output Y_t does not grow as fast—e.g. because of insufficient aggregate demand. In such cases of output lagging behind its potential, the unemployment rate increases. The most common view is that causality is running from the output gap to unemployment.[12]

One could say a lot about modifications on this bare storyline. There is, for instance, a discussion on the direction of causality: Is causality not running from unemployment to the output gap? Furthermore, the storyline is tightly linked to the representation of the economy through an aggregate production function. But we don't need to go down these lines. Among all the alternative formal models that instantiate the general story, I propose the following:[13]

1. Potential output Y_t^* grows at an exogenous rate $\hat{y}^*$;
2. Actual output Y_t grows at a rate $\hat{y}_t$ which is caused by factors not present in the model (e.g. aggregate demand);
3. Variation in the unemployment rate ΔU_t is linearly and deterministically caused by the output gap $g_t = \hat{y}^* - \hat{y}_t$:

$$\Delta U_t \Leftarrow \alpha g_t \tag{5.3}$$

[11] For a contemporary variation on this story, see Hoover (2012a, pp. 590-2).

[12] For instance, Okun (1962, p. 1) motivated his investigation by stating that our goal is full employment and that "policy measures designed to influence employment operate by affecting aggregate demand and production."

[13] Again, this is only one option. I discuss model variations below.

where α is a positive constant (independent of time) and $\Leftarrow$ is a directional equality sign to capture causal order.

As one can easily see from these three pieces, the inverse relationship between unemployment and actual output follows as a deductive consequence. This case of economic modeling thus fits one of the two elements of the vending-machine view: the expectation comes out of a deductive structure. It lacks however the second element: all of the three propositions above *are not believed* if taken as claims about real economies.

There are three options once this point is recognized. The first option is to say that this model is just a bad model, that we should do better. We should either de-idealize it or replace the pieces by something totally different such that, in the end, we can believe all the pieces. There are good reasons to be skeptical about this strategy as it incarnates the "perfect model model" (Teller, 2001). Successful modeling does not require that all the propositions forming the model (or that are true of the model) be true of the target system (Cartwright, 1999; Morgan and Morrison, 1999; Giere, 2006). Furthermore, a 'perfect model' of a real-world situation remains an unfulfilled promise—at least, it is clearly so in economics. What appears to be the case is that any model represents a target by being similar in some respects with it, but will always be dissimilar in other respects.

This view of modeling does not mean that the model presented above is as good as any other. It is quite likely that modifying it in specific ways would lead to a more adequate model for the task of understanding and predicting the relationship between output and unemployment. The view of modeling endorsed here implies however that this more adequate model will still be related to propositions that one should not believe as statements about real economies. In other words, if we stick to a model with a deductive structure, the second element of the vending-machine view will remain out of reach.

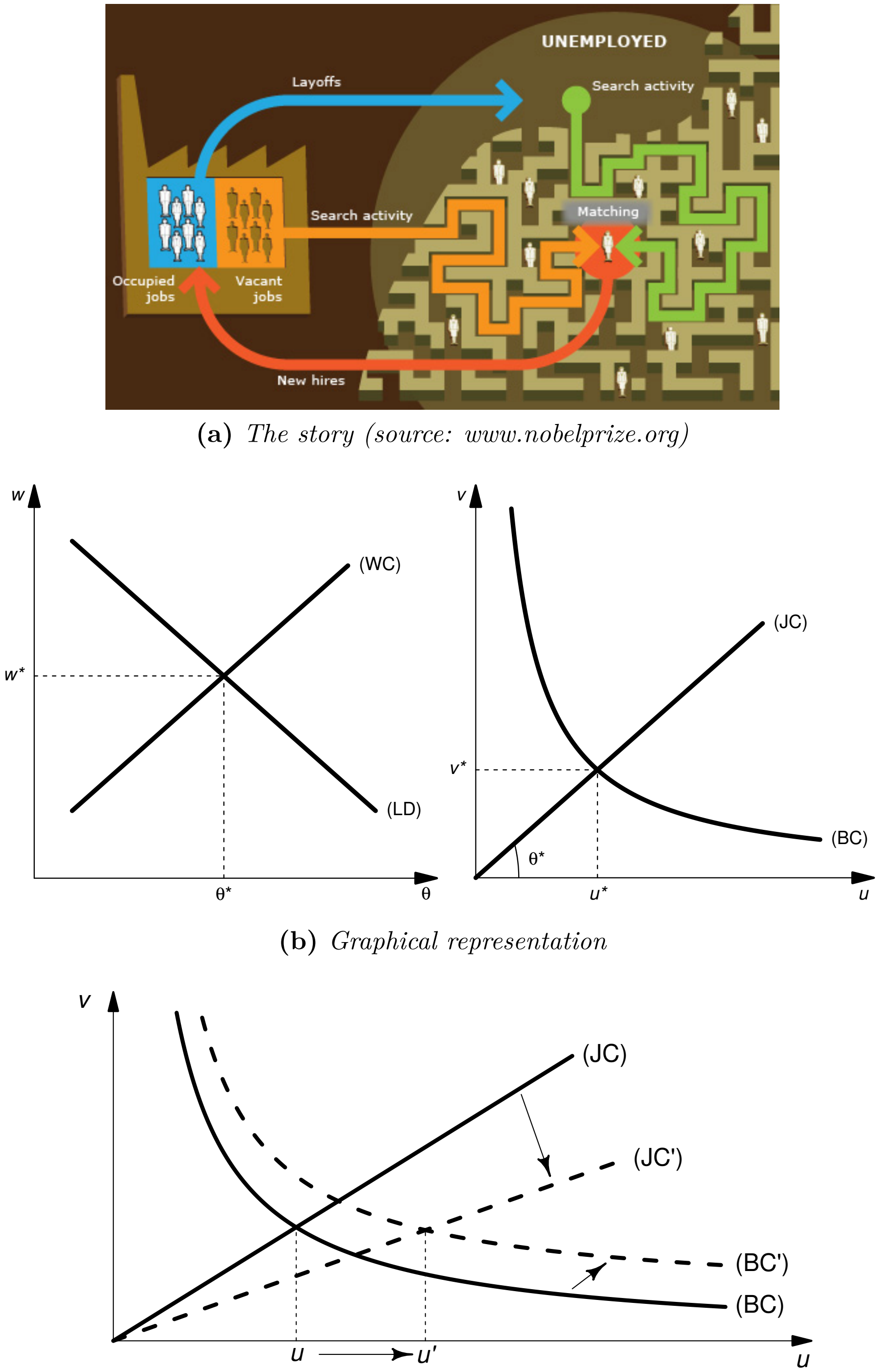
The second option is to claim that one should not have too strict a reading of propositions 1-3. These propositions might be statements about how things tend to be, or they might include an implicit proviso like 'most of the time'. In this watered-down form, they could be true of the target system. Such a strategy is generally acceptable, but it cannot however save the vending-machine view because the resulting structure is no more deductive. The statements being transformed into *generic* propositions, we now have a set of generics which cannot deductively imply the proposition 'unemployment rises when output growth drops'. If one is not wedded to the vending-machine view, this situation might be fine: there is still a link among these propositions even though it is weaker.

The last option, my personal favorite, is to accept the verdict, to keep the deductive structure, and to maintain that an adequate model for a given purpose does not need to be a 'perfect model'. There are multiple ways for an imperfect representation to be knowledge-conducive for its target, and some of these ways will become clear later on in this chapter.[14]

The second storyline won the Nobel Prize in Economics in 2010. It is formally developed in the DMP model of the labor market (DMP standing for Diamond, Mortensen and Pissarides). While the first story is centered on the notion of potential output, the core idea this time is that the labor market is a matching system with search frictions. The Royal Swedish Academy of Sciences made a nice picture to represent the storyline, which I reproduce in panel (a) of figure 5.2. The labor market is like a labyrinth where job seekers and employers look for each other. There is continuously an inflow in the pool of unemployed, mainly because specific firms terminate contracts for a variety of reasons. There is also an outflow as employers find the appropriate job seekers for their vacancies. At any time, the unemployment rate is a reflection of how many job seekers are lost in the labyrinth looking for a firm which might value their skills.

The mathematics associated to the DMP model is a bit more involved than the one associated to the first story (see Pissarides, 2000; Cahuc and Zylberberg, 2004). It involves three main equations and has an equilibrium in a 3-dimensional space (wage, vacancies, unemployment). This equilibrium is graphically represented in panel 5.2b, where w is the wage, v is the number of vacancies, u is the number of unemployed, and θ is the tightness of the labor market defined as the ratio of vacancies to the unemployed (v/u). The first two equations—the wage curve (WC) and labor demand (LD)—determine the equilibrium values for the wage and labor-market tightness. This relationship is depicted in the upper graph of panel 5.2b. The equilibrium tightness can then be combined with the Beveridge curve (BC)—a relationship between the number of unemployed and the number of vacancies meant to capture how efficient the labor market is at matching job seekers with firms—to determine the equilibrium value of the unemployment rate, as shown in the lower graph of panel 5.2b.

[14] There is a burgeoning literature on models and their epistemic functions. See, for instance, Morgan and Morrison (1999), Sugden (2000), Teller (2001), Giere (2006), Alexandrova (2008).



(a) *The story (source: www.nobelprize.org)*

(b) *Graphical representation*

(c) *Prediction for a recession*

Figure 5.2: *The DMP model*

The DMP model has been extremely influential. Its associated story is how most (labor) economists would informally depict the labor market. Policy proposals and generic 'shocks' are now routinely evaluated just by inserting them into the story. So what does it say about the behavior of unemployment when a recession hits?

The standard narrative of how lower (expected) output affects unemployment would be something like this. A recession can be understood as starting by a shock to the profitability of job matches—i.e. employers expect to get less profits both from their current workers and from new jobs they could create. This drop in profitability leads them to fire more workers and, especially, to open fewer vacancies. Panel 5.2c represents graphically the generic prediction (focusing only on the lower graph of panel 5.2b). Since firms are less willing to employ workers at a given wage, job creation drops—i.e. the job curve rotates clockwise. The Beveridge curve also shifts outward for two reasons. First, the spike in dismissals means that the pool of unemployed suddenly grows. Second, the ones fired may not have the skills sought for by the few firms having unfilled vacancies, which means that the matching process will be sluggish. In sum, a drop in output growth causes an upward change in unemployment.

The narrative might not seem tight enough, but the underlying model has a deductive structure. We are thus in the same situation as with the first story: the inverse relationship between output growth and unemployment can be *deduced* from the model. I claim that my second point also holds for this model: one should not believe the premises as true of real labor markets. It includes, for instance, that the only characteristic that job seekers care about is the wage, that they search at random for a new job, that the cost to firms of having an unfilled vacancy is proportional to time, that the wage rate for each match is given by the generalized Nash bargaining solution, and so on.[15]

[15] If the reader is still in the grip of the vending-machine view, I might have to add that these specific assumptions can be, and have been, tinkered with. It is however pure faith to assert that there will be an end to this process, that we will get the perfect model. All the improvements on the benchmark model result in propositions that are true of the model, but that are obviously false of the real world. This fact did not stop the Nobel Committee to praise the DMP model because it purportedly "help[s] us understand the ways in which unemployment, job vacancies, and wages are affected by regulation and economic policy." (source: www.nobelprize.org, accessed on 31-01-2012)

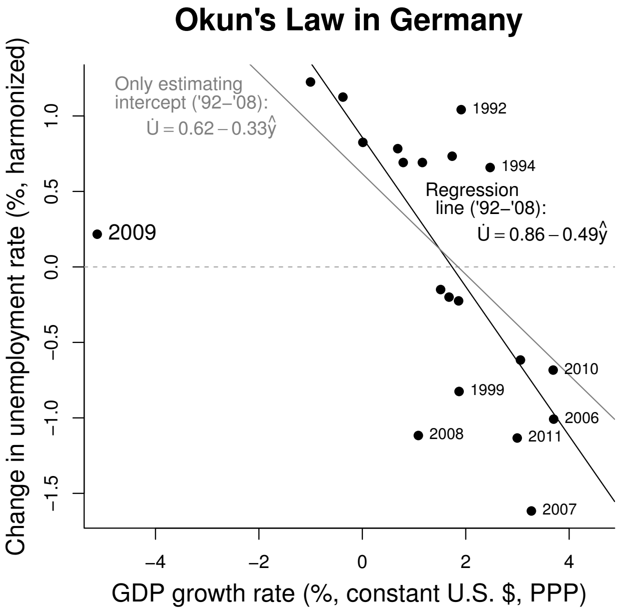
5.3.3 Dual-source expectation formation

When the financial crisis of 2008 came on the radar, economists started to expect higher unemployment in developed economies, including in Germany. The two previous subsections offer an explanation of this firm expectation. There is a clear empirical pattern known as Okun's law. There are also abstract models associated to stories about why output fluctuations are inversely related to changes in the unemployment rate.

These two elements were by far sufficient to crystallize expectations, although they do not square with the vending-machine view. Interestingly, each element met one, and only one, of the criteria in the vending-machine view. The expectation could not be deduced from Okun's law, but it was inferred from a believed premise (Okun's law itself).[16] In contrast, when abstract models are used to form expectations, we can ensure deductive entailment, but we fail regarding the criterion that the premises be believed. The next section will show that this particular structure of expectation formation has implications for the goal of deviant-case research.

It is about time I exhibit what happened in Germany during the crisis. The German deviance comes out starkly in figure 5.3, which plots the combinations of the German GDP growth rate and the change in its unemployment rate for every year since reunification. The observation for 2009 is a clear outlier. It was by far the worst year of reunified Germany if measured in output growth (−5.1 %). Nevertheless, the unemployment rate barely increased (+.2 %). There are multiple ways to illustrate how deviant this case is. Estimating Okun's law on the sample 1992-2008 gives the black diagonal line in figure 5.3 (since the slope is especially steep, I also include the gray line which imposes Okun's 3-to-1 link). Using the black line to predict the unemployment change in 2009, one would get the value of 3.3 % (2.3 % with the gray line), well off the mark. An interesting contrast is with the United States. If one takes the whole recession period, the German GDP dropped more than in the U.S. (6.6 % versus 4.5 %). This more severe decrease is associated with no change at all in the German unemployment rate (over the recession as a whole), while the U.S. unemployment rate jumped by 4.5 percentage points. No wonder that economists call this episode the German job miracle (e.g.

[16] To see where this claims come from, first remember (see section 5.2) that deduction is a small subset of all inferences. Second, according to my argument in subsection 5.3.1, Okun's law should be interpreted as a generic proposition and such propositions cannot play the role of universal statements in a deduction. Generics are nevertheless relied on to generate expectations, i.e. to infer what will happen.



Okun's Law in Germany

Source: OECD Statistics

Figure 5.3: *The German deviance*

Zimmermann, 2009; Boysen-Hogrefe and Groll, 2010; Möller, 2010; Burda et al., 2011; Klinger et al., 2011).

5.4 What is the epistemic goal?

The previous section offered a detailed analysis of how expectations crystallize in a specific scientific community which does not have the luck to have a theory working like a vending machine. This section and the next turn to the analysis of *research* on cases deviating from expectations. I first ask what the *epistemic* goal of this research is.

An implication of my analysis of expectation formation is that this goal is not to restore consistency among one's set of beliefs. I do not mean to say that, in eclectic sciences, restoring consistency is never one aim of deviant-case research. I mean that the strategy of looking for the false proposition—i.e. the strategy in the vending-machine view—is not generally applicable. For instance, the goal of researchers in my case study cannot be to restore consistency since there is no inconsistency in the first place. Okun's law being a generic proposition, it is not shown false by the German job miracle. We only know that the generic proposition misled us in this case—a possibility which is built in generic propositions. Similarly, the German case is not 'falsifying' our models. Mathematical models are not the sort of things which get falsified. We knew already that many propositions are true of our models but blatantly false of our target systems. Again, the models simply misled us in this case. Obviously, if our generic propositions and our mathematical models mislead us on too many occasions—'too many' being intentionally vague—we should revise both of them. But that's clearly not (yet) the situation for the relationship between unemployment and output fluctuations.

In the case of the German job miracle, I take economists *primarily* to seek a reasonably well-supported causal proposition accounting for *this* case. The titles of the articles on which I will draw seem sufficient evidence of this goal—e.g. 'The German labor market response in the world recession – de-mystifying a miracle' (Möller, 2010) and 'What Explains the German Labor Market Miracle in the Great Recession?' (Burda et al., 2011).

Do researchers care only about this case? Obviously not. What makes the German miracle particularly worth studying is the hope that it teaches us something more generic about what influence the job-output relationship. This hope is in turn linked to an *interventionist* goal: economists would love to repeat the miracle.[17] Should we then say that the *primary* epistemic goal is the formulation and justification of a new generic proposition? I think not. The easy answer is that getting the causal attribution right for this case is a necessary condition to be entitled to the generic claim—it is primary in this sense. The generic claim can indeed be understood as an upward extrapolation from the deviant case to the whole population.

This answer seems a bit too easy for two reasons. First, we will see that, when it comes to *justifying* the causal claim, it is convenient to rely on a comparability assumption across cases (Gerring, 2006, p. 718-23). This assumption makes the evidence for the deviant case, simultaneously evidence for the generic proposition. It is thus somewhat misleading to present deviant-case research as made of two separate steps—first establishing the claim for the deviant case, then extrapolating. Second, some

[17] Take, for instance, the OECD (2010, p. 16):

> [J]ob losses and the size of the increase in unemployment have ... differed markedly in countries where the fall in real GDP has been similar, raising the possibility that the right package of policies and institutions can significantly reduce the vulnerability of workers to cyclical unemployment.

scholars are so enthusiastic with the prospects of a new generic proposition that they almost skip the step of supporting the case-specific causal claim—Paul Krugman will be my example below.

In any case, my analysis below takes deviant-case research as having the justification of a case-specific causal claim as its primary epistemic goal. Achieving this goal remains a necessary condition to establishing the related generic proposition, although it is not strictly *temporally* prior and although some enthusiastic scholars might inappropriately skip the step.

5.5 How should research proceed?

One can articulate a causal proposition by unpacking its contrastive structure (Woodward, 2003, pp. 145-6; Schaffer, 2005). For deviant-case research, the proposition is about *actual* causation—i.e. it states that some causing actually occurred. Furthermore, it is about one system—in this case Germany. The contrastive structure of such a proposition is captured by the following schema:

$$c\,[C'] \hookrightarrow e\,[E'], \tag{5.4}$$

which reads '(For the system of interest,) the instantiation of $c \notin C'$ rather than some element in the set C' caused $e \notin E'$ to be instantiated rather than some element in the set E''. I take the 'elements' in question to be properties. Contemporary Germany, for instance, instantiates many properties like 'having a majority if its population speaking German' (a potential c or e) and fails to instantiate others like 'having a majority if its population speaking Esperanto' (a potential element of C' or E').[18] Now, deviant-case research can be reduced to

1. filling in the four elements of the causal schema (5.4);

2. justifying these choices.

I take each step in turn.

[18] I thus conceive the causal relata to be events in Kim's (1976) sense of "property exemplification", and I opt for an explicit contrastive structure. See (Schaffer, 2008) for a discussion of the metaphysics of causation.

5.5.1 Filling in the causal relata

The effect side—i.e. the explanandum—is fixed by the situation. Indeed, the values of e and E' are suggeted by the deviant/expected relationship. For the German job miracle, an obvious candidate for e is the actual unemployment change in 2009—i.e. $\Delta U_{2009}^{actual} = +.2$ %. The contrastive set E' includes an interval of potential values for ΔU_{2009} substantially above ΔU_{2009}^{actual}. While the *bounds* of this interval are not clearly specified, the interval must include only the values that would have been 'expected'. One way to represent this interval would be to center it at the point estimate given by Okun's law (e.g. the black or gray lines in figure 5.3) and to allow for some deviation around this value—i.e. $\Delta U_{2009}^{Okun} \pm \varepsilon$, where ε is a positive constant small enough to keep the actual value well outside the interval.

Disagreement arises when we turn to the cause side—i.e. the explanans. For the German job miracle, there is a long list of proposals for c and the elements of C'. These terms can be conjunctions of simpler elements and, indeed, each article in this literature argues for a few elements, and often argues against other elements.[19]

Before moving any further, I want to introduce the two examples of causal claims that I will use in the remainder of this chapter:

- **Krugman's claim.** In his New York Times column, the famous

[19] Here is a list of these simpler elements found in the literature on the German job miracle (the contrast classes C' are in squared brackets):

- strong employment protection legislation [a weak legislation];
- a short time work scheme [no scheme];
- use of 'working-time accounts' [no such accounts];
- wage moderation before the recession [normal wage growth];
- hiring shortfall before the recession [more jobs created];
- cyclical demand shock on export-oriented companies [structural crisis];
- recent skill shortages [ease to find appropriate workers];
- flexibility of working time [lower flexibility];
- room for negotiated wage adjustments [inflexible unions];
- 2005 reforms of the German labor market [no reform].

Each of these purported causes is identified in at least one of the following papers: Holland et al. (2009), Krugman (2009), Zimmermann (2009), Boysen-Hogrefe and Groll (2010), Elsby et al. (2010), Möller (2010), Boeri and Brücker (2011), Burda et al. (2011), Dietz et al. (2011), Klinger et al. (2011), Rinne and Zimmermann (2011).

economist Paul Krugman (2009) offered his explanation of Germany's success:

> Germany came into the Great Recession with strong employment protection legislation [EPL]. This has been supplemented with a "short-time work scheme," [STW] which provides subsidies to employers who reduce workers' hours rather than laying them off. These measures didn't prevent a nasty recession, but Germany got through the recession with remarkably few job losses.

Krugman fills in the causal schema (5.4) as

$$\mathrm{EPL}^{\mathrm{high}}\&\,\mathrm{STW}\,\left[\mathrm{EPL}^{\mathrm{low}}\&\,\neg\mathrm{STW}\right] \hookrightarrow \Delta U_{2009}^{\mathit{actual}}\,\left[\Delta U_{2009}^{\mathit{Okun}} \pm \varepsilon\right]. \tag{5.5}$$

- **Wage moderation.** Here is another claim which appeared early in the literature:

  > [W]age moderation in the years before the crisis is one of the most important factors explaining German labour market performance during the crisis. (Boysen-Hogrefe and Groll, 2010, p. R39)

  A rough way of expressing this claim in (5.4) is

  $$\hat{w}_{\mathrm{B2009}}^{\mathrm{low}}\,\left[\hat{w}_{\mathrm{B2009}}^{\mathrm{high}}\right] \hookrightarrow \Delta U_{2009}^{\mathit{actual}}\,\left[\Delta U_{2009}^{\mathit{Okun}} \pm \varepsilon\right] \tag{5.6}$$

  where $\hat{w}$ is the growth rate of wages and B2009 refers to "the years before the crisis".

Before turning to the challenging part of deviant-case research—i.e. justifying claims like (5.5) and (5.6)—it is worth asking how economists came up with these claims as *hypotheses* in the first place. Remember that, in the example of the boiling point of water, one could formulate a simple rule to fix the starting point of the research: start with the least-believed proposition. But this rule should be changed when, like in an eclectic science, deviance does not necessarily imply inconsistency of beliefs. In such cases, rather than orienting research toward finding what was falsely believed, one can seek what was true of the situation but has been so far overlooked. A reasonable starting question is thus: Which overlooked variables have a generic causal influence on the effect variable? The factors identified—which obviously come from background beliefs—are then further investigated.

The example about wage moderation fits perfectly this pattern. Faced with the deviant behavior of German unemployment, one is prompted to ask what is missing in Okun's law. What might be relevant to unemployment in addition to aggregate output? That an economist comes up with 'price' should not surprise anyone, as the relationship between the price of a good—the wage of workers—and the quantity exchanged is what Economics 101 is all about. It is both embedded in all economic models and literally expressed in the Law of Demand. See how the hypothesis flows out naturally:

> Okun's Law is critically flawed in the sense that it completely abstracts from the cost of labour, amongst other things. Clearly, the demand for labour also depends on the wage the employer has to pay. (Boysen-Hogrefe and Groll, 2010, p. R44)

A similar story could be given for the hypothesis involving employment protection legislation (EPL). Both of these hypotheses reveal part of the eclecticism of economics: researchers have a rich set of background knowledge on which they draw when deviance occurs. But what about the little known short time work schemes (STW)? What is most plausible in this case is that policy makers whispered the hypothesis in the ear of eager-to-listen economists like Krugman. Early in the crisis, the German government presented the use of a short time work scheme as its strategy to fight unemployment. When the deviant unemployment figures became public, the popular press presented it as a fact that the short time work scheme was preserving jobs (e.g. Atkins, 2009). To put it bluntly, Krugman seems to have simply jumped on the bandwagon.

In short, background beliefs about overlooked causal factors play a role in determining the starting point of deviant-case research, but sometimes the starting hypothesis comes out of peculiar locations.

5.5.2 Justifying the causal claim

Now that the causal schema (5.4) is filled in, we can turn to the justification of the resulting causal proposition. A warning: justification is a messy business in an eclectic science. To produce a principled discussion, I start by giving the skeleton of the justificative process. I then illustrate each step drawing on my two examples introduced above. My discussion will be both descriptive and normative: the steps in the justificative process are what the scientist and her epistemic community ought to do—and ought to do well—but scientists are already roughly aiming at this target, and if they fail, they are typiclly reminded of their duties by their fellows.

Skeleton of the justificative process. Justifying a claim about *actual* causation is done in two steps:

1. Justification of the description: Provide evidence that c and e actually occurred.
2. Difference-making justification: Provide evidence that the occurrence of c *made the difference* to the occurrence of e instead of (some elements of) E'.

One might think that the first step is trivial. It is not. But let me keep that for later. There are two paradigmatic strategies for the difference-making justification. Each strategy has its danger:

2a. Using established generic propositions.

> Danger: The propositions can be plain false or misleading for the case under study.

2b. Using inductive methods by drawing on other cases.

> Danger: The implicit comparability assumption may be unwarranted.

The difference-making justification can draw on one or on both strategies, and can do so with varying degrees of sophistication.

All of what has been said so far can be achieved by a single scientist, but the last and richest element in the justificatory process is properly *social*:

3. The dynamics of reasons: Other scientists ought to systematically question the proposed justification, others ought to question this questioning, and so forth.

Although all the elements of the skeleton are implicitly prescriptive, I put this last step as explicitly prescriptive because *it is the most important step.* I will not provide strong evidence for this bold claim—only weak evidence—but it is also not a novel claim: the crucial role of sustained peer criticism is generally recognized (e.g. Popper, [1963] 2002). And this role must be even bigger in eclectic sciences because students of these sciences cannot but use gappy reasonings—they have after all to rely on generic propositions, model worlds, and extrapolation. Someone ought to look in the gaps.

I now take each part of this skeleton in turn and put some flesh by borrowing from my examples.

Justification of the description. It is true that this step is a quick one for some causal factors. Krugman, for instance, can look up the OECD's index of employment protection to find that Germany is listed as having strong employment protection. He can also rely on the German government which is openly advertising its short time work scheme.[20]

The step is more demanding for other causal factors. Take the idea of wage moderation in my second example. To operationalize it, one needs data on average wages and a concept of 'normality'. Figure 5.4 captures what the authors identify as the normality and the moderation: the real wage (black line) had been steadily increasing from 1970 to 2003 (the normality), but it plateaued from 2003 to the onset of the recession. Wage moderation refers to this plateau. The identified moderation is thus a deviance of wage growth from what would have been expected based on a simple historical extrapolation. Putting the finger on this development is already an achievement.[21]

Difference-making justification. There are two paradigmatic strategies at this step, each with associated dangers. Krugman's short column—when we give it a charitable reading—provides a great illustration because the two strategies are present in simple forms.

First, relying on a generic proposition, Krugman states that employment protection and short time work schemes are "policies that support private-sector employment" (Krugman, 2009). It is indeed the case that the primary *goal* of the two policies is to prevent dismissals. Krugman

[20] There is one descriptive mistake in Krugman's quotation (see above p. 168): the German STW scheme is not a creation of the 2008-09 recession. This fact is important for the dynamics of reasons (see below).

[21] Beyond my two examples, other scholars do even more work at this justificative step. Möller (2010), for instance, deploys great efforts to establish that the German firms badly hit by the crisis were also the ones which suffered from labor shortages before the crisis. He puts so much effort at this step that he forgets the other step, difference-making justification.

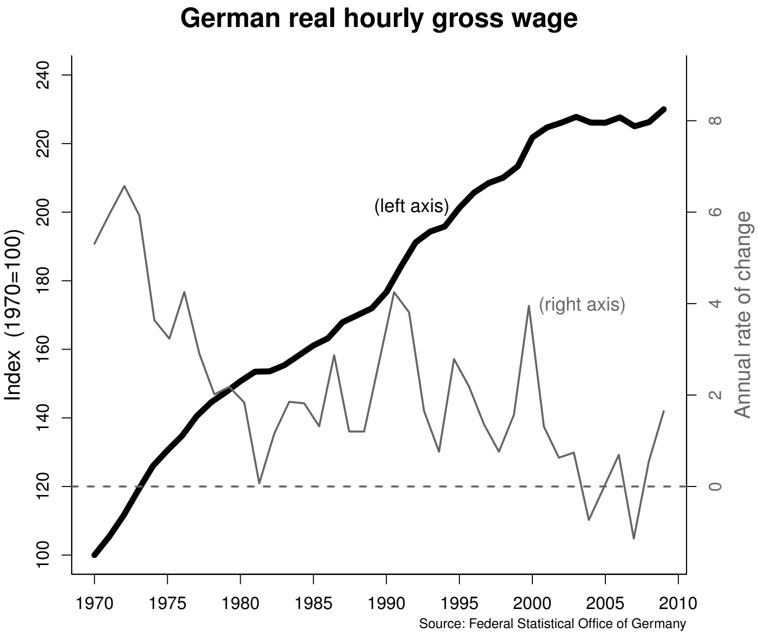


Figure 5.4: *Justifying the description in Boysen-Hogrefe and Groll (2010)*

seems to think that the policies achieve this goal. Second, Krugman makes an explicit comparison between Germany and the United States: the former has strong employment protection, a short-time work scheme, and experienced almost no rise in unemployment, the latter has weak employment protection, no short-time work scheme, and saw its unemployment rate skyrocket.

Krugman's justification is rather unsophisticated—it's a newspaper article after all. The dangers with the two strategies come out clearly. The first risk of using the generic proposition is that it might be false—the history of policy making is full of examples of ineffective policies. The reader might believe that policies such as employment protection and short time work schemes *must*, almost by definition, preserve jobs, but there are plenty of reasons—already present in the economic literature—for why they would be totally ineffective at keeping unemployment down.[22] The second risk is that, although true as a generic claim, the proposition could be misleading for Germany. Germany has already demarcated itself with

[22] For instance, employers might hire less knowing that employment protection makes it costly to fire; a short time work scheme might cause a further drop in production by inhibiting the restructuring process of the economy.

its unemployment performance in 2009, why should it follow the herd when it comes to the (generic) effect of these policies?

The inductive method that Krugman uses is also risky. One can think of it as a very rough attempt to use Mill's method of difference (Mill, 1886, p. 255). We would ideally want to observe 'Germany 2008-09' twice: this temporally-situated system with $\text{EPL}^{\text{high}} \& \text{STW}$, and the same system with $\text{EPL}^{\text{low}} \& \neg\text{STW}$. The impossibility of performing these two observations is what Paul Holland (1986, p. 947) famously called the Fundamental Problem of Causal Inference. This problem means that we have to compare recent Germany to units that will inevitably be unlike it in some respects. Krugman draws a simple comparison with the United States, but we all know that the two countries differ in many more respects than the three highlighted by Krugman. It does not mean that the comparability assumption must be dropped, but it certainly means that relying on it is risky.

Krugman proposes two weak arguments for his causal claim. He also uses a pervasive strategy in eclectic sciences: presenting arguments side by side such that his claim can benefit from the credibility boost coming with evidential variety (see part II of this thesis). Krugman's arguments remain shaky and have been disputed. The ensuing dynamics of reasons is fascinating. But first, I use my second example to show that the difference-making justification can be more sophisticated.

What arguments were given to support the claim that wage moderation made the difference for the German job miracle? Many arguments, some more sophisticated than others. I will focus on two arguments which revisit the two sources of expectations discussed in section 5.3.

The first one—not central but still present—relies on a formal model.[23] In the model world, the unemployment rate is shown to react less strongly to a negative shock in times following a period of wage moderation.[24] In other words, the authors introduce a different condition in the model: assume wages to be below their equilibrium value when the recession hits. With this assumption, the unemployment response to a negative shock is milder than without it (when wages take their equilibrium value). This new version of the model is then used in a *comparison* with actual Germany: if wage moderation causes a mild unemployment response in the model world, it might well have done the same in real Germany. We can also put the argument in terms of expectations: if we had taken

[23] It is a New Keynesian model to which labor market frictions are appended; see Boysen-Hogrefe et al. (2010), who rely on the model of Lechthaler et al. (2010).

[24] Boysen-Hogrefe and Groll (2010, p. R45) offer a simple narrative for their model.

wage moderation in due consideration, our model world (at least the one that the authors propose) would have led us to expect the German unemployment performance (cf. subsection 5.3.2).

The second argument—which is central in Boysen-Hogrefe and Groll (2010)—revisits the other source of expectations: the empirical pattern. The strategy is again to maintain that the deviance would have been expected provided the right factors had been considered:

> We would like to answer the following question: in the first quarter of 2008, how would we have predicted the evolution of total hours worked and employment throughout 2008 and 2009, had we known in advance the actual development of real GDP and real wages during the forecasting period? In other words, just how surprising was the German labour market given the actual evolution of the explanatory variables? (Boysen-Hogrefe and Groll, 2010, p. R42)

Roughly put, the substantive innovation of the authors is to add wages in a modernized specification of Okun's relationship.[25] After estimating the parameters of this specification with German data between 1970 and early 2008, they forecast the evolution of employment in the recession (2008 to early 2010) given the actual evolution of the other variables. What they report is that the 'deviant' German unemployment would have been (roughly) expected if we had used their empirical model (instead of Okun's law and model narratives) to form our expectations.

Both arguments, although more sophisticated, share the weak point of Krugman's comparison between the United States and Germany: the comparability assumption. First, the comparison between the model world and Germany might be inappropriate: wage moderation could well fail to cause a mild unemployment response in Germany, though it succeeds in the model world. Second, the modernized empirical analysis—when interpreted as not simply a forecasting device but primarily as a detector of causation—requires that 'old' Germany and 'new' Germany be comparable in the relevant respects. In particular, it requires that the new behavior of wages (i.e. flattening) and the new behavior of unemployment (i.e. mildly responding to a drop in the growth rate) are not effects of a common cause left out of the comparison.

[25] I say that the specification is 'modernized' because it draws on vector autoregression techniques—i.e. it includes a lag structure and an error correction term, and is checked for autocorrelation of the residuals and for structural stability.

The dynamics of reasons. While the causal claim about wage moderation relies on more sophisticated arguments than Krugman, these arguments are still gappy. The last step of the justificative process—the social step which can well be unending—is when these gaps are scrutinized. I will not use the wage moderation example to illustrate this step.[26] The debate following Krugman's explanation is rich enough. In fact, I will only look at the part about the short-time work scheme.[27]

Krugman's implicit generic proposition is something like 'In a recession, having a short-time work scheme saves a *substantial* number of jobs.' What should one expect based on this proposition? Critics have pointed out that, during a recession, countries with short time work schemes should experience a smaller rise in unemployment than countries without them. This is something to be expected if the generic proposition is true.

Some scholars looked across countries and back in time whether this expected pattern is present (Boysen-Hogrefe and Groll, 2010). It turns out that Germany's short time work scheme is a century old, and that many countries have one too. Furthermore, the percentage of German workers participating in the scheme during the recent recession is comparable to previous German recessions and to a few other countries during the 2008-9 recession. In these other cases, unemployment fluctuations were however not deviant. Why only Germany in 2009?

This sounds like a powerful argument against Krugman's generic proposition. Why would the proposition be true if short time work schemes do not seem to make a difference elsewhere? The argument is nevertheless inconclusive with respect to Krugman's causal claim for Germany. Say that, instead of upholding his generic proposition, we opt for the opposite proposition: 'In a recession, having a short-time work scheme does *not* save a *substantial* number of jobs.' It might still be that Krugman is right for Germany since, being generic, this new proposition allows for exceptions. Perhaps Germany in 2009 is such an exception. Another way to put it is that the comparability assumption implicit in the above argument might not hold—perhaps Germany in 2009 cannot be straightforwardly compared to the German past or to other countries with a short time work scheme. Indeed, quite a few economists argue just that

[26] The interested reader can start by Burda et al. (2011) in which it is argued that many variables beside wages behave 'abnormally' from (at least) 2005 to 2008. These include average hours worked, employment itself, and business expectations. Why point to wage moderation as the primary culprit?

[27] For a criticism of the claim about employment protection legislation, see, inter alia, Möller (2010).

(see, for instance, Boeri and Brücker, 2011; Dietz et al., 2011; Rinne and Zimmermann, 2011).

In substance, these economists claim that Germany in 2009 is doubly deviant—it deviates both from Okun's law and from the generic ineffectiveness of short time work. They do not only claim it, they provide justifications for their view. According to these authors, both the characteristics of the German short time work scheme (which was reformed before and early in the recession) and the characteristics of the German recession would have made the short time work scheme an important contributor to the job miracle.[28]

These justifications rely on other generic propositions and on inductive methods, and all these justifications can be questioned and further investigated. No need of discussing these justifications since my main point with this example should be clear enough: In eclectic sciences like I consider here, deviant-case research relies on shaky propositions at every junction. By probing the gaps in one's reasoning, the dynamics of reasons generates a web of propositions of a great complexity. The intellectual contributions of the various researchers and the resulting web of propositions are other manifestations of eclecticism.

5.6 Conclusion

Post-positivist philosophers of science have a nice story about deviant-case research. This story is however at odds with the actual research on deviant cases in eclectic sciences. This would be the conclusion if a bold generalization was drawn based on the example studied in this chapter.

The reason why the post-positivist story does not fit the example of this chapter is that economists cannot rely on a neat theory, made of trustworthy proposition, which works like a vending machine. Their expectations have to come from elsewhere. I have argued that they come from generic propositions about empirical patterns and from what is found to happen in model worlds (or their associated stories). As a consequence, the recognition of a case as deviant does not imply that one must solve an inconsistency problem among beliefs.

When the goal is not to restore consistency—like in the research on the German job miracle—the target is rather to formulate and justify a causal claim accounting for the deviance. In attempting to accomplish this task,

[28] No one claims that having a short time work scheme is a sufficient cause. Every scholar gives a multiple factor list for the German job miracle.

economists come to rely on a host of generic propositions which were left in the background when everything was running smoothly. These background propositions are both used to generate hypotheses and to justify causal claims. The other justificative strategy produces new beliefs through inductive methods. All the justificative strategies are gappy. This fact calls for the critical scrutiny of the community. The resulting dynamics of reasons is what prevents us from accepting quickly a claim which could well turn out to be false.

These points are the main lines of my alternative story. It is a story about variety and combination, a story about eclecticism. Is this story appropriate for all instances of deviant-case research in eclectic sciences? Most probably not. After all, one key point of the emerging view to which this chapter contributes is that science is diverse. It remains to be seen how far my story can travel.

Conclusion

The five core chapters of this thesis covered many aspects of causal reasoning in economics. They, however, touched on some interesting aspects only too briefly, if at all. In this brief conclusion, I first offer a short recapitulation of the overall narrative and the main points. I then gesture toward future lines of investigation that would contribute to our understanding of causal reasoning in economics.

If there is something that can be said about science in general—and perhaps about the distinction between science and non-science—it is that a practice which is meant to be scientific ought to be pursued with a critical attitude. The propositions generated in the course of this practice must be subject to sustained probing. A healthy scientific community is ready to give reasoned arguments for the propositions that it accepts, and encourages its members to request such reasons. Karl Popper, among others, emphasized this characteristic of science:

> My thesis is that what we call 'science' is differentiated from the older myths not by being something distinct from a myth, but by being accompanied by a second-order tradition—that of critically discussing the myth. (Popper, [1963] 2002, p. 170)

Critically discussing a scientific practice and its products requires an understanding of this practice. The second-order reflection of a scientific community can go seriously wrong if its self-understanding is limited, if it entertains widely implausible beliefs about its modes of reasoning. To me, a central goal of philosophy of science is to foster a better understanding of scientific practices, such that scientific communities can rely on this improved understanding in their self-reflective moments. Philosophy of science achieves this goal to the extent that it supplies abstract conceptual frameworks that do justice to the variety of scientific practices.

Philosophy of science can also impinge upon lucid self-reflection if its conceptual frameworks amount to tying some scientific practices to a Procrustean bed. Indeed, the material in this thesis leads me to the conclusion that philosophers of science have to work hard if they want to do justice to causal reasoning in economics. Looking especially at the research on the causes of aggregate unemployment, I develop, in the previous five chapters, conceptual frameworks to better understand semantic, epistemic, and dynamical aspects of causal reasoning.

For the semantic aspect, Luis Mireles-Flores and I argue, in chapter 1, that the meaning of causal generalizations in policy-oriented economics is best captured by using the concept of 'inferential relation' rather than the one of 'referential relation'. We further distinguish between different types of inferential relations, which constitute together the meaning of a causal generalization. Our framework has the advantage of showing why the widespread practice of demanding and supplying causal generalizations is epistemically reasonable. It also guards one against requiring fundamental reforms of this practice for it to comply with the standards of a semantics focused on the referential relation.

For the epistemic aspect (part II), my key concept is ‘evidential variety’. I argue that the justification of a causal claim tends to rely on multiple sources of evidence. It is thus important to transcend single-source assessment, which has typically been the focus in methodological discussions. Evidential variety is typically understood to be closely connected with another concept, i.e. ‘independent evidence’. I offer an interpretation of this concept in terms of error or reliability independence. I further investigate, in a Bayesian framework (chapter 4), whether more reliability independence is always conducive to higher confirmation, *ceteris paribus* (i.e. the variety-of-evidence thesis). I find out that, in extreme epistemic situations, the relationship between more independence and more confirmation is reversed.

For the dynamical aspect, I focus on deviant-case research. I try to answer three questions about this type of research. What makes a case deviant? What is the epistemic goal of deviant-case research? Given this goal, how should research proceed? I argue that we will fail to understand (at least) some deviant-case research as long as our answers to these questions are based on an influential, theory-centered conception of science—what Nancy Cartwright calls the “vending machine view”. I propose an alternative conception which I label ‘eclectic science’. The key difference between the two conceptions when it comes to deviant cases is that expectations are formed by deduction from believed premises according to the first, and from the combination of different informational sources according to the second. In chapter 5, I articulate answers to my three questions about deviant-case research in an eclectic science.

This thesis, I believe, contributes to a better understanding of causal reasoning in economics. I am, however, deeply aware that much more work is required before we might feel that we understand sufficiently this rich practice. One task for me is to develop further the conceptual frameworks proposed in this thesis. The inferentialist semantics of causal generalizations in part I is in an early phase of development. It needs much more careful thoughts before it can claim to have attained the degree of conceptual refinement of the referentialist approach. The study of evidential variety in part II only scratches the surface of this complex topic. I have, for instance, avoided cases in which the different evidential elements are discordant; but there is no doubt that these cases largely outnumber the cases of fully concordant evidence. In part III, studying an instance of deviant-case research has been my way to contribute to an understanding of the dynamics of beliefs among economists. The broad question was: How are causal beliefs affected by incoming facts? My contribution falls quite short of a complete answer to this question. It is first important to evaluate whether my answer works for other instances of deviant-case research. Second, my answer should be more tightly connected to the research dynamics of eclectic sciences at large, not only deviant-case research.

More work is also required on topics which recur in the three parts of my thesis but are discussed only superficially. I am thinking in particular about three topics: the commonalities between scientific generalizations and common general statements (generics), the nature and roles of models, and the community dimension of scientific reasoning. Further investigations on these three topics promise to be fascinating. I conclude by shortly discussing each topic in turn.

Philosophers of science have been debating for a long time the status of generalizations in the 'special' sciences. The basic challenge is that their surface structure suggests that they are universally quantified statements, but they are certainly false under this reading. It might come as a relief for philosophers of science to learn that philosophers of natural languages face a similar, if not identical, challenge in trying to analyze the meaning of what they call 'generics' (Krifka et al., 1995).[29] Unfortunately, philosophers of language to not have a consensual answer for the semantics of generics, but bringing the two literatures together is promising. For instance, it seems to me that psychological research on generics could be brought to bear on our understanding of scientific generalizations (e.g. through the work of Leslie, 2007, 2008). In my interpretation, the result of this research brings support to an inferentialist semantics of scientific generalizations, but more work is required to support this interpretation.

Modeling is a highly popular topic in contemporary philosophy of science, and for good reasons. Models are obviously important in economics, but my thesis does not analyze them in any great detail. It does include short discussions of the DMP model (subsections 2.5.1, 3.5.2, and 5.3.2), but much more could be said. One thing is that I present this model as being a source of evidence and expectations, but I do not justify this

[29] I point to the literature on generics in subsections 1.4.1 and 5.3.1.

interpretation. Many philosophers would maintain, in an empiricist fashion, that such a highly idealized model cannot be a source of evidence for empirical propositions. Though I think that my interpretation can be defended, no defense as been given here. Another thing to add is an investigation on the nature of models. There are, in particular, connections to be made between my treatment of generalizations and a plausible understanding of models as “instruments of investigation” (Morgan and Morrison, 1999). Both objects would be given their status—as generalizations or as models—by our use of them in our inferential practices. In other words, our use of them would be constitutive of their identity.

I have kept the most exciting and ambitious extension for the end. The idea that scientific practices are community-level phenomena—not something accomplished by a researcher in isolation—is present throughout my thesis but little explored. It comes up first in the semantic part in which generalizations are understood through their role in the communication process from expert economists to policy makers. It is also implicit in the epistemic part since the various evidential elements are supplied by different researchers. It finally returns in the part on dynamics, most clearly in my discussion of the dynamics of reasons. A more systematic investigation of the social dimension of causal reasoning will require a real engagement with the burgeoning literature in social epistemology (Fuller, 1988; Longino, 1990; Goldman, 1999). It should be able to investigate various issues including (i) the possibility of shifts in the meaning of a statement as it travels in the epistemic community, (ii) the research dynamics generated by disagreements over the evidential strength of some result, (iii) the possibility of belief convergence even though the relevant information is complex and *prima facie* discordant. Addressing these issues seems of primary importance to better understand causal reasoning in economics.

Appendices

Appendix A

Proofs for Chapter 4

A.1 Bovens and Hartmann's version

A.1.1 Posterior in the likelihood-ratio form

By Bayes' rule and the Law of Total Probability, we can rewrite this posterior

$$\begin{aligned} P^*(h) =& P(h|e_1, e_2) \\ =& \frac{P(e_1, e_2|h)P(h)}{P(e_1, e_2)} \\ =& \frac{P(e_1, e_2|h)P(h)}{P(e_1, e_2|h)P(h) + P(e_1, e_2|\neg h)P(\neg h)}. \end{aligned}$$

Dividing the numerator and the denominator by $P(e_1, e_2|h)$, we get

$$\begin{aligned} =& \frac{P(h)}{P(h) + P(\neg h)\frac{P(e_1,e_2|\neg h)}{P(e_1,e_2|h)}} \\ =& \frac{h_0}{h_0 + \bar{h}_0 L} \end{aligned}$$

where the last line gives us expression (4.5) by replacing $P(h)$ and $P(\neg h)$ by their values.

A.1.2 Likelihood ratios for the two versions

We get the two likelihood ratios L_I and L_S by using the probabilistic information encoded in Figure 4.1 together with equations (4.1), (4.2) and (4.4). Starting with the likelihood ratio for the independent-reliability version:

$$\begin{aligned}
L_I =& \frac{P_I(e_1, e_2|\neg h)}{P_I(e_1, e_2|h)} \\
=& \frac{\sum\limits_{R_1,R_2} P_I(e_1, e_2|\neg h, R_1, R_2) P_I(R_1) P_I(R_2)}{\sum\limits_{R_1,R_2} P_I(e_1, e_2|h, R_1, R_2) P_I(R_1) P_I(R_2)} \\
=& \frac{\sum\limits_{R_1,R_2} P_I(e_1|\neg h, R_1) P_I(R_1) P_I(e_2|\neg h, R_2) P_I(R_2)}{\sum\limits_{R_1,R_2} P_I(e_1|h, R_1) P_I(R_1) P_I(e_2|h, R_2) P_I(R_2)}
\end{aligned}$$

Given that the terms in the multiplications are either solely about source 1 or source 2, we can factorize by source:

$$\begin{aligned}
=& \frac{\prod\limits_{i=\{1,2\}} \sum\limits_{R_i} P_I(e_i|\neg h, R_i) P_I(R_i)}{\prod\limits_{i=\{1,2\}} \sum\limits_{R_i} P_I(e_i|h, R_i) P_I(R_i)} \\
=& \frac{\prod_i \left[P_I(e_i|\neg h, r_i) P_I(r_i) + P_I(e_i|\neg h, \neg r_i) P_I(\neg r_i)\right]}{\prod_i \left[P_I(e_i|h, r_i) P_I(r_i) + P_I(e_i|h, \neg r_i) P_I(\neg r_i)\right]} \\
=& \frac{\prod_i [0\rho_i + \alpha_i \bar{\rho}_i]}{\prod_i [1\rho_i + \alpha_i \bar{\rho}_i]} \\
L_I =& \frac{\alpha_1 \alpha_2 \bar{\rho}_1 \bar{\rho}_2}{(\rho_1 + \alpha_1 \bar{\rho}_1)(\rho_2 + \alpha_2 \bar{\rho}_2)} \qquad \text{(A.1)}
\end{aligned}$$

The third line results from plugging in the parameter values; simplifying in the fourth line gives us our final equation.

Now for the shared-reliability version:

$$\begin{aligned}
L_S =& \frac{P_S(e_1, e_2|\neg h)}{P_S(e_1, e_2|h)} \\
=& \frac{\sum_R P_S(e_1, e_2|\neg h, R) P_S(R)}{\sum_R P_S(e_1, e_2|h, R) P_S(R)} \\
=& \frac{\sum_R P_S(e_1|\neg h, R) P_S(e_2|\neg h, R) P_S(R)}{\sum_R P_S(e_1|h, R) P_S(e_2|h, R) P_S(R)} \\
=& \frac{P_S(e_1|\neg h, r) P_S(e_2|\neg h, r) P_S(r) + P_S(e_1|\neg h, \neg r) P_S(e_2|\neg h, \neg r) P_S(\neg r)}{P_S(e_1|h, r) P_S(e_2|h, r) P_S(r) + P_S(e_1|h, \neg r) P_S(e_2|h, \neg r) P_S(\neg r)} \\
=& \frac{0 * 0\rho + \alpha_1 \alpha_2 \bar{\rho}}{1 * 1 * \rho + \alpha_1 \alpha_2 \bar{\rho}}
\end{aligned}$$

$$L_S = \frac{\alpha_1\alpha_2\bar{\rho}}{\rho + \alpha_1\alpha_2\bar{\rho}} \tag{A.2}$$

A.1.3 *Ceteris paribus* condition

$P(h|e_i)$ can be expressed in the likelihood-ratio form:

$$P(h|e_i) = \frac{h_0}{h_0 + \bar{h}_0 L^i}$$

This new likelihood ratio L^i can be easily computed by looking at the two likelihood ratios that were produced above and contract them to be applicable to a single e_i:

$$L^i = \frac{\alpha_i\bar{\rho}_i}{\rho_i + \alpha_i\bar{\rho}_i} \tag{A.3}$$

From this likelihood ratio, we see that giving the same values for ρ_i and α_i to the evidential sources across models will also equalize L^i across models. The likelihood ratios of the previous subsection can thus be rewritten:

$$L_I = \frac{(\alpha\bar{\rho})^2}{(\rho + \alpha\bar{\rho})^2} \qquad\qquad L_S = \frac{\alpha^2\bar{\rho}}{\rho + \alpha^2\bar{\rho}}$$

A.1.4 Independence versus shared reliability

Imposing the *ceteris paribus* condition, the variety-of-evidence thesis implies $P_I^*(h) > P_S^*(h)$. Using the results above, we have:

$$\frac{h_0}{h_0 + \bar{h}_0 L_I} > \frac{h_0}{h_0 + \bar{h}_0 L_S} \quad\Leftrightarrow\quad L_S > L_I \quad\Leftrightarrow\quad \frac{\alpha^2\bar{\rho}}{\rho + \alpha^2\bar{\rho}} > \frac{\alpha^2\bar{\rho}^2}{(\rho + \alpha\bar{\rho})^2}$$

$$\Leftrightarrow\quad (\rho + \alpha\bar{\rho})^2 > \bar{\rho}(\rho + \alpha^2\bar{\rho}) \quad\Leftrightarrow\quad \rho^2 + 2\alpha\bar{\rho}\rho + \alpha^2\bar{\rho}^2 > \bar{\rho}\rho + \alpha^2\bar{\rho}^2$$

$$\Leftrightarrow\quad (1 - \bar{\rho}) + 2\alpha\bar{\rho} > \bar{\rho} \quad\Leftrightarrow\quad 1 > 2\bar{\rho} - 2\alpha\bar{\rho} \quad\Leftrightarrow\quad .5 > \bar{\alpha}\bar{\rho}.$$

A.2 Model with unreliability as systematic bias

A.2.1 Likelihood ratio for the independent-reliability version

$$
\begin{aligned}
L_{I'} =& \frac{P_{I'}(e_1, e_2|\neg h)}{P_{I'}(e_1, e_2|h)} \\
=& \frac{\prod\limits_{i=\{1,2\}} \sum\limits_{R_i} P_{I'}(e_i|\neg h, R_i) P_{I'}(R_i)}{\prod\limits_{i=\{1,2\}} \sum\limits_{R_i} P_{I'}(e_i|h, R_i) P_{I'}(R_i)} \\
=& \frac{\prod_i \left[P_{I'}(e_i|\neg h, r_i)P_{I'}(r_i) + P_{I'}(e_i|\neg h, b_i^h)P_{I'}(b_i^h) + P_{I'}(e_i|\neg h, b_i^{\neg h})P_{I'}(b_i^{\neg h})\right]}{\prod_i \left[P_{I'}(e_i|h, r_i)P_{I'}(r_i) + P_{I'}(e_i|h, b_i^h)P_{I'}(b_i^h) + P_{I'}(e_i|h, b_i^{\neg h})P_{I'}(b_i^{\neg h})\right]} \\
=& \frac{\prod_i [0\rho_i + 1\alpha_i\bar{\rho}_i + 0\bar{\alpha}_i\bar{\rho}_i]}{\prod_i [1\rho_i + 1\alpha_i\bar{\rho}_i + 0\bar{\alpha}_i\bar{\rho}_i]} \\
=& \frac{\alpha_1\alpha_2\bar{\rho}_1\bar{\rho}_2}{(\rho_1 + \alpha_1\bar{\rho}_1)(\rho_2 + \alpha_2\bar{\rho}_2)}
\end{aligned}
$$

The last line is the same as equation (A.1) which proves that my model and Bovens and Hartmann's model agree when sources are reliability independent.

A.2.2 Likelihood ratio for the shared-reliability version

$$
\begin{aligned}
L_{S'} =& \frac{P_{S'}(e_1, e_2|\neg h)}{P_{S'}(e_1, e_2|h)} \\
=& \frac{\sum_R P_{S'}(e_1|\neg h, R)P_{S'}(e_2|\neg h, R)P_{S'}(R)}{\sum_R P_{S'}(e_1|h, R)P_{S'}(e_2|h, R)P_{S'}(R)} \\
=& \frac{0 * 0\rho + 1 * 1\alpha\bar{\rho} + 0 * 0\bar{\alpha}\bar{\rho}}{1 * 1\rho + 1 * 1\alpha\bar{\rho} + 0 * 0\bar{\alpha}\bar{\rho}} \\
=& \frac{\alpha\bar{\rho}}{\rho + \alpha\bar{\rho}}
\end{aligned}
$$

The last line is not equal to equation (A.2)—i.e., this version of the shared-reliability situation does not concord with Bovens and Hartmann's version. In fact, it is equal to equation (A.3), which expresses the likelihood ratio for a *single* positive report.

A.2.3 Independent reliability is always better

To fulfill the *ceteris paribus* clause, I again assume that the prior h_0 is the same for both models, that $\alpha_1 = \alpha_2 = \alpha$, and that $\rho_1 = \rho_2 = \rho$. Then we know that

$$P^*_{I'}(h) > P^*_{S'}(h) \quad \Longleftrightarrow \quad L_{I'} < L_{S'}$$

and the last inequality is easily proven to hold for all admissible parameter values:

$$L_{I'} < L_{S'} \quad \Leftrightarrow \quad \frac{\alpha^2\bar{\rho}^2}{(\rho+\alpha\bar{\rho})^2} < \frac{\alpha\bar{\rho}}{\rho+\alpha\bar{\rho}}$$

$$\Leftrightarrow \quad \alpha\bar{\rho} < \rho + \alpha\bar{\rho} \quad \Leftrightarrow \quad 0 < \rho$$

Thus, as soon as the prior probability that the evidential source(s) is (are) reliable is non-null, reliability-independent sources are epistemically preferable.

A.3 Extended Model

A.3.1 Posterior belief in the hypothesis

The posterior given two positive reports:

$$P^*_F(h) = \frac{h_0}{h_0 + \bar{h}_0 L_F}$$

Focusing on the likelihood ratio:

$$\begin{aligned}
L_F =& \frac{P_F(e_1, e_2|\neg h)}{P_F(e_1, e_2|h)} = \frac{\sum\limits_{R_1,R_2} P_F(e_1, e_2, R_1, R_2|\neg h)}{\sum\limits_{R_1,R_2} P_F(e_1, e_2, R_1, R_2|h)} \\
=& \frac{\sum\limits_{R_1,R_2} P_F(e_1|\neg h, R_1)P_F(e_2|\neg h, R_2)P_F(R_1, R_2)}{\sum\limits_{R_1,R_2} P_F(e_1|h, R_1)P_F(e_2|h, R_2)P_F(R_1, R_2)} \\
=& \frac{P_F(b^h_i, b^h_j)}{P_F(r_i, r_j) + 2P_F(r_i, b^h_j) + P_F(b^h_i, b^h_j)} \\
=& \frac{\omega_{hh}}{\omega_{rr} + 2\omega_{rh} + \omega_{hh}} = \left[1 + \frac{\omega_{rr} + 2\omega_{rh}}{\omega_{hh}}\right]^{-1}
\end{aligned}$$

where the second-to-last line uses the information in Table 4.2 and the last line uses Table 4.3a.

A.3.2 Conditions for the off-diagonal elements

The fact that the elements in Table 4.3a must sum up to 1 gives us the following restriction:

$$\omega_{rr} + \omega_{hh} + \omega_{\neg h \neg h} + 2(\omega_{rh} + \omega_{r\neg h} + \omega_{h\neg h}) = 1 \tag{A.4}$$

The *ceteris paribus* condition also requires that we keep the confirmatory power of each evidential element constant, i.e., $P_F(h|e_i)$. This amounts to keep the following likelihood ratio constant:

$$\begin{aligned} L^i_F =& \frac{P_F(e_i|\neg h)}{P_F(e_i|h)} = \frac{\sum\limits_{R_i,R_j} P_F(e_i, R_i, R_j|\neg h)}{\sum\limits_{R_i,R_j} P_F(e_i, R_i, R_j|h)} \\ =& \frac{\sum\limits_{R_i,R_j} P_F(e_i|\neg h, R_i) P_F(R_i, R_j)}{\sum\limits_{R_i,R_j} P_F(e_i|h, R_i) P_F(R_i, R_j)} \\ =& \frac{\sum_{R_j} P_F(b^h_i, R_j)}{\sum_{R_j} [P_F(r_i, R_j) + P_F(b^h_i, R_j)]} \\ =& \frac{\omega_{hr} + \omega_{hh} + \omega_{h\neg h}}{\omega_{rr} + \omega_{rh} + \omega_{r\neg h} + \omega_{hr} + \omega_{hh} + \omega_{h\neg h}} = \left[1 + \frac{\omega_{rr} + \omega_{rh} + \omega_{r\neg h}}{\omega_{hr} + \omega_{hh} + \omega_{h\neg h}}\right]^{-1} \end{aligned} \tag{A.5}$$

where the second-to-last line uses the information in Table 4.2 and the last line uses Table 4.3a.

We can also express L^i_F in terms of our two parameters ρ and α by using one of the two extreme cases of fully-shared or fully-independent reliability (see Tables 4.3b and 4.3c):

$$L^i_F = \left[1 + \frac{\rho}{\bar{\rho}\alpha}\right]^{-1}. \tag{A.6}$$

The following restriction follows from equating (A.5) and (A.6):

$$\frac{\rho}{\bar{\rho}\alpha} = \frac{\omega_{rr} + \omega_{rh} + \omega_{r\neg h}}{\omega_{hr} + \omega_{hh} + \omega_{h\neg h}}$$

One way for this equality to hold is when the numerator on the left-hand-side is equal to the numerator on the right-hand-side and the same for the denominators. Equating the numerators and the denominators in this fashion and using condition A.4, we reach this system of equation:

$$\begin{aligned}\omega_{rr} + \omega_{rh} + \omega_{r\neg h} &= \rho\\ \omega_{rh} + \omega_{hh} + \omega_{h\neg h} &= \bar{\rho}\alpha\\ \omega_{r\neg h} + \omega_{\neg hh} + \omega_{\neg h\neg h} &= \bar{\rho}\bar{\alpha}\end{aligned}$$

which is, in fact, saying that the marginal probabilities $P(r_i)$, $P(b_i^h)$ and $P(b_i^{\neg h})$ are kept constant as the degree of independence varies. We can solve this system of equation for the off-diagonal elements in terms of the diagonal elements, ρ and α (I omit the simple algebraic manipulations):

$$\begin{aligned}\omega_{rh} &= \rho + \bar{\rho}\alpha - .5(1 + \omega_{rr} + \omega_{hh} - \omega_{\neg h\neg h})\\ \omega_{r\neg h} &= \rho + \bar{\rho}\bar{\alpha} - .5(1 + \omega_{rr} - \omega_{hh} + \omega_{\neg h\neg h})\\ \omega_{h\neg h} &= .5(1 + \omega_{rr} - \omega_{hh} - \omega_{\neg h\neg h}) - \rho\end{aligned} \tag{A.7}$$

A.3.3 The derivative of the likelihood ratio

We can rewrite the likelihood ratio in (4.11) by using information from (A.7)

$$\begin{aligned}L_F =&\frac{\omega_{hh}}{\omega_{rr} + 2\omega_{rh} + \omega_{hh}}\\ =&\frac{\omega_{hh}}{\omega_{rr} + \omega_{hh} + 2(1 - \bar{\rho}\bar{\alpha}) - 1 - \omega_{rr} - \omega_{hh} + \omega_{\neg h\neg h}}\\ =&\frac{\omega_{hh}}{1 - 2\bar{\rho}\bar{\alpha} + \omega_{\neg h\neg h}}\\ =&\frac{(\bar{\rho}\alpha)^{1+\delta}}{1 - 2\bar{\rho}\bar{\alpha} + (\bar{\rho}\bar{\alpha})^{1+\delta}}\end{aligned}$$

where the last line uses condition (4.12). I take the derivative with respect to δ:

$$\begin{aligned}\frac{\partial L_F}{\partial \delta} =&\frac{(\bar{\rho}\alpha)^{1+\delta}\ln(\bar{\rho}\alpha)(1 - 2\bar{\rho}\bar{\alpha} + (\bar{\rho}\bar{\alpha})^{1+\delta}) - (\bar{\rho}\alpha)^{1+\delta}(\bar{\rho}\bar{\alpha})^{1+\delta}\ln(\bar{\rho}\bar{\alpha})}{[1 - 2\bar{\rho}\bar{\alpha} + (\bar{\rho}\bar{\alpha})^{1+\delta}]^2}\\ =&\frac{(\bar{\rho}\alpha)^{1+\delta}[(1 - 2\bar{\rho}\bar{\alpha})\ln(\bar{\rho}\alpha) + (\bar{\rho}\bar{\alpha})^{1+\delta}(\ln(\bar{\rho}\alpha) - \ln(\bar{\rho}\bar{\alpha}))]}{[1 - 2\bar{\rho}\bar{\alpha} + (\bar{\rho}\bar{\alpha})^{1+\delta}]^2}\\ =&\frac{(\bar{\rho}\alpha)^{1+\delta}[(1 - 2\bar{\rho}\bar{\alpha})\ln(\bar{\rho}\alpha) + (\bar{\rho}\bar{\alpha})^{1+\delta}\ln(\alpha/\bar{\alpha})]}{[1 - 2\bar{\rho}\bar{\alpha} + (\bar{\rho}\bar{\alpha})^{1+\delta}]^2}\end{aligned}$$

The variety-of-evidence thesis maintains that $\partial L_F/\partial \delta < 0$ for all admissible parameter values. Verifying this:

$$\frac{\partial L_F}{\partial \delta} < 0 \quad \Leftrightarrow \quad (1 - 2\bar{\rho}\bar{\alpha})\underbrace{\ln(\bar{\rho}\alpha)}_{<0} + \underbrace{(\bar{\rho}\bar{\alpha})^{1+\delta}}_{>0}\ln(\alpha/\bar{\alpha}) < 0, \tag{A.8}$$

which does not hold for some combination of parameter values (see Figure 4.4). The fact that two distinct regions of $\alpha \times \rho$ lead to a reversal of the inequality can be seen from expression (A.8). The first term is positive (i.e., contributing to a reversal of the relationship) iff $(1 - 2\bar{\rho}\bar{\alpha}) < 0$, or more intuitively, $.5 < \bar{\rho}\bar{\alpha}$. The second term is positive iff $\alpha > .5$. It follows that the two terms cannot be positive at the same time.

A.3.4 Posterior belief in reliability

Having a single reliable source is sufficient for the two positive reports e_1 and e_2 to be truth revealing. The posterior belief that at least one source is reliable is

$$\begin{aligned} P(r_1 \cup r_2|e_1, e_2) &= P(r_1, r_2|e_1, e_2) + 2P(r_i, b_j^h|e_1, e_2) + 2\underbrace{P(r_i, b_j^{\neg h}|e_1, e_2)}_{=0} \\ &= \frac{P(e_1, e_2|r_1, r_2)P(r_1, r_2) + 2P(e_1, e_2|r_i, b_j^h)P(r_i, b_j^h)}{P(e_1, e_2)} \\ P^*(r_1 \cup r_2) &= \frac{h_0(\omega_{rr} + 2\omega_{rh})}{h_0(\omega_{rr} + 2\omega_{rh}) + \omega_{hh}} = \left[1 + \frac{\omega_{hh}}{h_0(\omega_{rr} + 2\omega_{rh})}\right]^{-1}. \end{aligned}$$

For ease of manipulation, I use the last equality to define the variable DT (for distrust):

$$DT = P^*(r_1 \cup r_2)^{-1} - 1 = \frac{\omega_{hh}}{h_0(\omega_{rr} + 2\omega_{rh})}$$

Reusing a result in Appendix A.3.1, I also define a variable C (for strength of confirmation):

$$C = L_F^{-1} - 1 = \frac{\omega_{rr} + 2\omega_{rh}}{\omega_{hh}}$$

Distrust and confirmation are related as

$$C = \frac{1}{h_0 DT}$$

from which it follows that

$$\frac{\partial\, C}{\partial\, \delta} > 0 \quad \Leftrightarrow \quad \frac{\partial\, DT}{\partial\, \delta} < 0 \quad \Leftrightarrow \quad \frac{\partial\, P^*(r_1 \cup r_2)}{\partial\, \delta} > 0.$$

In words, confirmation increases with reliability independence iff posterior trust in the sources increases.

Appendix B

Samenvatting (Dutch Summary)

Veel algemene vragen met zeker filosofische belang over causale redeneringen in economische wetenschappen zijn vooralsnog gebrekkig beantwoord. Ten eerste, wat is de betekenis van causale redeneringen? Dit is een semantische vraag. Ten tweede, hoe kan een causale claim adequaat ondersteund worden door bewijsmateriaal? Dit is een epistemologische vraag. Ten derde, hoe kunnen causale overtuigingen beïnvloed worden door nieuwe feiten? Dit is een vraag over de dynamiek van overtuigingen.

Deze thesis draagt bij aan het beantwoorden van deze vragen door gebruik te maken van een op een casusstudie gebaseerde aanpak. De casusstudie die gebruikt is in deze thesis is het wetenschappelijk onderzoek naar de oorzaken van werkeloosheid. Door het bestuderen van de wetenschappelijke praktijk ben ik tot de formulering en verdediging gekomen van antwoorden op de drie hierboven genoemde vragen. Ik beweer niet dat de antwoorden algemene geldigheid hebben—waarschijnlijk zijn ze niet van betrekking op alle gevallen van causale redeneringen. Toch, draagt de conceptuele arbeid die verricht is in deze thesis bij aan een beter begrip van causale redeneringen in de economische wetenschappen en daarbuiten.

In deel I, het semantische gedeelte dat samen met Luis Mireles-Flores geschre-ven is, onderzoeken wij de betekenis van causale generalisaties in de economische theorie van werkloosheidsvraagstukken. We betogen dat de standaard benadering van betekenis misleidend is omdat zij ten onrechte de *referentiële* relatie aanwijst als hetgeen betekenis constitueert. Om een beter begrip te krijgen van de wijdverbreide praktijk van de vraag naar en het aanbod van causale generalisaties in wetenschappelijke disciplines zoals de economische, hebben we een andere benadering nodig die prioriteit geeft aan de *inferentiële* relatie boven de referentiële relatie. Wij dragen bij aan de ontwikkeling van deze alternatieve benadering door verschillende types inferentiële relaties te identificeren, die samen de betekenis van een uitdrukking bepalen.

In deel II, het epistemologische deel, beargumenteer ik dat rechtvaardiging in wetenschappen als de economische vaak afhankelijk is, en afhankelijk zou moeten zijn, van verscheidenheid in bewijsvoering—ofwel, de combinatie van bewijzen uit verschillende bronnen. Het herkennen van het belang van verscheidenheid in bewijsvoering is cruciaal om in het methodologische debat afstand te nemen van één bron beoordeling. Dit deel, het langste van mijn thesis, bestaat uit drie hoofdstukken. In hoofdstuk 2 beargumenteer ik dat een levendig debat in hedendaagse econometrie tussen de ontwerpgebaseerde en structurele benaderingen lijdt onder de voorkeur voor één bron beoordeling. In hoofdstuk 3 wend ik mijtot een debat in de filosofie dat zich afspeelt rondom wat bekend is geworden als de Russo-Williamson Thesis. Ik beweer dat Russo and Williamson (2007) op het verkeerde spoor zitten met hun lezing dat de zoektocht naar zowel verschilmakend als ook mechanistisch bewijs incompatibel is met standaard monistische benaderingen van causaliteit. Ik beweer dat deze zoektocht simpelweg een epistemologische strategie is om verscheidenheid in bewijsvoering te genereren. Uit deze wijdverbreide epistemologische strategie zijn er geen conclusies te trekken over de semantiek en de metafysica van causaliteit. In hoofdstuk 4 gebruik ik een Bayesiaans model om het waarheidsgehalte te onderzoeken van de variëteit van bewijsvoering these. De variëteit van bewijsvoering these stelt dat, *ceteris paribus*, de kracht van de confirmatie van een these door een verzameling van bewijsmateriaal toeneemt met de mate van de diversiteit aan de elementen in die verzameling. Met een aangepast model van Bovens and Hartmann (2002, 2003) kom ik tot de conclusie dat, ondanks dat de variëteit van bewijsvoering these een goede leidraad is in typische omstandigheden, het een slechte leidraad is in extreme omstandigheden (wanneer bronnen van bewijsmateriaal hoogstwaarschijnlijk onbetrouwbaar zijn).

In deel III, het deel over overtuigingsdynamiek, bestudeer ik afwijkende gevallen onderzoek. Een casus is afwijkend wanneer het zich niet gedraagt zoals verwacht. De gedragingen van de Duitse werkloosheidscijfers na de financiële crisis van 2008 is zo'n afwijkende casus. Afwijkende casussen hebben verschillende labels gekregen in post-positivistische wijsbegeerte van een wetenschapsgebied—bijvoorbeeld, falsificeerders of anomalieën. In hoofdstuk 5 beargumenteer ik dat de invloedrijke wetenschappelijk zienswijze die ik de “verkoopautomaat kijk” (Cartwright, 1999) noem, een misleidende indruk geeft van afwijkende gevallen onderzoek in wetenschappen zoals de economische. De kern van het probleem is dat de verwachtingen in deze wetenschappen niet het resultaat zijn van deducties uit overtuigende premissen. In dit hoofdstuk werk ik een alternatieve visie uit op afwijkende gevallen onderzoek in wetenschappen die ik eclectisch noem. Deze wetenschappen worden gekarakteriseerd door variëteit en combinatie; ze zijn niet gebouwd rondom een monolithische theorie.

Appendix C

Curriculum Vitae

François Claveau
claveau@fwb.eur.nl

Education

PhD candidate, Erasmus Institute for Philosophy and Economics, Erasmus University Rotterdam (September 2010 - December 2012 [expected]), Rotterdam (Netherlands).

Research Master in Philosophy and Economics, Erasmus University Rotterdam (September 2008 - August 2010), Rotterdam (Netherlands). *Cum Laude*

Thesis: *Policy Conversation in Economics: A Critical Appraisal of the New Mainstream Approaches*, supervised by Jack Vromen.

M.A., Economics, McGill University (2007-2008), Montréal. GPA 4.0/4.0

Essay: *Interdependent Preferences in Economic Theory*, supervised by J.C.R. Rowley and Francisco Alvarez-Cuadrado.

Qualifying Year for M.A. in Economics, McGill University (2006-2007), Montréal. GPA 4.0 / 4.0

B.A., Political Science and Philosophy, Université de Montréal (2003-2006). GPA 4.2 / 4.3

Research

AOS: Philosophy of Science; Philosophy of Economics; Philosophy of Causality; Philosophy of Statistics; Formal Epistemology

AOC: Philosophy of Medicine; Foundations of Decision and Game Theories; History of Economics; Econometrics; Labor Economics

Appointments

Current appointments

Postdoctoral researcher, Centre interuniversitaire de recherche sur la science et la technologies (CIRST), Montreal, since July 2012.

Co-editor, *Erasmus Journal for Philosophy and Economics*, `ejpe.org`, since January 2010.

Previous appointments

Co-organizer, *Graduate Conference in Philosophy of Science*, Rotterdam, 8-9 March 2012.

Research Assistant for Prof. Mary MacKinnon, Department of Economics, McGill University, September 2007 – August 2008.

Research Assistant for Prof. Peter Dietsch, Department of Philosophy, Université de Montréal, May 2006 – August 2008.

Grants and Awards

1. Social Sciences and Humanities Research Council (SSHRC), Postdoctoral Fellowship (January 2013 - December 2015), CAD 76,000.

2. Erasmus Institute for Philosophy and Economics (EIPE), Doctoral Scholarship (September 2010 - August 2013), CAD 80,000.

3. Social Sciences and Humanities Research Council (SSHRC), Doctoral Fellowship (September 2009 - August 2013), CAD 80,000.

4. Social Sciences and Humanities Research Council (SSHRC), Joseph-Armand Bombardier CGS Doctoral Scholarship (September 2009 - August 2012), CAD 105,000, *Declined.*

5. Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC), PhD Scholarship (September 2009 - August 2012), CAD 60,000, *Declined.*

6. Faculty of Philosophy, Erasmus University Rotterdam, Master's scholarship (September 2008 - August 2009), CAD 14,500.

7. Netherlands organization for international cooperation in higher education (Nuffic), HSP Huygens (September 2008 - August 2010), CAD 44,750.

8. Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC), Research Master's Scholarship (September 2007 - August 2009), CAD 30,000, *Accepted only for fall 2008.*

9. Social Sciences and Humanities Research Council (SSHRC), Master's Scholarship, (September 2007 - August 2008), CAD 17,500.

Publications

Peer-reviewed articles

1. Claveau, François (forthcoming), The Independence Condition in the Variety-of-Evidence Thesis, *Philosophy of Science.*

2. Claveau, François (2012), The Russo-Williamson Theses in the Social Sciences: Causal Inference Drawing on Two Types of Evidence, *Studies in History and Philosophy of Biological and Biomedical Sciences.*

3. Claveau, François (2011), Evidential Variety as a Source of Credibility for Causal Inference: Beyond Sharp Designs and Structural Models, *Journal of Economic Methodology* 18 (3): 231-253.

4. Claveau, François (2009), Interdependent Preferences and Policy Stances in Mainstream Economics, *Erasmus Journal for Philosophy and Economics* 2 (1): 1-28.

5. Dietsch, Peter and François Claveau (2008), Concurrence fiscale et responsabilité étatique, *Éthique publique* 10 (1): 34-44.

6. Claveau, François (2006), L'abondance chez Locke, *Arguments* 1 (1): 64-79.

Other contributions to academic journals

1. Bassett, David and François Claveau (2011), The Economic Entomologist: An Interview with Alan Kirman, *Erasmus Journal for Philosophy and Economics* 4 (2): 42-66.

2. Claveau, François (2010), Review of 'Conversations on Ethics' written by Alex Voorhoeve, *Éthique et Économique/Ethics and Economics* 7 (2).

3. Claveau, François (2006), Review of 'Revenu minimum garanti : comparaison internationale, analyses et débats' written by Lionel-Henri Groulx, *Canadian Journal of Political Science* 39 (3) : 695-698.

Presentations

1. *The Independence Condition in the Variety-of-Evidence Thesis,*

    - 14th Annual Pitt-CMU Graduate Student Philosophy Conference, Pittsburgh (USA), April 6-7 2012.
    - TiLPS-EIPE graduate workshop in Philosophy of Science, Tilburg (Netherlands), November 17th 2011.

2. *Investigating Deviant Cases in Economics: A Methodological Reflection*, INEM Conference, Helsinki (Finland), September 1-3 2011.

3. *Semantic Analysis of Causal Generalizations in Policy-Oriented Social Sciences* (with Luis Mireles-Flores),

    - SPSP Conference, Exeter (UK), June 22nd 2011.
    - CiTS Work-in-Progress Meeting, Campus of Kent University in Paris, June 10th 2011.
    - EUR-FWB Lunchtime Seminar, Rotterdam, October 14th 2011.

4. *Causal Inference about Macro Social Phenomena: What Makes Inference Credible?*, CiTS Work-in-Progress Meeting, Brussels (Belgium), January 10-11 2011.

5. *From Multiple Sources of Evidence to Warranted Inference: Conditions for Successful Triangulation*, First Dutch-Flemish Graduate Conference on Philosophy of Science and/or Technology , Ghent (Belgium), November 25-26 2010.

6. *Multiple Evidential Sources in Economics: Consensual Unemployment Policies by Triangulation*, INEM Conference, Birmingham, Alabama, November 12-14 2010.

7. *Causal Claims for Unemployment Policy: Weak Evidential Elements, Strong Evidential Set?* Poster presented at CiBaSS conference, Rotterdam, October 6-8 2010.

8. *Causality and Macroeconomic Policies Concerning Unemployment: Combining Multiple Sources of Evidence*, Work in Progress in Causal and Probabilistic Reasoning, Campus of Kent University in Paris, June 28-29 2010.

9. *Policy-Oriented Behavioral Economics: Discursive Analysis of a New Phase in the Economics & Psychology Movement*, HISRECO Conference, École normale supérieure de Cachan, June 3rd 2010.

10. *Interdependent Preferences in Economic Theory: Efficiency-Based Debates with the Extended Utility Approach*, EIPE PhD/ReMa Seminar, Erasmus University Rotterdam, October 20th 2008.

11. *Choosing our Story of Fiscal Interdependence*, Tax Competition: How to meet the normative and political challenge, Université de Montréal, August 28th-29th 2008.

12. *Tax Competition and the Responsibilities of the State* (with Peter Dietsch), Intentions & Motivations in International Relations, Université de Montréal, May 23th 2008.

13. *Concurrence fiscale et responsabilité étatique* (with Peter Dietsch), Centre de Recherche en Éthique de l'Université de Montréal (CRÉUM), October 24th 2007.

14. Reply to *A Brief History of Liberty*, presented by David Schmidtz (University of Arizona) at Montreal Political Theory Workshop, March 16th 2007.

Working and Unpublished Papers

1. Claveau, François (February 2012), *Deviant Cases in a Dappled World: Considerations from Recent Economics.*

2. Claveau, François and Luis Mireles-Flores (September 2011), *Semantic Analysis of Causal Generalizations in Policy-Oriented Social Sciences.*

3. Claveau, François (May 2010), *Policy-Oriented Behavioral Economics: Discursive Analysis of a New Phase in the Economics & Psychology Movement.*

4. Claveau, François (July 2008), *Choosing our Story of Fiscal Interdependence.*

5. Claveau, François (April 2008), *Reference-Dependent Preferences in Bosnia and Herzegovina.*

Teaching

Lecturer for *Histoire de la pensée économique* (ECN1600), Université de Montréal, October 2012-January 2013.

Tutorials for *Philosophy of Economics* (FEB12002X-10, main lecturer: Julian Reiss), Erasmus University Rotterdam, May-June 2011 and again in 2012 (5 x 1.5 hour per year).

What's Philosophy of Economics at EIPE? Introductory Seminar for the new Research Master Students (taught with Sine Bagatur, Luis Mireles Flores and Attilia Ruzzene), September 15th 2010.

Miscellaneous

Econometrics

Competent with the following statistical packages: R, Gretl, Stata, SPSS.

Claveau, François. (2009). *R Programs for Applied Time Series Analysis*. My workbook as an introduction to using R for time series analysis.

Additional Courses Attended (without being graded)

1. Evidence and the Foundations of Statistics (with Julian Reiss, EUR), spring 2012.
2. Philosophy of Matter (with Fred Muller, EUR), spring 2012.
3. Model Thinking (with Scott E. Page, coursera.org), spring 2012.
4. Decision and Game Theory (with Conrad Heilmann, EUR), fall 2011.
5. Current Affairs: Global Warming, Financial Crisis and Terror (with Julian Reiss, EUR), winter 2011.
6. Philosophy of Science and Social Science (with Conrad Heilmann, EUR), winter 2011.
7. Living Philosophical Insights from Dead Economists (with Geoffrey M. Hodgson, EUR), 2010-11.
8. Advanced Labor Economics (with Jan van Ours and Matteo Picchio, Tilburg U.), fall 2010.
9. Explanation in the Social Sciences (with Rogier De Langhe, EUR), spring 2010.
10. Introduction to Social Choice Theory: Rationality (with Harrie de Swart, EUR), spring 2010.
11. Applied Time Series Analysis (with Michael Sampson, Concordia University), fall 2009.
12. Macroéconométrie (with Alain Guay, UQAM), fall 2009.

13. Causality (with Julian Reiss, EUR), spring 2009.

14. Evolutionary Origins of Morality (with Jack Vromen, EUR), spring 2009.

Languages

French (native)
English (expert)
Dutch (advanced; diploma *Staatexamen II*)
Spanish (beginner)

References

Julian Reiss
Associate Professor
Faculty of Philosophy
Erasmus Universiteit
P.O. Box 1738
3000 DR Rotterdam
Netherlands
phone:
+31 10 408 8962
reiss@fwb.eur.nl

Kevin D. Hoover
Professor of Economics and Philosophy
Duke University
Box 90097
Durham, North Carolina
27708-0097 USA
phone: (919) 660-1876
kd.hoover@duke.edu

Peter Dietsch
Associate Professor
Department of Philosophy
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, Québec
H3C 3J7 Canada
phone: (514) 343-6482
peter.dietsch@umontreal.ca

Bibliography

Achinstein, P. (2005), "Four Mistaken Theses about Evidence, and How to Correct Them," in *Scientific Evidence: Philosophical Theories and Applications*, ed. Achinstein, P., Baltimore: Johns Hopkins Press, pp. 35–50.

Aldrich, J. (2006), "When Are Inferences Too Fragile to Be Believed?" *Journal of Economic Methodology*, 13, 161.

Alexandrova, A. (2008), "Making Models Count," *Philosophy of Science*, 75, 383–404.

Angrist, J. D. and Pischke, J. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

— (2010), "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24, 3–30.

Atkins, R. (2009), "Eurozone feels benefit of short-time work schemes," *Financial Times*, 6, october 29th.

Baker, D., Glyn, A., Howell, D. R., and Schmitt, J. (2005), "Labor Market Institutions and Unemployment: Assessment of the Cross-Country Evidence," in *Fighting Unemployment: The Limits of Free Market Orthodoxy*, ed. Howell, D. R., New York: Oxford University Press, pp. 72–118.

Bassanini, A. and Duval, R. (2006), "Employment Patterns in OECD Countries: Reassessing the Role of Policies and Institutions," *OECD Economic Studies*, 42, 7–86.

Bechtel, W. and Abrahamsen, A. (2005), "Explanation: A Mechanist Alternative," *Studies in History and Philosophy of Science Part C: Stud-*

ies in History and Philosophy of Biological and Biomedical Sciences, 36, 421–441.

Beebee, H. (2007), "Hume on Causation: The Projectivist Interpretation," in *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, eds. Price, H. and Corry, R., Oxford: Oxford University Press, pp. 224–249.

Belot, M. and van Ours, J. C. (2001), "Unemployment and Labor Market Institutions: An Empirical Analysis," *Journal of the Japanese and International Economies*, 15, 403–418.

Blanchard, O. (2006), "European Unemployment: The Evolution of Facts and Ideas," *Economic Policy*, 21, 5–59.

— (2007), "Review of "Unemployment: Macroeconomic Performance and the Labour Market"," *Journal of Economic Literature*, 45, 410–418.

— (2008), "The State of Macro," *NBER Working Paper Series.*

Blanchard, O. and Wolfers, J. (2000), "The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence," *Economic Journal*, 110, C1–C33.

Block, N. (1998), "Semantics, Conceptual Role," in *Routledge Encyclopedia of Philosophy*, ed. Craig, E., London: Routledge.

Boeri, T. and Brücker, H. (2011), "Short-Time Work Benefits Revisited: Some Lessons from the Great Recession," *IZA Discussion Paper.*

Boeri, T. and van Ours, J. (2008), *The Economics of Imperfect Labor Markets*, Princeton: Princeton University Press.

Bovens, L. and Hartmann, S. (2002), "Bayesian Networks and the Problem of Unreliable Instruments," *Philosophy of Science*, 69, 29–72.

— (2003), *Bayesian Epistemology*, Oxford: Oxford University Press.

Boysen-Hogrefe, J. and Groll, D. (2010), "The German Labour Market Miracle," *National Institute Economic Review*, 214, R38 –R50.

Boysen-Hogrefe, J., Groll, D., Lechthaler, W., and Merkl, C. (2010), "The Role of Labor Market Institutions in the Great Recession," *Applied Economics Quarterly (formerly: Konjunkturpolitik)*, 56, 65–88.

Brandom, R. (2007), "Inferentialism and Some of Its Challenges," *Philosophy and Phenomenological Research*, 74, 651–676.

Brigandt, I. (2010), "Scientific Reasoning Is Material Inference: Combining Confirmation, Discovery, and Explanation," *International Studies in the Philosophy of Science*, 24, 31–43.

Burda, M. C., Hunt, J., Elsby, M. W. L., and Haltiwanger, J. (2011), "What Explains the German Labor Market Miracle in the Great Recession? [with Comments and Discussion]," *Brookings Papers on Economic Activity*, 273–335.

Burian, R. M. (2001), "The Dilemma of Case Studies Resolved: The Virtues of Using Case Studies in the History and Philosophy of Science," *Perspectives on Science*, 9, 383–404.

Cahuc, P. and Zylberberg, A. (2004), *Labor Economics*, Cambridge, MA: MIT Press.

Campbell, D. T. and Fiske, D. W. (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56, 81–105.

Cartwright, N. (1979), "Causal Laws and Effective Strategies," *Noûs*, 13, 419–437.

— (1999), *The Dappled World: A Study of the Boundaries of Science*, Cambridge, UK: Cambridge University Press.

— (2007), *Hunting Causes and Using Them: Approaches in Philosophy and Economics*, Cambridge, UK: Cambridge University Press.

CERN (2012), "Neutrinos Sent from CERN to Gran Sasso Respect the Cosmic Speed Limit," Press release, 8 June, press.web.cern.ch/Press/PressReleases/Releases2011/PR19.11E.html (last accessed on 29 June 2012).

Chang, H. (2007), "The Myth of the Boiling Point," http://www.hps.cam.ac.uk/people/chang/boiling/.

— (2011), "Beyond Case-Studies: History as Philosophy," in *Integrating History and Philosophy of Science*, eds. Mauskopf, S., Schmaltz, T., Cohen, R. S., Renn, J., and Gavroglu, K., Springer Netherlands, vol. 263 of *Boston Studies in the Philosophy of Science*, pp. 109–124.

Contaldi, C. R. (2011), "The OPERA neutrino velocity result and the synchronisation of clocks," *arXiv:1109.6160.*

Craver, C. F. and Bechtel, W. (2007), "Top-down Causation Without Top-down Causes," *Biology & Philosophy*, 22, 547–563.

Cuaresma, J. C. (2008), "Okun's Law," in *The New Palgrave Dictionary of Economics*, eds. Durlauf, S. N. and Blume, L. E., Basingstoke: Nature Publishing Group, 2nd ed., pp. 178–180.

Culp, S. (1994), "Defending Robustness: The Bacterial Mesosome as a Test Case," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1994, 46–57.

— (1995), "Objectivity in Experimental Inquiry: Breaking Data-Technique Circles," *Philosophy of Science*, 62, 438–458.

Deaton, A. (2010), "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–455.

Denzin, N. K. (1978), *The Research Act: A Theoretical Introduction to Sociological Methods, Second Edition*, New York: McGraw-Hill.

Dessler, A. and Parson, E. (2010), *The Science and Politics of Global Climate Change, Second Edition*, Cambridge: Cambridge University Press.

Dietz, M., Stops, M., and Walwei, U. (2011), *Safeguarding Jobs in Times of Crisis – Lessons from the German Experience*, Geneva: International Labour Organization.

Dowe, P. (2000), *Physical Causation*, Cambridge: Cambridge University Press.

Downward, P. and Mearman, A. (2007), "Retroduction as Mixed-Methods Triangulation in Economic Research: Reorienting Economics into Social Science," *Cambridge Journal of Economics*, 31, 77 –99.

Dupré, J. (1984), "Probabilistic Causality Emancipated," *Midwest Studies In Philosophy*, 9, 169–175.

Earman, J. (1992), *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.

Earman, J., Roberts, J., and Smith, S. (2002), "Ceteris Paribus Lost," *Erkenntnis*, 57, 281–301.

Eells, E. (1991), *Probabilistic Causality*, Cambridge, MA: Cambridge University Press.

Elsby, M. W. L., Hobijn, B., and Şahin, A. (2010), "The Labor Market in the Great Recession [with Comments and Discussion]," *Brookings Papers on Economic Activity*, 1–69.

Feyerabend, P. K. (1999), *Knowledge, science, and relativism: 1960-1980*, Cambridge, UK: Cambridge University Press.

Fisher, R. A. (1935), *The Design of Experiments*, New York: Hafner Press.

Fitelson, B. (1996), "Wayne, Horwich, and Evidential Diversity," *Philosophy of Science*, 63, 652–660.

— (2001), "A Bayesian Account of Independent Evidence with Applications," *Philosophy of Science*, 68, S123–S140.

Fredriksson, P. and Holmlund, B. (2006), "Improving Incentives in Unemployment Insurance: A Review of Recent Research," *Journal of Economic Surveys*, 20, 357–386.

Freeman, R. B. (2005), "Labour Market Institutions Without Blinders: The Debate Over Flexibility and Labour Market Performance," *International Economic Journal*, 19, 129.

Frege, G. (1892), "On Sense and Reference," in *Translations from the Philosophical Writings of Gottlob Frege (1960)*, eds. Geach, P. and Black, M., Oxford: Basil Blackwell, p. 56–78.

Fuller, S. (1988), *Social Epistemology*, Bloomington: Indiana University Press.

Gerring, J. (2006), "Single-Outcome Studies," *International Sociology*, 21, 707 –734.

Giere, R. N. (2006), *Scientific Perspectivism*, University of Chicago Press.

Gillies, D. (2010), "The Russo-Williamson Thesis and the Question of Whether Smoking Causes Heart Disease," in *Causality in the Sciences*, eds. Illari, P. M., Russo, F., and Williamson, J., Oxford University Press.

Glennan, S. (1996), "Mechanisms and the nature of causation," *Erkenntnis*, 44, 49–71.

— (2002), "Rethinking Mechanistic Explanation," *Philosophy of Science*, 69, S342–S353.

— (2011), "Singular and General Causal Relations: A Mechanist Perspective," in *Causality in the Sciences*, eds. Illari, P. M., Russo, F., and Williamson, J., Oxford: Oxford University Press, pp. 789–817.

Goldman, A. I. (1999), *Knowledge in a Social World*, Oxford: Oxford University Press.

Gordon, R. A. (1965), "Full Employment as a Policy Goal," in *Employment Policy and the Labor Market*, ed. Ross, A. M., Berkeley: University of California Press, pp. 25–55.

Granger, C. W. (2007), "Causality in Economics," in *Thinking About Causes: From Greek Philosophy to Modern Physics*, eds. Machamer, P. and Wolters, G., Pittsburgh: University of Pittsburgh Press, pp. 284–296.

Haack, S. (2008), "Proving Causation: The Holism of Warrant and the Atomism of Daubert," *Journal of Health & Biomedical Law*, 4, 253–289.

Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Econometrica*, 12, iii–115.

Hall, N. (2004), "Two Concepts of Causation," in *Causation and counterfactuals*, eds. Collins, J., Hall, N., and Paul, L. A., Cambridge, MA: MIT Press.

Hands, D. W. (2001), *Reflection without Rules : Economic Methodology and Contemporary Science Theory*, New York: Cambridge University Press.

Hartmann, S. (2008), "Modeling in Philosophy of Science," in *Representation, Evidence, and Justification: Themes from Suppes*, eds. Frauchiger, M. and Essler, W. K., Heusenstamm: ontos verlag, pp. 95–121.

Hartmann, S. and Sprenger, J. (2011), “Bayesian Epistemology,” in *Routledge Companion to Epistemology*, eds. Bernecker, S. and Pritchard, D., London: Routledge, pp. 609–620.

Hausman, D. M. (1992), *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.

— (1998), *Causal Asymmetries*, Cambridge University Press.

— (2010), “Probabilistic Causality and Causal Generalizations,” in *The Place of Probability in Science*, eds. Eells, E. and Fetzer, J., Dordrecht: Springer Netherlands, vol. 284, pp. 47–63.

Heckman, J. J. (1992), “Randomization and Social Policy Evaluation,” in *Evaluating Welfare and Training Programs*, eds. Manski, C. F. and Garfinkel, I., Cambridge, MA: Harvard University Press, pp. 201–30.

— (2000), “Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective,” *Quarterly Journal of Economics*, 115, 45–97.

— (2005), “The Scientific Model of Causality,” *Sociological Methodology*, 35, 1–98.

— (2008), “Econometric Causality,” *International Statistical Review*, 76, 1–27.

— (2010), “Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy,” *Journal of Economic Literature*, 48, 356–398.

Heckman, J. J. and Vytlacil, E. J. (2007), “Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation,” in *Handbook of Econometrics, Volume 6B*, eds. Heckman, J. J. and Leamer, E. E., New York: Elsevier, pp. 4779–4874.

Hedström, P. and Ylikoski, P. (2010), “Causal Mechanisms in the Social Sciences,” *Annual Review of Sociology*, 36, 49–67.

Heim, I. and Kratzer, A. (1998), *Semantics in Generative Grammar*, Malden, MA: Blackwell.

Hitchcock, C. (2001a), “Causal Generalizations and Good Advice,” *The Monist*, 84, 218–241.

— (2001b), “A Tale of Two Effects,” *The Philosophical Review*, 110, 361–396.

— (2007), “How to Be a Causal Pluralist,” in *Thinking About Causes: From Greek Philosophy to Modern Physics*, eds. Machamer, P. and Wolters, G., Pittsburgh: University of Pittsburgh Press, pp. 200–221.

Holland, D., Kirby, S., and Whitworth, R. (2009), “Labour Markets in Recession: An International Comparison,” *National Institute Economic Review*, 209, 35–41.

Holland, P. W. (1986), “Statistics and Causal Inference,” *Journal of the American Statistical Association*, 81, 945–960.

Hood, W. C. and Koopmans, T. C. (eds.) (1953), *Studies in Econometric Method*, no. 14 in Cowles Commission Monograph, New York: John Wiley & Sons.

Hoover, K. D. (1990), “The Logic of Causal Inference: Econometrics and the Conditional Analysis of Causation,” *Economics and Philosophy*, 6, 207–234.

— (1991), “Scientific Research Program or Tribe? A Joint Appraisal of Lakatos and the New Classical Macroeconomics,” in *Appraising Modern Economics: Studies in the Methodology of Scientific Research Programs*, eds. De Marchi, N. and Blaug, M., Aldershot: Edward Elgar, pp. 364–394.

— (2001), *Causality in Macroeconomics*, Cambridge, MA: Cambridge University Press.

— (2004), “Lost Causes,” *Journal of the History of Economic Thought*, 26, 149–164.

— (2006), “Economic Theory and Causal Inference (FORTHCOMING),” in *Handbook of the Philosophy of Economics*, ed. Mäki, U., Amsterdam: Elsevier/North-Holland.

— (2008), “Causality in Economics and Econometrics,” in *The New Palgrave Dictionary of Economics, Second Edition*, eds. Durlauf, S. N. and Blume, L. E., Basingstoke: Palgrave Macmillan.

— (2009), "Review of Nancy Cartwright's 'Hunting Causes and Using Them: Approaches in Philosophy and Economics'," *Journal of Economic Literature*, 47, 493–495.

— (2011), "Counterfactuals and Causal Structure," in *Causality in the Sciences*, eds. Illari, P. M., Russo, F., and Williamson, J., Oxford: Oxford University Press, pp. 338–360.

— (2012a), *Applied Intermediate Macroeconomics*, New York: Cambridge University Press.

— (2012b), "Causal Structure and Hierarchies of Models," *Studies in History and Philosophy of Biological and Biomedical Sciences.*

Hoover, K. D. and Perez, S. J. (2004), "Truth and Robustness in Cross-Country Growth Regressions," *Oxford Bulletin of Economics and Statistics*, 66, 765–98.

Horwich, P. (1982), *Probability and evidence*, Cambridge: Cambridge University Press.

— (1998), "Wittgensteinian Bayesianism," in *Philosophy of Science: The Central Issues*, eds. Curd, M. and Cover, J. A., New York and London: W.W. Norton, pp. 607–624.

Howell, D. R. (ed.) (2005), *Fighting Unemployment: The Limits of Free Market Orthodoxy*, New York: Oxford University Press.

Howson, C. (1997), "A Logic of Induction," *Philosophy of Science*, 64, 268–90.

Howson, C. and Urbach, P. (1993), *Scientific Reasoning: The Bayesian Approach (Second Edition)*, Chicago: Open Court.

Hume, D. ([1739] 1975), *A Treatise of Human Nature*, Oxford: Clarendon Press, 2nd ed.

ICARUS Collaboration (2012), "Measurement of the neutrino velocity with the ICARUS detector at the CNGS beam," *Physics Letters B*, 713, 17–22.

Illari, P. M. (2011), "Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis," *International Studies in the Philosophy of Science*, 25, 139–157.

ILO (1982), “Resolution Concerning Statistics of the Economically Active Population, Employment, Unemployment and Underemployment,” The Thirteenth International Conference of Labour Statisticians: International Labour Organization, pp. 1–10.

Imbens, G. W. (2010), “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),” *Journal of Economic Literature*, 48, 399–423.

Imbens, G. W. and Angrist, J. D. (1994), “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.

Imbens, G. W. and Wooldridge, J. M. (2009), “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.

IPCC (2007), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge, UK: Cambridge University Press.

Istituto Nazionale di Fisica Nucleare (2011), “Particles Appear to Travel Faster Than Light: OPERA Experiment Reports Anomaly in Flight Time of Neutrinos,” *ScienceDaily*, retrieved October 26, 2011, from www.sciencedaily.com/releases/2011/09/110923084425.htm.

Keane, M. P. (2010), “A Structural Perspective on the Experimentalist School,” *Journal of Economic Perspectives*, 24, 47–58.

Kelly, T. (2008), “Evidence,” in *The Stanford Encyclopedia of Philosophy*, ed. Zalta, E. N., fall 2008 ed.

Kim, J. (1976), “Events as Property Exemplifications,” in *Action Theory*, eds. Brand, M. and Walton, D., Dordrecht: D. Reidel Publishing, pp. 159–77.

Kincaid, H. (1996), *Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research*, Cambridge, UK: Cambridge University Press.

— (2009), “Explaining Growth,” in *The Oxford Handbook of Philosophy of Economics*, eds. Kincaid, H. and Ross, D., Oxford: Oxford University Press, pp. 455–476.

Klinger, S., Rebien, M., Heckmann, M., and Szameitat, J. (2011), "Did Recruitment Problems Account for the German Job Miracle?" *International Review of Business Research Papers*, 7, 265–281.

Koopmans, T. C. (ed.) (1950), *Statistical Inference in Dynamic Economic Models*, no. 10 in Cowles Commission Monograph, New York: John Wiley & Sons.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Link, G., and Chierchia, G. (1995), "Genericity: An Introduction," in *The Generic Book*, eds. Carlson, G. N. and Pelletier, F. J., University of Chicago Press, pp. 1–124.

Krugman, P. (2009), "Free to Lose," *The New York Times*, http://www.nytimes.com/2009/11/13/opinion/13krugman.html.

Kugler, A. and Pica, G. (2008), "Effects of Employment Protection on Worker and Job Flows: Evidence from the 1990 Italian Reform," *Labour Economics*, 15, 78–95.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kuorikoski, J. (2009), "Two Concepts of Mechanism: Componential Causal System and Abstract Form of Interaction," *International Studies in the Philosophy of Science*, 23, 143–160.

Kuorikoski, J., Lehtinen, A., and Marchionni, C. (2010), "Economic Modelling as Robustness Analysis," *The British Journal for the Philosophy of Science*, 61, 541 –567.

Lakatos, I. (1978), *The Methodology of Scientific Research Programmes, Philosophical Papers*, vol. I, Cambridge, UK: Cambridge University Press.

Lalive, R., Van Ours, J., and Zweimüller, J. (2006), "How Changes in Financial Incentives Affect the Duration of Unemployment," *Review of Economic Studies*, 73, 1009–1038.

Lange, M. (1993), "Natural Laws and the Problem of Provisos," *Erkenntnis*, 38, 233–248.

— (2000), *Natural Laws in Scientific Practice*, New York: Oxford University Press.

Leamer, E. E. (1983), “Let’s Take the Con Out of Econometrics,” *American Economic Review*, 73, 31–43.

— (2010), “Tantalus on the Road to Asymptopia,” *Journal of Economic Perspectives*, 24, 31–46.

Lechthaler, W., Merkl, C., and Snower, D. J. (2010), “Monetary Persistence and the Labor Market: A New Perspective,” *Journal of Economic Dynamics and Control*, 34, 968–983.

Leslie, S. (2007), “Generics and the Structure of the Mind,” *Philosophical Perspectives*, 21, 375–403.

— (2008), “Generics: Cognition and Acquisition,” *Philosophical Review*, 117, 1–47.

Lester, R. A. (1964), “Part III What Can We Learn from European Experience? Discussion,” in *Unemployment and the American Economy*, ed. Ross, A. M., New York: John Wiley & Sons, pp. 195–198.

Levins, R. (1966), “The Strategy of Model Building in Population Biology,” *American Scientist*, 54, 421–431.

Lewis, D. (1973), “Causation,” *Journal of Philosophy*, 70, 556–567.

— (1979), “Counterfactual Dependence and Time’s Arrow,” *Noûs*, 13, 455–76.

Liebesman, D. (2011), “Simple Generics,” *Noûs*, 45, 409–442.

Longino, H. (1990), *Science as Social Knowledge*, Princeton: Princeton University Press.

Longworth, F. (2010), “Cartwright’s Causal Pluralism: a Critique and an Alternative,” *Analysis*, 70, 310 –318.

Machamer, P. (2004), “Activities and Causation: The Metaphysics and Epistemology of Mechanisms,” *International Studies in the Philosophy of Science*, 18, 27–39.

Machamer, P., Darden, L., and Craver, C. F. (2000), “Thinking about Mechanisms,” *Philosophy of Science*, 67, 1–25.

Mackie, J. (1974), *The Cement of the Universe: A Study of Causation*, Oxford: Clarendon.

Marshall, A. (1920), *Principles of Economics: An Introductory Volume*, London: Macmillan.

Martin, J. P. (1996), "Measures of Replacement Rates for the Purpose of International Comparisons: A Note," *OECD Economic Studies*, 99–115.

Meyer, B. D. (1995), "Lessons from the U.S. Unemployment Insurance Experiments," *Journal of Economic Literature*, 33, 91–131.

Mill, J. S. (1886), *A System of Logic: Ratiocinative and Inductive*, London: Longmans, Green, and Co.

Mireles-Flores, L. (2013), "Economics for Use," PhD thesis, Erasmus University Rotterdam, Rotterdam (Netherlands).

Mitchell, S. D. (2009), *Unsimple Truths: Science, Complexity, and Policy*, Chicago: University of Chicago Press.

Morgan, M. S. (1990), *The History of Econometric Ideas*, Cambridge: Cambridge University Press.

Morgan, M. S. and Morrison, M. (eds.) (1999), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge, UK: Cambridge University Press.

Morgan, S. L. and Winship, C. (2007), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge, UK: Cambridge University Press.

Myers, R. J. (1964), "Unemployment in Western Europe and the United States," in *Unemployment and the American Economy*, ed. Ross, A. M., New York: John Wiley & Sons, pp. 172–186.

Myrvold, W. C. (1996), "Bayesianism and Diverse Evidence: A Reply to Andrew Wayne," *Philosophy of Science*, 63, 661–665.

Möller, J. (2010), "The German labor market response in the world recession – de-mystifying a miracle," *Zeitschrift für ArbeitsmarktForschung*, 42, 325–336.

Nevo, A. and Whinston, M. D. (2010), "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," *Journal of Economic Perspectives*, 24, 69–82.

Nickel, B. (2010), "Ceteris Paribus Laws: Genericity and Natural Kinds," *Philosophers' Imprint*, 10, 1–25.

Nickell, S. (1997), "Unemployment and Labor Market Rigidities: Europe versus North America," *Journal of Economic Perspectives*, 11, 55–74.

Nickell, S., Nunziata, L., and Ochel, W. (2005), "Unemployment in the OECD since the 1960s. What Do We Know?" *Economic Journal*, 115, 1–27.

Norton, J. D. (2003), "A Material Theory of Induction," *Philosophy of Science*, 70, 647–670.

Novack, G. (2007), "Does Evidential Variety Depend on How the Evidence Is Described?" *Philosophy of Science*, 74, 701–711.

Nye, M. J. (1972), *Molecular Reality: A Perspective on the Scientific Work of Jean Perrin*, London: Macdonald.

OECD (1992), "Final Communiqué – Ministerial Meeting 18-19 May," Tech. Rep. SG/PRESS(92)43, Organisation for Economic Co-operation and Development, Paris.

— (1994a), *OECD Jobs Study: Evidence and Explanations*, Paris: Organisation for Economic Co-operation and Development.

— (1994b), *OECD Jobs Study: Facts, Analysis, Strategies*, Paris: Organisation for Economic Co-operation and Development.

— (1995), *The OECD Jobs Study: Implementing the Strategy*, Paris: Organisation for Economic Co-operation and Development.

— (1996a), *OECD Economic Surveys: Italy 1996*, Paris: Organisation for Economic Co-operation and Development.

— (1996b), *OECD Economic Surveys: Netherlands 1996*, Paris: Organisation for Economic Co-operation and Development.

— (1996c), *OECD Economic Surveys: United States 1996*, Paris: Organisation for Economic Co-operation and Development.

— (1998), "The OECD Jobs Strategy: Progress Report on Implementation of Country-Specific Recommendations," *OECD Economics Department Working Papers*.

— (1999), *Implementing the OECD Jobs Strategy: Assessing Performance and Policy*, Paris: Organisation for Economic Co-operation and Development.

— (2006a), *Boosting Jobs and Incomes - Policy Lessons from Reassessing the OECD Jobs Strategy*, Paris: Organisation for Economic Co-operation and Development.

— (2006b), *OECD Employment Outlook: Boosting Jobs and Incomes*, Paris: Organisation for Economic Co-operation and Development.

— (2010), *OECD Employment Outlook: Moving Beyond the Jobs Crisis*, Paris: Organisation for Economic Co-operation and Development.

Okun, A. M. (1962), "Potential GNP: Its Measurement and Significance," *Proceedings of the Business and Economic Statistics Section of the American Statistical Association.*

Oreskes, N. (2007), "The Scientific Consensus on Climate Change: How Do We Know We're Not Wrong?" in *Climate Change*, eds. DiMento, J. and Doughman, P., Cambridge, MA: MIT Press, pp. 65–99.

Orzack, S. H. and Sober, E. (1993), "A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966)," *The Quarterly Review of Biology*, 68, 533–546.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco: Morgan Kaufmann.

— (2009), *Causality: Models, Reasoning and Inference, Second Edition*, Cambridge, UK: Cambridge University Press.

Peregrin, J. (2012), "What Is Inferentialism?" in *Inference, Consequence, and Meaning: Perspectives on Inferentialism*, ed. Gurova, L., Cambridge, UK: Cambridge Scholars Publishing, pp. 3–16.

Perrin, J. (1913), *Les atomes*, Paris: Félix Alcan.

Pissarides, C. A. (2000), *Equilibrium Unemployment Theory, Second edition*, Cambridge, MA: MIT Press.

Popper, K. ([1959] 1992), *The Logic of Scientific Discovery*, London: Routledge.

— ([1963] 2002), *Conjectures and Refutations: The Growth of Scientific Knowledge*, London: Routledge.

Reiss, J. (2007), *Error in Economics: Towards a More Evidence-Based Methodology*, London: Routledge.

— (2009), “Causation in the Social Sciences: Evidence, Inference, and Purpose,” *Philosophy of the Social Sciences*, 39, 20–40.

— (2011a), “Empirical Evidence: Its Nature and Sources,” in *The SAGE Handbook of the Philosophy of Social Sciences*, eds. Jarvie, I. C. and Zamora-Bonilla, J., London: SAGE, pp. 551–576.

— (2011b), “Third Time’s a Charm: Causation, Science and Wittgensteinian Pluralism,” in *Causality in the Sciences*, eds. Illari, P. M., Russo, F., and Williamson, J., Oxford: Oxford University Press, pp. 907–927.

— (2012), “Causation in the Sciences: An Inferentialist Account,” *Studies in History and Philosophy of Biological and Biomedical Sciences*, forthcoming.

Reutlinger, A., Schurz, G., and Hüttemann, A. (2011), “Ceteris Paribus Laws,” in *The Stanford Encyclopedia of Philosophy*, ed. Zalta, E. N., spring 2011 ed.

Rinne, U. and Zimmermann, K. F. (2011), “Another Economic Miracle? The German Labor Market and the Great Recession,” *IZA Discussion Paper*.

Ross, A. M. (1964), “Conclusions,” in *Unemployment and the American Economy*, ed. Ross, A. M., New York: John Wiley & Sons, pp. 199–211.

Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.

— (1980), “Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment,” *Journal of the American Statistical Association*, 75, 591–593.

— (1986), “Statistics and Causal Inference. Comment: Which Ifs Have Causal Answers,” *Journal of the American Statistical Association*, 81, 961–962.

— (1990), “Formal Mode of Statistical Inference for Causal Effects,” *Journal of Statistical Planning and Inference*, 25, 279–292.

Russo, F. and Williamson, J. (2007), “Interpreting Causality in the Health Sciences,” *International Studies in the Philosophy of Science*, 21, 157–170.

— (2010), “Epistemic Causality and Evidence-Based Medicine,” Working Paper, Kent University, http://philsci-archive.pitt.edu/8351/.

— (2011), “Generic Versus Single-Case Causality: The Case of Autopsy,” *European Journal for Philosophy of Science*, 1, 47–69.

Salmon, W. C. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

— (1998), *Causality and Explanation*, New York: Oxford University Press.

Schaffer, J. (2005), “Contrastive Causation,” *The Philosophical Review*, 114, 327–358.

— (2008), “The Metaphysics of Causation,” in *The Stanford Encyclopedia of Philosophy*, ed. Zalta, E. N., fall 2008 ed.

Schneider, F. and Enste, D. H. (2000), “Shadow Economies: Size, Causes, and Consequences,” *Journal of Economic Literature*, 38, 77–114.

Scriven, M. (1959), “Truisms as the Grounds for Historical Explanations,” in *Theories of history*, ed. Gardiner, P., New York: The Free Press, pp. 443–475.

Sellars, W. (1953), “Inference and Meaning,” *Mind*, LXII, 313–338.

— (1954), “Some Reflections on Language Games,” *Philosophy of Science*, 21, 204–228.

Shogenji, T. (2005), “Justification by Coherence from Scratch,” *Philosophical Studies*, 125, 305–325.

Simon, H. A. (1957), “Causal Order and Identifiability,” in *Models of Man*, New York: Wiley.

Sims, C. A. (1980), “Macroeconomics and Reality,” *Econometrica*, 48, 1–48.

— (2010), “But Economics Is Not an Experimental Science,” *Journal of Economic Perspectives*, 24, 59–68.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, London: Methuen.

Sober, E. (1989), “Independent Evidence About a Common Cause,” *Philosophy of Science*, 56, 275–287.

Solow, R. M. (1962), “Technical Progress, Capital Formation, and Economic Growth,” *The American Economic Review*, 52, 76–86.

Speaks, J. (2011), “Theories of Meaning,” in *The Stanford Encyclopedia of Philosophy*, ed. Zalta, E. N., summer 2011 ed.

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search, Second Edition*, Cambridge, MA: MIT Press.

Spohn, W. (2006), “Causation: An Alternative,” *The British Journal for the Philosophy of Science*, 57, 93 –119.

Staley, K. W. (2004), “Robust Evidence and Secure Evidence Claims,” *Philosophy of Science*, 71, 467–488.

Steel, D. (2008), *Across the Boundaries: Extrapolation in Biology and Social Science*, New York: Oxford University Press.

— (2011), “Causality, Causal Models, and Social Mechanisms,” in *The SAGE Handbook of Philosophy of Social Science*, eds. Jarvie, I. C. and Zamora-Bonilla, J., SAGE.

Stegenga, J. (2009), “Robustness, Discordance, and Relevance,” *Philosophy of Science*, 76, 650–661.

— (2011), “Is Meta-Analysis the Platinum Standard of Evidence?” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42, 497–507.

Stock, J. H. (2010), “The Other Transformation in Econometric Practice: Robust Tools for Inference,” *Journal of Economic Perspectives*, 24, 83–94.

Sugden, R. (2000), “Credible Worlds: The Status of Theoretical Models in Economics,” *Journal of Economic Methodology*, 7, 1–31.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, Amsterdam: North-Holland.

Teller, P. (2001), "Twilight of the Perfect Model Model," *Erkenntnis*, 55, 393–415.

Thurmond, V. A. (2001), "The Point of Triangulation," *Journal of Nursing Scholarship*, 33, 253–258.

Tobin, G. A. and Begley, C. M. (2004), "Methodological Rigour Within a Qualitative Framework," *Journal of Advanced Nursing*, 48, 388–396.

Tobin, J. (1987), "Okun, Arthur M. (1928–1980)," in *The New Palgrave Dictionary of Economics*, eds. Durlauf, S. N. and Blume, L. E., Basingstoke: Nature Publishing Group, 2nd ed.

Tyrowicz, J. and Cichocki, S. (2011), "Employed unemployed? On shadow employment in transition," *Empirica*, 38, 259–281.

van Fraassen, B. C. (1980), *The Scientific Image*, Oxford: Clarendon Press.

Vanoli, A. (2008), "national accounting, history of," in *The New Palgrave Dictionary of Economics*, eds. Durlauf, S. N. and Blume, L. E., Basingstoke: Nature Publishing Group, 2nd ed., pp. 838–843.

Wayne, A. (1995), "Bayesianism and Diverse Evidence," *Philosophy of Science*, 62, 111–121.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (1966), *Unobtrusive Measures: Nonreactive Research in the Social Sciences*, Chicago: Rand McNally.

Weber, E. (2009), "How Probabilistic Causation Can Account for the Use of Mechanistic Evidence," *International Studies in the Philosophy of Science*, 23, 277–295.

Weber, E. and Leuridan, B. (2008), "Counterfactual Causality, Empirical Research, and the Role of Theory in the Social Sciences," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 41, 197–201.

Weber, M. (2005), *Philosophy of Experimental Biology*, Cambridge, UK: Cambridge University Press.

Weingast, B. R. and Wittman, D. (2008), *The Oxford Handbook of Political Economy*, Oxford: Oxford University Press.

Weisberg, M. (2006), "Robustness Analysis," *Philosophy of Science*, 73, 730–742.

Wheeler, G. (2009), "Focused Correlation and Confirmation," *The British Journal for the Philosophy of Science*, 60, 79 –100.

Wheeler, G. and Scheines, R. (2011), "Causation, Association and Confirmation," in *Explanation, Prediction, and Confirmation*, eds. Dieks, D., Gonzalez, W. J., Hartmann, S., Uebel, T., and Weber, M., Dordrecht: Springer Netherlands, pp. 37–51.

Whiting, D. (2009), "Conceptual Role Semantics," *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/conc-rol/.

Williamson, J. (2005), "Epistemic Causality," in *Bayesian Nets and Causality: Philosophical and Computational Foundations*, Oxford: Oxford University Press, pp. 130–151.

Wilson, E. O. (1998), *Consilience: The Unity of Knowledge*, New York: Vintage Books.

Wimsatt, W. C. (2007a), *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Cambridge, Mass.: Harvard University Press.

— (2007b), "Robustness, Reliability, and Overdetermination," in *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Cambridge, MA: Harvard University Press, pp. 43–74.

Wittgenstein, L. ([1953] 2001), *Philosophical Investigations*, Oxford: Blackwell, 3rd ed.

— (1958), *Preliminary studies for the "Philosophical investigations": generally known as the Blue and Brown books*, Oxford: Blackwell Publishing.

Woodford, M. (2009), "Convergence in Macroeconomics: Elements of the New Synthesis," *American Economic Journal: Macroeconomics*, 1, 267–279.

Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

— (2006), “Some Varieties of Robustness,” *Journal of Economic Methodology*, 13, 219–240.

— (2010), “Causation in biology: stability, specificity, and the choice of levels of explanation,” *Biology & Philosophy*, 25, 287–318.

— (2011), “Mechanisms Revisited,” *Synthese*, 183, 409–427.

Zimmermann, K. F. (2009), “Germany’s New Labor Market Miracle,” Slides of a presentation in Stockholm (Sweden), http://www.slideshare.net/tankesmedjanfores/klaus-zimmermann (last accessed June 28, 2012).