RESEARCH METHODS SERIES

Case Studies and Causal Inference

An Integrative Framework

Ingo Rohlfing

ecpr

Case Studies and Causal Inference

Research Methods Series

General Editors: **Bernhard Kittel**, Professor of Social Science Methodology, Department of Social Sciences, **Carl von Ossietzky**, Universität Oldenburg, Germany and **Benoît Rihoux,** Professor of Political Science, Université catholique de Louvain (UCL), Belgium.

In association with the European Consortium for Political Research (ECPR), Palgrave Macmillan is delighted to announce the launch of a new book series dedicated to producing cutting-edge titles in Research Methods. While political science currently tends to import methods developed in neighbouring disciplines, the series contributes to developing a methodological apparatus focusing on those methods which are appropriate in dealing with the specific research problems of the discipline.

The series provides students and scholars with state-of-the-art scholarship on methodology, methods and techniques. It comprises innovative and intellectually rigorous monographs and edited collections which bridge schools of thought and cross the boundaries of conventional approaches. The series covers both empirical-analytical and interpretive approaches, micro and macro studies, and quantitative and qualitative methods.

Titles include:

Joachim Blatter and Markus Haverland
DESIGNING CASE STUDIES
Explanatory Approaches in Small-N Research

Alexander Bogner, Beate Littig and Wolfgang Menz (*editors*)
INTERVIEWING EXPERTS

Bernhard Kittel, Wolfgang J. Luhan and Rebecca B. Morton (*editors*)
EXPERIMENTAL POLITICAL SCIENCE
Principles and Practices

Audie Klotz and Deepa Prakash (*editors*)
QUALITATIVE METHODS IN INTERNATIONAL RELATIONS
A Pluralist Guide

Lane Kenworthy and Alexander Hicks (*editors*)
METHOD AND SUBSTANCE IN MACROCOMPARATIVE ANALYSIS

Ingo Rohlfing
CASE STUDIES AND CAUSAL INFERENCE
An Integrative Framework

Research Methods Series
Series Standing Order ISBN 978–0230–20679–3–hardcover
ISBN 978–0230–20680–9–paperback
(outside North America only)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and one of the ISBNs quoted above.

Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England

Case Studies and Causal Inference

An Integrative Framework

Ingo Rohlfing
*Assistant Professor of Comparative Social Research,
University of Cologne, Germany*

palgrave
macmillan

Softcover reprint of the hardcover 1st edition 2012 978-0-230-24070-4

First published 2012 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries

ISBN 978-1-349-31657-1 ISBN 978-1-137-27132-7 (eBook)
DOI 10.1057/9781137271327

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

10 9 8 7 6 5 4 3 2 1
21 20 19 18 17 16 15 14 13 12

For Stephanie and Adrian

This page intentionally left blank

Contents

Figures

Tables

Preface

'Do not select cases that do not vary on the outcome; you cannot learn anything from them.' This is one of the main messages that I took from my research design course a decade ago, a course that heavily built on what is arguably the most influential book on social science methods, King, Keohane, and Verba's (1994) *Designing Social Inquiry*. At that point in time, as I had completed some statistics education, this advice struck me as quite reasonable. How would I do a correlational analysis if the cases do not vary on the dependent variable in the first place? In retrospect, I recognize that I was fully subscribing to a 'statistical worldview' (McKeown 1999) on the social and political sphere and on social science methods. However, as is the case with many, I was not aware that I had made this subscription, as we did not discuss any of the numerous possible responses and replies to *Designing Social Inquiry* in the research design course.

It was only at the beginning of my PhD studies that a methods workshop exposed me to the methods debate that had evolved following the publication of *Designing Social Inquiry*. In retrospect, these workshops also constituted the beginning of my own work on various aspects of social science methods and methodology. Appearing after a considerable delay, the present book is one major outcome of my engagement with the methods literature and will, I hope, also make some contribution to it. (For those who might be thinking, 'Please, not another book doing some infighting with *Designing Social Inquiry*', I can assure the reader that this is not the style of my book. Although I periodically draw on *Designing Social Inquiry*, this book represents a single, though certainly valuable, contribution to the development of the social science methods to which I refer in my book.)

Many decisions about the structure of a book, its key messages, and so on, have to be made in the course of the writing process. In this case, one important decision concerned the question of how to handle empirical examples. Some people understand arguments on methods better on an abstract level and by referring to Y, X1, X2, and so on, while others prefer specific empirical examples. I decided to use empirical examples whenever possible in order to demonstrate the relevance of my arguments for empirical research. To the extent that it is possible, I draw on published studies from a wide variety of fields of research. Several notes are in order in relation to the cited empirical research (in particular, the researchers that are cited). First, some of the studies are not case studies. For instance, Drezner's (2000) study on the enforcement of multilateral economic sanctions relies on a regression analysis. Nevertheless, I draw on Drezner's piece because it is a good example of a study that modifies an existing hypothesis that is flawed in some respect. Second, I distinguish between research that aims to build, test, and modify a hypothesis (see Chapter 1). In some instances, I discuss an empirical study in a different theoretical context so as to clarify an argument. Again, Drezner's study can be taken to illustrate this point. Drezner notes that an existing explanation of the effectiveness of multilateral sanctions is flawed because the explanation is only sometimes confirmed. He then proposes a modified hypothesis and tests it quantitatively. In terms of the framework that I introduce in Chapter 1, this is a hypothesis-testing study. For purposes of illustration, I present it as an example of a hypothesis-modifying case study, that is, an inductive study at the end of which a proposition is modified in light of the collected empirical evidence.

Finally, I sometimes needed to fictionalize cases or causes for purposes of illustration. In the chapter on cross-case comparisons, I repeatedly rely on Eckert's (2010) case study of delegation in the postal sectors of the United Kingdom, Germany, and France. Since I could not show all the arguments that I wanted to make with these three cases, I modified the study in various ways. For example, one modification involved assuming that the competitiveness of the postal companies in Belgium and Italy was high, which may or may not have been the case a decade ago.

Of course, this book would not have been possible without the support of many people. The biggest thanks go to my wife Stephanie, who has been an invaluable support and a source of encouragement throughout my entire career. My little son Adrian also must be mentioned, as he did everything, in the best and worst ways, to keep me from writing this book. On the academic side, I would like to thank Bernhard Kittel for arousing my interest in social science methods and, together with Benoît Rihoux, for giving me the opportunity to write this book. Hans-Jürgen Andreß, Philipp Genschel, and André Kaiser deserve some credit for having given me the opportunity to pursue my methodological interests.

Many friends and fellows made very helpful comments on various parts of the manuscript. I extend my thanks to Derek Beach, Joachim Blatter, Payam Ghalehdar, Lukas Haffert, Peter Hall, Annika Hennl, Michael Kaeding, Andreas Kammer, David Kühn, Dirk Leuffen, James Mahoney, Alexander Reutlinger, Saskia Ruth, Jan Sauermann, André Schaffrin, Carsten Q. Schneider, Sebastian Sewerin, Peter Starke, Paul Thurner, Andreas Warntjen, Tobias Wickern, Gregor Zons, and Christina Zuber. My thanks go to the numerous participants of the following courses that helped me to shape my thoughts and constantly revise my perspective on small-n methods: the case study courses at the University of Cologne in the winter terms of 2009, 2010, and 2011 and the case study courses at ECPR Summer School in Methods and Techniques (Ljubljana, 2007–2011). In relation to the summer school courses, special thanks go to my teaching assistants Payam Ghalehdar, David Kühn, Natasja Reslow, and Christina Zuber, as they kept dragging

me into discussions that allowed me see some things more clearly. Finally, I want to show my appreciation for the special role that Andreas Kammer, David Kühn, and Peter Starke played during the writing process. Whenever I was struck with a particular issue and could not get my thoughts straight, I could rely on one of the three for a second opinion and to serve as a target at which I could shoot thoughts from my hip.

On the administrative side, Liz Holwell from Palgrave deserves credit for being patient with me and for organizing the publication process very efficiently. Ross McCalden and Nancy Deyo provided invaluable language assistance. Finally, I applaud my student assistant, Tobias Schafföner, who was in charge of editing my manuscript. As can easily be imagined, the editing process has many intricacies, both anticipated and unforeseen; Tobi managed to master the intricacies successfully and complete the process. Needless to say, I am to blame for any errors that remain or misleading arguments.

1 Introduction

There are many different ways to define the social sciences, their aims, and the manner in which social and political research should be carried out in order to achieve these aims. Some argue that it is possible to discover regular patterns in the social and political world by crafting a suitable research design and using appropriate methods (King et al. 1994; Mahoney and Rueschemeyer 2003a; Pollins 2007). Others contend that there are regularities but that they are impossible to detect because the social world is too complex (Bhaskar 1975; Kurki 2007; Lane 1996). Going one step farther, some argue that the wwhole concept of thinking in terms of regular cause-effect relationships is misleading because there are no regularities whatsoever. According to this reading, all that the social sciences can and should do is reconstruct the intentions and meaning of actor behavior (Ferejohn 2004; Howe 2011; Jackson 2010, chap. 1).

This is, of course, not meant to be more than a sketch of three different philosophies of social science, usually broadly referred to as neopositivism, critical realism, and constructivism, which certainly do not exhaust the range of existing philosophies. Whichever philosophy one ascribes to, it is important to reflect on it because of the implications for how empirical research should be carried out (Friedrichs and Kratochwil 2009; Hay 2006; Jackson 2010; Johnson 2006; Pollins 2007).

The intention of this book is to provide a comprehensive elaboration of the case study method interested in inferences about *regular* causal relationships.[1] This means that case studies are based on the ontological premise that at least some empirical relationships are regular, that is invariant (regular without exception) or at least systematic (regular with exceptions), and that one can learn something about these relationships via systematic small-n research.[2] An additional implication is that case studies are understood to be *theory centered* as opposed to *case centered*.[3] A case study is theory centered when it contributes to the advancement of general theory; such case studies can take any of three different forms (see Section 1.3). (1) An exploratory case study can generate conclusions such as ‘In regions with high export rates, there are strong organized business interests pressing regional governments to promote their exports abroad' (see Blatter et al. 2010). (2) It can test hypotheses such as 'Regulatory agencies established by governments that face political uncertainty are more insulated from political oversight than those established under stable regimes' (see Yesilkagit and Christensen 2010). (3) It can refine existing hypotheses in order to resolve a puzzle such as 'Why have gender quotas been adopted by parliaments in countries where women have a low status?' (see Bush 2011). These examples show that specific cases – for example, the adoption of a gender quota by a national parliament in a specific country at a specific point in time – are instrumental for producing general theoretical statements extending beyond the cases that one examined empirically.[4]

In case-centered case studies, on the other hand, theory is instrumental for the formulation of a comprehensive explanation of a single case. Two related features of case-centered case studies are, first, that insights derived from the case study are not taken for the advancement of general theory and, second, that the explanation formulated for the case at hand is not generalized to other cases.

This understanding of theory-centered and case-centered case studies shows that their underlying goals are compatible. The formulation of general inferences on the basis of qualitative case studies does not preclude one from also gaining a comprehensive understanding of the examined cases (and the other way round). However, it should be clearly understood that the focus of this book is on theory-centered case studies aiming at general statements such as the three illustrative ones mentioned above. The emphasis on the analysis of regular relationships is also reflected in and accounts for my definition of 'case study' (elaborated in Chapter 2), which is conceptualized as *the empirical analysis of a small sample of bounded empirical phenomena that are instances of a population of similar phenomena*.

So far, the discussion has referred to general causal relationships as the objects of interest. The broad understanding of a regular causal relationship has two components that play a vital role in the social sciences and in this book, as well. The notion of a regular cause-effect relationship entails that a certain cause (X) has a *causal effect* on an outcome (Y) and is connected to it via *causal processes* and one or more *causal mechanisms* in a specified population of cases (Kühn and Rohlfing 2010). There has been ongoing debate in the social sciences about what is required for valid causal inference (Beck 2010; Beck 2006; Brady et al. 2006; Collier et al. 2010; Gerring 2010); that is, the generation of correct claims that a certain factor is indeed a cause of a given outcome.

In principle, a well-crafted design allows it to generate causal inferences by finding evidence for the systematic effect of a cause (Woodward 2011). This is particularly pertinent to experiments that allow for the randomization of the treatment to units (Morton and Williams 2010, chap. 2), as randomization allows one to eliminate the effects of potential confounders. As is well known, however, causal inference is much more protracted in observational research because of the inability to engage in randomization (Przeworski 2007). Moreover, causal inference via the analysis of causal effects falls short of offering a causal explanation of *why* an outcome occurs, that is, it remains unclear as to what the causal process and mechanism are that underlie the effect and tie the cause to the outcome (Dessler 1991).[5] Because of the problems of causal inference from the analysis of causal effects alone and the additional merit of explaining phenomena, there is an emerging consensus that causal analysis best combines inquiries into causal effects and causal mechanisms alike (Cartwright 2004; Hedström and Ylikoski 2010; Johnson 2006; Lipton 1991, 32; Waldner 2007, 145–6).[6] The joint focus on effects and mechanisms has been also referred to as *integrative theory* (Dessler 1991) as it blends the two facets of causal relationships into one theoretical account.

This twofold perspective on causal relationships and inference permits it to concretize the major goal of this book, which is to evaluate the potential of qualitative case studies for the advancement of general theory by the generation of inferences about causal effects and causal mechanisms. In fact, the case study is the only method that allows one to approach an empirical phenomenon from two sides simultaneously.[7] Of course, the opportunity to examine causal effects and mechanisms does not speak to the degree to which small-n research is able to generate valid integrative theories. In order to shed light on this issue, the following chapters include a systematic discussion of all salient stages of the research process wherein case studies create specific challenges. These stages concern case selection, comparisons, process tracing, the actual generation of inferences on causal effects and mechanisms, and the generalization of these inferences to other cases in the population.[8] In order to lay the groundwork for these chapters, the following sections develop an integrative framework that serves as a template for the discussion of the case study method.

1.1 An integrative framework for case studies

The discussion of the case study method settled in a regularities perspective does not mean that all ontological and epistemological issues are settled. The ontological premise that empirical relationships are regular still allows for diverse views on a multitude of salient aspects (Jackson 2010, chap. 3). Among researchers who share a belief in regular social phenomena, there is a long-standing and vibrant discussion about causal inference along the dimensions of qualitative versus quantitative research and large-n versus small-n research (Mahoney 2010; Prakash and Klotz 2007). Given this discussion and the increasingly rich body of literature on small-n methods, one may wonder whether there is still much left to say about case studies. Considering the last decades (Collier 1993) and the last 15 years in particular (Mahoney 2010), the case study method made a huge leap forward.

Just to mention a few aspects, there is now a much greater awareness of the distinct features of research concerning necessary conditions (Braumoeller and Goertz 2000; Goertz and Starr 2003b), temporality (Büthe 2002; Grzymala-Busse 2011), the combination of set relations and temporality (Mahoney et al. 2009), case selection (Collier and Mahoney 1996; Seawright and Gerring 2008), and the use of sources (Lustick 1996; Thies 2002).

Certainly, the treatment of these and other topics contributed to the development of the case study method. At the same time, they led to a remarkable expansion of the field and to the formulation of seemingly contradictory arguments that may leave empirical small-n researchers a little bit puzzled. For instance, research advocating process tracing and causal mechanisms is a growth industry (for example, Checkel 2008; Falleti and Lynch 2008; Falleti and Lynch 2009; George and Bennett 2005, chaps 7 and 10; McAdam et al. 2008; Tilly 2004), yet we also find the opposing claim that mechanisms are not needed for causal inference (Gerring 2010). Similarly, we find pleas for the adoption of an experimental (i.e., correlational) perspective (Gerring 2007a, chap. 6) and a set-relational perspective on causation (George and Bennett 2005, chap. 11), two perspectives that are very difficult if not impossible to reconcile (see Section 2.4 and Ragin 1981; Ragin and Zaret 1983).

In light of this heterogeneity in the field, the goal of this book is the formulation of an integrative framework that allows for a self-contained and balanced discussion of causal inference via case studies. Four dimensions discussed in more detail in the remainder of this chapter constitute that framework.[9] First, the *research purpose*, or *goal*, of a case study can be the formation of an entirely new hypothesis, a plain test of a hypothesis, or the refinement of a hypothesis. Second, the *level of analysis* can be the cross-case level and/or the within-case level. A case study interested in the within-case level is concerned with causal mechanisms and processes, while a cross-case study aims to infer whether a given factor has a causal effect and, if it does, of what sort it is. The third dimension is specifically related to the cross-case level and covers the *nature of the causal effect*. The question that is central to this dimension is whether the effect of a cause on an outcome is assumed to be correlational or set relational.[10] Fourth and specifically related to case studies that test hypotheses, one can rely on both *frequentist* and *Bayesian* modes of causal inference. Frequentism means that causal inferences are based only on the number of collected observations that are and are not in line with a given hypothesis. In contrast, Bayesianism asks for the likelihood of collecting specific observations prior to the empirical analysis. Confidence in a hypothesis is then updated, depending on how possible it was to gather the observations that one actually collected in the empirical analysis.

The application of these four dimensions to different stages of the research process distinguishes the present book from discussions that tend to segment the field and focus on the practice of small-n research. For instance, one finds the distinction between a covariational approach, a causal process tracing approach, and a congruence approach in small-n research (Blatter and Blume 2010). In light of the framework underlying the present book, this threefold distinction artificially separates elements of empirical research that can go together. For instance, covariational case studies as one form of cross-case analysis are separated from process tracing. In contrast, the comprehensive approach that underlies the following chapters allows for the viable combination of both elements in a stand-alone case study, which is, in fact, common in empirical research (see, for example, Lieberman 2009; Ziblatt 2009). Moreover, the covariational approach contrasts with the assertion that most small-n research takes a set-relational view on causation (Goertz and Mahoney 2012; Mahoney and Goertz 2006). At the same time, however, set-relational reasoning might dominate small-n research (Mahoney and Terrie 2008), but this is only a matter of practice and should not conceal that the case study method permits it to generate correlational or set-relational inferences and that both cross-case inferences can coincide with process tracing and the analysis of temporality and causal mechanisms. The integrative perspective that is taken here aims to work against the segmentation of the field by offering a comprehensive discussion of the case study method and its potential for rigorous causal inference.

The four-dimensional framework underlying this book does not aim to synthesize the multitude of arguments on small-n research. In fact, one major insight is that there are noteworthy differences between the ways in which case studies can be implemented. A single best case study method and way to perform empirical small-n research do not exist; but there are various equally viable ways to complete case studies. The application of the four dimensions to all research stages shows that their interplay is crucial to shaping both the research design and the route case study researchers should take in their empirical analysis and in the generation of causal inferences. The integrative framework thus sorts the field of case study research and methods without segmenting it into different camps, paradigms, and so on.

Sorting the field without segmenting it yields the additional benefit of highlighting the differences and, more importantly, commonalities among the various ways of doing case study research. As is elaborated on in more detail below, a balanced discussion of the case study method shows that small-n research faces similar challenges regardless of where a given analysis is located in the four-dimensional framework.

In addition to this, the integrative approach yields benefits by clarifying the current state of the methodological debate and pointing to some inconsistencies and blind spots. For instance, one generally finds the distinction between case studies that test hypotheses and that build hypotheses via exploratory process tracing (Beach and Pedersen, 2012; George and Bennett 2005; Mahoney 2010). The latter category lumps together what I refer to as case studies that build new hypotheses from scratch and that address puzzling cases in order to refine existing propositions. The following chapters show that it is important to keep the two theoretical goals separate because they differ in important respects related to case selection and comparison, for example. Since puzzle solving and the analysis of hitherto unexplored phenomena are generally taken as the key domains of qualitative case studies (Eisenhardt 1989; Gerring 2004; Odell 2004), it is unfortunate that the implications of the two research goals for the case study method have not been sufficiently spelled out so far.

Moreover, what seems to be a contradiction in the debate (e.g., as regards the inferential value of single-case studies [Gerring 2004; Rueschemeyer 2003]) often proves to be a different perspective on the same issue – for example, the nature of the causal effect in terms of correlations and set relations.[11] One central premise of this book is that the method and the research design should follow theory and ontological premises about how the social and political world works (Hay 2006). Once one allows ontology and theory to play that role, it becomes evident that there are many ways to do case studies.

Finally and on a more general level, the application of the integrative framework to the case study method shows that one should beware of formulating and following seemingly hard-and-fast rules. The debate about the case study method is rife with misconceptions derived from the transfer of principles from a context in which they make sense to another where they are misplaced. Arguably, the most famous example is the admonition of King, Keohane, and Verba (1994, 147–9) not to select cases that lack variance on the outcome. This case selection strategy must be avoided if one aims to test a hypothesis that postulates a correlation between cause and outcome, which is King et al.'s key concern. In response to this claim, though, it has been repeatedly and correctly argued that no variance on the outcome can have inferential merit (Collier and Mahoney 1996) and that it is particularly warranted if one is interested in necessary conditions (Braumoeller and Goertz 2000; Dion 1998). However, the widely upheld principle that research on necessary conditions requires invariance on the outcome falls short of being a universal principle, too. As is shown in detail in Chapter 5, small-n research that aims to solve the puzzle of why the outcome occurs in the absence of a necessary condition must do more than achieve invariance on the phenomenon of interest. A systematic elaboration of the interplay between the four dimensions guides empirical researchers through their analyzes and avoids the production of invalid causal inference by following fallacious hard and fast rules.

1.2 Two problems for causal inference and two solutions

On one hand, the framework that underlies this book is salient for the proper realization of empirical case studies. On the other, an examination of the case study method along the four dimensions shows that it faces two uniform problems regardless of the location of a small-n analysis in this framework. First, an indeterminacy problem plagues the formulation of inferences that exhibit *internal validity*. Indeterminacy means that multiple inferences can make equal sense of the same empirical phenomenon (Collier et al. 2004c, 47). For instance, the hypothesis that a democratic dyad goes along with peace is indeterminate as regards the underlying determinants of this outcome because multiple mechanisms such as norms and domestic institutional constraints can account for this cross-case regularity (Rosato 2003). Second, *external validity* is questionable because of the generalization of causal inferences from a small number of examined cases to, probably, many cases that one did not study.

These two conclusions may come as no surprise because similar criticisms have been raised before vis-à-vis the case study method (see, for example, Lieberson 1991; Steinmetz 2004; Zelditch 1971). However, it should be noted that there are two salient differences between the arguments made in the literature and those made in this book. First, some of the earlier critiques were based on misleading or incomplete characterizations of the case study method, partly owing to the transfer of the template of quantitative research to the realm of qualitative case studies (Collier et al. 2004a). In contrast, my conclusions are derived from a small set of premises that are, in my reading, widely shared in the case study literature to which I refer. The fundamental assumption is that there are regularities that can be detected in empirical research, a premise that is shared with quantitative research (Jackson 2010, chap. 3). Unless one is able to analyze all cases in the population, this assumption accounts for the generalization problem because one generates inferences for more cases that one could examine empirically.

The diagnosis of an indeterminacy problem is equally based on two widely held premises: different causes can produce the same empirical phenomenon, which is known as equifinality, and multiple causes can interact in generating the outcome, also referred to as conjunctural and configurational causality. Without both assumptions, one makes the strong and usually hard-to-defend assumption of monocausation (Franzese 2008), meaning that there is only a single cause that can bring about the outcome.[12] There may be phenomena that can be explained in monocausal terms, but it is much more compelling to presume that equifinality and interaction effects are pervasive in the social world (Bennett and Elman 2006, 457; Checkel 2008, 126; Gerring 2007a, 61; Mahoney 2007b, 135). This premise, in my view, is reflected in the laudable development of social science theory. For instance, instead of arguing that only power or norms or agency or structure or a single institution matters for a given outcome, it is now widely acknowledged that structure *and* agency matter (Sil 2000), that actor behavior is shaped by material interests *and* norms *and* beliefs (Hall 2008), and that multiple institutions jointly structure the opportunities that are available to actors (Kaiser 1997; Lijphart 1999; Tsebelis 2002).[13] While this development of social science theory is appealing, as it moves away from the assumption of monocausation, it will become apparent in later chapters that this comes at the cost of manifest problems for causal inference in case studies (and empirical research more generally).

The second aspect in which this book differs from similar evaluations of the case study method is related to the potential of small-n research for causal inference. Notwithstanding the two problems that one confronts, I claim that there is neither need to disavow the case study method nor to declare case studies inherently deficient. Small-n studies can play an important role in the advancement of knowledge in the context of research interested in regularities. In particular, one can rely on two tools for strengthening internal and external validity. First, as regards internal validity, more elaborate theory can help to reduce the number of causes that one has to include in the analysis and the range of interaction effects that are logically possible (Dür 2007a), thereby diminishing indeterminacy. Second, the transformation of potential causes into scope conditions eases the indeterminacy problem. Scope conditions, also known as boundary conditions, delineate the domain within which a specific causal relationship is expected to exist (Walker and Cohen 1985). The transformation of causes into scope conditions reduces indeterminacy because the latter specify only the domain for which hypotheses and causal inferences are supposed to hold true. They are not part of the actual empirical analysis and are nor directly factored into the causal inferences. Furthermore, additional scope conditions increase the homogeneity of cases because they are members of the population only when meeting each of the specified scope conditions. This means that the more scope conditions one specifies, the less need there is for control of potential causes in the actual empirical analysis.

With regard to generalization, second, the transformation of causes into scope conditions is equally viable to promote external validity simply because the size of the population decreases as the number of scope conditions increases. This means that the ratio of examined to nonexamined cases becomes more favorable and claims of external validity more credible.

On the downside, one may counter that invoking more scope conditions means a diminished size of the population and that causal inferences cover fewer cases, and so, the empirical analysis becomes less interesting. In essence, one faces a trade-off here between one's confidenc in the internal and external validity of causal inferences, on the one hand, and the breadth of inferences in terms of the number of cases, on the other. Every case study researcher has to decide where to position the case study in this trade-off. From the perspective of causal inference, though, there are clear benefits to many scope conditions and few cases.

Furthermore, when one invokes scope conditions for the improvement of internal and external validity, it seems to be that study researchers are condemned to study small populations. However, this impression would be incorrect. As is explained in Chapter 9, the strategy of *layered generalization* allows one to gradually relax boundary conditions in a systematic manner and thereby adds one small layer of cases after another to the original population. At every round of the generalization stage, this strategy permits one to keep the external validity problem within acceptable limits and to increase, in a stepwise fashion, the number of cases to which causal inferences are extended.

The application of these instruments certainly does not eliminate all inferential problems that one confronts in small-n research, and some uncertainty always remains in producing causal inferences (King, Keohand, and Verba 1994, chap. 1). However, compared with case studies that rely on weak theory and aim for sweeping generalizations, one can diminish the challenges and substantially improve causal inference and generalization. The case study method therefore can be used for the building, testing, and modification of hypotheses on causal effects and causal mechanisms, and so, claims about the inferiority of small-n research or its abandonment are misplaced.

In the remainder of this chapter, I discuss each of the four dimensions that constitute the integrative framework in more detail so as to lay the foundations for the subsequent chapters. The introduction concludes with an outline of the book in which I present some of the key arguments in each chapter.

1.3 Three types of theory-centered research goals

In theory-centered research, one can distinguish designs that aim to build, test, or modify one or multiple hypotheses.[14] The three types of research aims can be distinguished according to two criteria. First, the case study is or is not built on an existing body of theory. Second, precise, testable hypotheses are formulated prior to or after the empirical analysis. The intersection of the two criteria creates the matrix presented in Table 1.1.

A case study builds a hypothesis from scratch if one does not draw on an elaborated body of theory and when one develops a hypothesis only after exploratory process tracing has been performed. Such case studies are particularly warranted for the analysis of new empirical developments

Table 1.1 Three theory-centered research goals

<table>
<tr><th></th><th></th><th colspan="2">Hypothesis is formed</th></tr>
<tr><th></th><th></th><th>After empirical analysis</th><th>Before empirical analysis</th></tr>
<tr><td rowspan="2">Existing theory on which one builds</td><td>No</td><td>Building hypothesis</td><td rowspan="2">Testing hypothesis</td></tr>
<tr><td>Yes</td><td>Modifying hypothesis</td></tr>
</table>

about which not much is known (Eisenhardt 1989, 532), for example, international negotiations about climate protection or the sources and effects of international criminal courts of justice. The social sciences offer theoretical frameworks such as rational choice and interpretivism that one can use for the formulation of testable hypotheses (Rueschemeyer 2009). However, the point is that no off-the-shelf hypotheses exist and one does not formulate any hypothesis in advance of the empirical analysis. One or multiple testable hypotheses come into existence only at the end of the case study, once empirical evidence has been collected and used to formulate propositions.

A hypothesis-testing case study puts to scrutiny a hypothesis that is specified in advance of the empirical analysis. Whether the prior state of theory is weak or strong is not relevant as long as one formulates a testable proposition prior to the collection and evaluation of data.[15] Although the use of case studies for the testing of hypotheses is disputed (King et al. 1994, 208–13; Steinmetz 2004), one finds countless empirical case studies that test propositions.

A case study is engaged in the modification of a hypothesis when it aims to improve a hypothesis that turned out to be deficient in previous research.[16] Quite often, researchers are motivated by the presence of a puzzle deserving a closer analysis (Alvesson and Karreman 2007; Grofman 2001); that is to say, one is modifying the hypothesis in the light of which a case represents a puzzle. Such a puzzle can be the observation of risky political reforms because such reforms create short-run costs and long-term benefits that political actors with short time horizons should eschew (Jacobs 2008; Vis 2009).

More precisely, modifying a hypothesis means adapting the original hypothesis by adding a cause so as to improve its performance (see Chapter 4). Drezner's (2000) analysis of effective economic sanctions is an example of this form of modification. Drezner notes that multilateral sanctions are sometimes effective and sometimes not, and this is the puzzle that he aims to solve. He does so by demonstrating that multilateral cooperation is effective only when it is supported by an international organization that enforces the sanctions. The original hypothesis 'Economic sanctions are effective when they are multilateral' is therefore transformed into the proposition 'Economic sanctions are effective when they are multilateral and if the sanctions are enforced by an international organization'.

Somewhat strikingly, there is a mismatch between the degree to which case studies are seen as suitable for each of the three research purposes and the extent to which they have been discussed before in the literature. There is hardly any disagreement across disciplines that case studies are appropriate for exploratory purposes, that is, the formation and refinement of hypotheses (Beck 2010; Eisenhardt 1991; Eisenhardt 1989; Flyvbjerg 2006; Gerring 2004; Ichniowski and Shaw 2011; Odell 2004, 66–7; Stryker 1996). On the other hand, it is strongly disputed whether case studies can be used to test hypotheses (Odell 2004, 70). Some scholars clearly argue in favor of hypothesis-testing small-n research (Anckar 2008; Bennett 2005; Savolainen 1994;

Seawright and Gerring 2008), while others are much more skeptical and deny the utility of case studies for this purpose (Beck 2006; Munck 2005).

The fact that the discussion of hypothesis-testing case studies has been a 'battlefield' (Kittel 2005) for decades now may explain why much thought has been invested in this research purpose. Moreover, one may believe that an in-depth discussion of case studies on the formation and refinement of hypotheses is not necessary, as these are exploratory endeavors that do not require many principles. This would be a misleading impression, though, because the research-design tasks that one confronts in small-n research are very similar across all three research goals. Figure 1.1 highlights this through the comparison of the stylized research process for each research purpose.

All three variants of theory-centered case studies necessarily start with the formulation of concepts (Sartori 1970). Small-n research testing hypotheses proceeds with the specification of at least one proposition, while the hypothesis-modifying variant identifies the hypothesis deserving improvement. Subsequently, all case studies continue with the choice of cases and the empirical analysis, which is exploratory in hypothesis-building and hypothesis-modifying research and confirmatory otherwise. The final step involves the evaluation of the collected evidence in order to develop or modify a hypothesis or to evaluate the veracity of a hypothesis that was derived at an earlier stage.[17]

Comparison of the three types of research goals shows that they have to address very similar tasks. In light of this insight, one can imagine that case studies focused on the formation and modification of hypotheses do follow

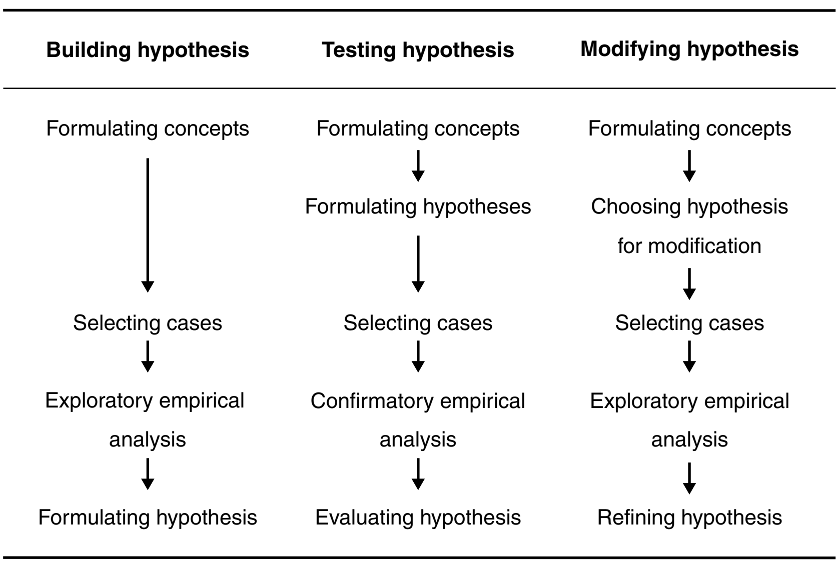

Figure 1.1 The research process for three theory-centered research goals

guidelines and principles as well. As will become clear throughout the book, genuine hypothesis-building case studies are particularly demanding because the basis for tests and modifications of hypotheses is theory. Theory is necessarily weaker for case studies that set out to generate new hypotheses in the first place. However, this does not imply that 'anything goes', as there are rules that distinguish better from worse exploratory case studies. Furthermore, a principle-based implementation of hypothesis-building case studies is important because it lays the best possible foundation for subsequent tests of propositions. This is not to say that nonstandardized small-n research is useless, as it is likely to lead to some interesting insights from which hypothesis-testing research can benefit. However, rigorously crafted case studies that build hypotheses make it much easier to test them in a follow-up study.

1.4 The cross-case level and the within-case level of analysis

It is essential to be clear about one's research purpose at the outset of a case study, but the research goal alone does not tell us anything about the level of analysis on which the case study operates. Here, 'level of analysis' refers to the level that is of theoretical interest and not the level on which the empirical analysis is located (see below). The two available levels of analysis thus understood are called the cross-case level and the within-case level.[18] In general, the cross-case level is the level on which a causal effect is theorized and examined, while the within-case level is concerned with causal mechanisms and causal processes (also called causal chains). Depending on the definition of causal mechanism, a within-case analysis is not necessarily dealing with processes and mechanisms at the same time, but can also focus on either of the two. In a within-case study of causal mechanisms, one asks what the factor or factors are that produce the cross-case relationship. An inquiry into causal processes is focused on a sequence of intervening factors that elucidate step by step how the cause brings about the outcome (see Section 2.2 and Chapter 6).

The link between the two levels and components of causal relationships can be exemplified with democratic peace theory, which tries to explain why two democracies – a democratic dyad – are almost certain to maintain peaceful relations (Rosato 2003). The central cross-case hypothesis of democratic peace theory simply states that two democracies are at peace with each other. The hypothesis is about the causal effect of a democratic dyad on the nature of foreign relations between the two countries in terms of war and peace.

A within-case hypothesis of the democratic peace phenomenon could specify the mechanism that explains why two democratic countries refrain from fighting each other, a question left implicit by the cross-case proposition. One possible mechanism could be 'commitment to democratic norms', as this entails that the political elite is committed to peaceful conflict resolution and trusts that other countries adhere to the same principle. The within-case analysis then serves to deliver observations that support this proposition, for example, by the analysis of internal documents on the decision-making process and interviews with members of the political elite. An observation that corroborates the within-case hypothesis might be a statement by a politician: 'We decided not to escalate the crisis with the other democracy because we knew that we could trust that they wouldn't opt for escalation either.'[19]

A hypothesis on the causal mechanism does not necessarily imply an expectation about the causal process. In this example, the mechanism 'commitment to democratic norms' leaves implicit the precise sequence of events in which a democratic dyad brings about peace. Events that could constitute a causal chain and inform causal inference on the process could be the mobilization of the military in response to increasing tension between the two countries, public demonstrations for peace, cabinet deliberations about the proper course of action, a summit in which the leaders of both countries meet, and so on (see Chapter 6). The democractic peace example demonstrates that it is analytically useful to distinguish between the cross-case level and causal effects, on the one hand, and the within-case level and causal mechanisms and processes, on the other, as doing so allows one to examine the democratic peace phenomenon from two different but complementary angles (Gerring 2005; Runde and de Rond 2010).

Equally important is the fact that even if the case study has a theoretical focus on one level, the other level also plays a vital role in the course of the case study. The varying roles that the cross-case and within-case levels might play can be described as a *means-end relationship*. The theoretical level of analysis corresponds with the theoretical end. The respective other level then constitutes the empirical means for achieving the theoretical end. Table 1.2 provides a sketch of the functions that the cross-case and within-case levels play if both are the end, or if one level is the theoretical end and the other one the empirical means.

The upper-right cell captures a constellation wherein the cross-case level is instrumental for the choice of cases for process tracing. Using the example of democratic peace theory, it is valuable to develop and/or test a within-case hypothesis that sheds light on why the democratic peace phenomenon exists. The theoretical end of such a study is therefore on the within-case level. But the cross-case level remains relevant because it is an empirical means for the within-case analysis in that it provides the basis for case selection.[20] As regards democratic peace theory, for instance, assume that you hypothesize that democratic norms are the mechanism accounting for peace among democracies. To test this hypothesis, one first has to identify two democracies that are at peace with each other because the selected case may otherwise be theoretically irrelevant (e.g., a nondemocratic dyad at war).

Table 1.2 Ends and means of cross-case and within-case levels

		Cross-case level	
		End	*Means*
	End	Cross-case and within-case analysis	Case selection for within-case analysis
Within-case level	*Means*	Establishing link between cause and outcome, or scoring of cause/outcome on cross-case level	–

The lower-left cell includes the reverse means-ends relationship where the theoretical end is to discern what the causes for a given outcome are and what their causal effect is. A qualitative within-case analysis is the empirical means because exploratory process tracing can hint at possible causes and their effect on the outcome. On the basis of the process tracing evidence, one might also conjecture about the causal mechanisms and processes that underlie the cross-case relationship. In principle, though, one can confine the within-case analysis to an instrument serving to generate a cross-case inference that a cause is or is not tied to the outcome.

Moreover, a within-case analysis presents the means for a cross-case analysis when it is difficult to assign a score to the cause and/or the outcome (Mahoney 2010, 125–8). For example, take Dür's cross-case hypothesis, that increased exporter lobbying prompts a government to negotiate a liberalizing trade agreement with a foreign country (2007b, 458). In this instance, the cause – an increase in lobbying activities – is notoriously difficult to measure with off-the-shelf indicators. In order to obtain a measure for lobbying intensity, Dür does a within-case analysis by consulting records of the US Congress and counting the number of exporters that gave testimony in its trade policy hearings.

The upper-left cell of Table 1.2 refers to the assessment of a *causal explanation* in the event that both the cross-case and within-case level constitute the theoretical end. In such analyzes, the cross-case level is also necessarily the empirical means for the within-case level, and vice versa. The lower-right cell of the table is left blank because it is impossible for the cross-case and the within-case level to represent only a means. If the case study has a theoretical ambition, which I suppose in this book, one of the two levels must capture the theoretical end of the analysis.

Building on the three types of means-ends relationships, it is now possible to derive six generic forms of case studies. The six types are obtained by

Table 1.3 Six generic types of case studies

		Level of analysis		
		Cross-case	*Within-case*	*Both together*
Number of cases	*One*	Single cross-case study	Single within-case study	Integrative single-case study
	More than one	Cross-case comparison	Within-case comparison	Integrative comparative case study

intersecting the three variants in Table 1.2 with the common distinction between *single-case* and *comparative case studies*. In the analysis of single cases as well as in comparative small-n research, it is feasible to have either the cross-case level or the within-case level as the theoretical end or to aim for an integrative analysis and theory. Altogether, this produces six generic forms of case studies (Table 1.3): single cross-case studies,[21] single within-case studies, cross-case comparisons, within-case comparisons, and integrative single- and comparative-case studies. In this book, 'case study' is therefore broadly defined and so subsumes a class of six designs dealing with different levels of analysis and numbers of cases.

1.5 Correlational v. set-relational causation

The third dimension important to consider in case study research pertains to the distinction between correlational and set-relational causal effects. This dimension constitutes a cleavage within the social sciences, as correlations and set relations represent two fundamentally different perspectives on what type of effect a cause has on an outcome. The distinction between correlations and set relations cuts across the more established differentiation between large-n and small-n research (see Beck 2010, 2006; Brady et al. 2006; Collier et al. 2010; Mahoney and Goertz 2006; Onwuegbuzie and Leech 2005; Prakash and Klotz 2007).[22] Thus, the ontologically grounded belief in a covariational or set-relational causal effect is independent of the number of cases one examines, and so, one can take two different perspectives on causal effects in case study research.

The differences between the two conceptions of causal effects can be exemplified with a case study seeking to explain the level of tariffs on agricultural imports in OECD countries (see Park and Jensen 2007). Assume that the tariff level hinges on the agricultural producers' capacity to organize for collective action, as this is directly related to lobbying success. Owing to the different nature of correlational and set-relational effects, this general expectation can be translated into two dissimilar hypotheses. In a correlational view, one hypothesizes that the higher the producers' degree of organization, the higher the tariffs on agricultural imports. In other words, the relationship is expected to be symmetric; tariffs increase as the level of organization increases and decline as the level of organization declines. In contrast, a set-relational hypothesis (stipulating a sufficient condition) posits a link between a high degree of producers' organization and high tariff levels. The implication that one is neither interested in the consequences of low levels of organization nor the determinants of low tariffs highlights the *asymmetric* nature of set relations (Grofman and Schneider 2009, 662).

The correlational and set-relational formulations of an effect of organizational capacity on tariff levels have apparent theoretical implications because they entail different causal inferences. Furthermore, in order to be able to generate valid cross-case inferences, it is mandatory to consider the implications of correlational and set relationships on the proper realization of case studies. One central goal of this book is to elaborate on these implications at various stages of the research process. Of course, covariational and set-relational causation figure prominently in the existing literature on the case study method (for example, George and Bennett 2005; Gerring 2007a; Goertz 2003). However, what was heretofore missing was a systematic and comparative clarification of the way in which covariational and set-relational small-n research ought to be implemented. Instead of choosing a side in the ontological dispute about whether correlations or set relations govern in the social world, I follow the principle that method should follow theory and aim to detail how to proceed in case study research *if* one adopts one conception of causal effects or the other.

1.6 Frequentist v. Bayesian causal inference

A final, salient distinction that pertains to hypothesis *tests* concerns the mode of causal inference in terms of frequentist versus Bayesian causal inference (Bennett 2006). Pending a more detailed discussion of the two modes in Chapter 8, they can be distinguished along three interrelated criteria (Table 1.4) (Howson and Urbach 2005; Bennett 2008).

The first element is epistemological and concerns the basis for the generation of causal inferences. In frequentist case studies, a hypothesis is judged on the basis of the number of observations that support and contradict it. Bayesian case studies, on the other hand, also emphasize the 'probative value' or expectability of an observation (Bennett 2008, 711), that is, how *likely* it is to collect in light of the hypothesis under scrutiny.

In order to illustrate this difference, assume that one aims to explain why some EU member states fail to transpose European directives within the mandated period of time (for example, Kaeding 2008). According to one

Table 1.4 Frequentist and Bayesian causal inference

	Mode of causal inference	
	Frequentism	*Bayesianism*
Empirical basis for inferences	Number of confirming and disconfirming observations	Likelihood of confirming and disconfirming observations
Central question	How likely is it to collect the evidence, given the assumption that the null hypothesis is correct?	How likely is it that the working hypothesis is correct, given the collected evidence?
Feasible inferences	Hypothesis either confirmed or disconfirmed	Degree of confidence in the hypothesis ranging from 0 (completely sure it is wrong) to 1 (completely sure it is correct)

line of reasoning, the misfit between the content of the directive and the national law(s) concerned by the directive can account for a delayed transposition because high misfit entails large adaptation costs. A frequentist case study might focus on a case where the misfit between the national law and the directive was high. The case study would consist of interviews with public officials and the analysis of primary documents in order to discern whether misfit and concerns about adaptation costs due to the transposition of the directive explain the huge delay.

A case study researcher thinking in Bayesian terms would criticize this case study because high misfit made it very likely that the within-case analysis delivered confirming evidence. Consequently, a Bayesian case study would focus on a case where a small misfit goes along with a huge delay. The rationale is that a minor misfit implies low adaptation costs that are unlikely to produce a massive delay in the transposition of directives. If interviews and primary sources nevertheless point to the eschewal of adaptation costs as the source for delay, this surprising finding leaves one with more confidence in the misfit hypothesis than in the frequentist case study (but see Chapter 8).

The second criterion pertains to the question that one asks with respect to the gathered evidence and the hypothesis. Frequentist case studies consider how likely it was to collect the observations that were assembled, given the assumption that the null hypothesis is correct. This means that the more observations are in line with the hypothesis, the less likely it is that the empirical evidence is attributable to chance. Bayesianism takes the opposite perspective and asks how likely it is that the hypothesis is correct given the collected evidence. Much as in quantitative research (Schrodt 2010), frequentist research makes inferences about the likelihood of evidence taking a hypothesis for granted, whereas Bayesian analyzes generate inferences about hypothesis in light of empirical evidence.

Third, the different perspectives that frequentism and Bayesianism take go along with different degrees of confidence in the generated inferences. Frequentists make an either/or inference, meaning that one either concludes that the evidence supports the hypothesis or not.[23] Although one may not be fully confident in whatever inference one makes – owing to limited access to sources, for example – there is no formalized way to express this uncertainty because frequentist case studies are limited to a yes/no conclusion in the end. Bayesian small-n research differs, as it allows one to attach a specific level or range of confidence to a hypothesis (Bennett 2006, 341). It is therefore possible to state that one is rather certain, say about 80 percent, that the hypothesis is correct, given the collected evidence and depending upon the level of confidence that one had prior to the empirical analysis.

Applying the two criteria to the transposition example, a frequentist case study would specifically need to consider how likely it is that many public officials mentioned high adaptation costs as the reason for a huge delay in the interviews if this were not the reason for the delay in fact. The more officials refer to large misfit and high costs, the less likely it becomes that these are not the actual causes for the transposition failure in the selected case. Consequently, one would conclude that misfit and adaptation costs are the cause of the delay.

A Bayesian case study would use the same evidence – internal documents and officials referring to adaptation costs as the cause of a delayed transposition – for the generation of a different causal inference. As explained above, the Bayesian case study would center on a case for which it is unlikely that misfit accounts for a delayed transposition. If we gather evidence now that supports the misfit hypothesis, we can be much more confident in the accuracy of the misfit hypothesis than before the empirical analysis because the proposition could master a relatively tough test (this argument will be qualified in Chapter 8).

On account of the apparently distinct and divergent nature of frequentist and Bayesian designs, the small-n literature tends to pit the two modes of causal inference against each other. Particular emphasis is put on the epistemological basis in terms of the number of observations in frequentist case studies and the probative value of observations in Bayesian research (Bennett 2008, 708). Chapter 8 shows that the differences between the two modes of causal inference are exaggerated because although they differ in some respects, they are similar in others. One salient difference rests in case selection because it follows different rationales and purposes (to be discussed in Chapter 3). However, principled arguments and an empirical example underscore an important similarity between frequentism and

Bayesianism, namely that, ceteris paribus, causal inferences are always more credible the larger the number of confirming observations. In other words, good Bayesian case studies take recognition of the fact that they entail an element of frequentism as well.[24]

In addition to a comparison of frequentism and Bayesianism, Chapter 8 includes a full exposition of Bayesian case in a simple and slightly formalized perspective. A formalized treatment of Bayesian small-n research sheds new light on some long-standing and intuitively plausible recommendations. On the basis of a formalized discussion and empirical examples, it is furthered explained how to craft a Bayesian case study in order to gain the greatest inferential value.

1.7 Outline of the book

The elaboration of the case study method broadly follows the research process as depicted in Figure 1.1.[25] In order to lay the groundwork for these chapters, it is necessary to clarify the key concepts and terms in Chapter 2, beginning with a definition of a case and a qualitative case study. In addition, the distinction between the cross-case and within-case level is related to the differentiation between *data set observations* and *causal process observations* as two types of evidence on which one can base causal inferences. Furthermore, the chapter elaborates on the meaning of and distinction between causal effects, causal processes, and causal mechanisms, followed by a detailed discussion of different definitions of 'mechanism' and their implications. Afterwards, the second chapter discusses in more detail correlational and set-relational causation as two manifestations of causal effects. This includes an introduction to two different measurement strategies in correlational case studies that have so far been largely ignored but are important to reflect on because they entail slightly different implications for causal inference. In addition, I give a general exposition of interaction effects in correlational case studies and introduce the reader to the full repertoire of varieties of set relations.

Chapter 3 deals with different *types of case studies,* for example, the typical case study and the most likely case study, and develops *case selection* guidelines for each type. It is shown that the research goal, the level of analysis, and the nature of the causal effect need to be considered concurrently to allow the appropriate choice of cases. In addition and with an eye on the distinction between frequentist and Bayesian causal inference, it is productive to distinguish between distribution-based and theory-based case selection strategies. The distribution-based choice of cases is tied to frequentist causal inference, whereas theory-based selection is linked to Bayesianism. The chapter also touches on the long-standing question of whether one can use the same case for the formation and test of a hypothesis.

Chapter 4 introduces the reader to cross-case comparisons and generation of inferences. Through reference to the distinction between correlations and set relations, it details how to perform intelligible cross-case comparisons for the formation, testing, and modification of a proposition. First, following the, often implicit, assumption that a comparison involves two cases and relies on dichotomous measurement of causes and outcomes, the chapter reflects on Mill's method of agreement and method of difference (1874) as the two arguably most famous forms of comparisons. In addition, the discussion extends to other variants of comparisons, as the method of agreement and the method of difference are not always the most adequate design. Chapter 4 shows that all cross-case comparisons face problems of establishing internally valid causal inferences. The problem is that the *property space,* understood as the logically possible number of scores on the causes, usually exceeds the number of cases to larger or smaller degree. This in turn implies that one can read more than causal inference into the cross-case pattern of scores. The problems of generating cross-case inferences have led to strong criticism of Mill's methods and the claim that they should be abandoned altogether. However, I argue that the criticism is overdrawn because the problems are independent of Mill's methods and one can rely on various tools for strengthening cross-case inferences. In addition, the critics of cross-case inferences ignore that the cross-case level constitutes the empirical means when the analysis of causal mechanisms and processes is the theoretical end.

Building on the insights in Chapter 4, Chapter 5 evaluates the pros and cons of five instruments for the improvement of cross-case inferences. The first section elaborates on how an explicit consideration of *time* and the *unit of analysis* can promote the comparability of cases. The remaining four sections examine the relationship between a large property space and a small number of cases. With regard to the latter aspect of the indeterminacy problem, an often recommended but limited means to improve comparisons is an increase in the number of cases. The third instrument pertains to the *level of measurement aggregation* of the causes and outcome. The level of measurement aggregation asks whether causes and/or the outcome are measured binarily or multicategorically. It is shown that binary measurement is inferentially advantageous because the property space is reduced. However, this upside has to be weighed against the downside that binary measures only offer a coarse-grained conceptualization of empirical phenomena. The fourth tool is, quite simply, better theory because the more possible causal inferences one can plausibly dismiss on theoretical grounds, the better. Finally, one can transform potential causes into *scope conditions* by deliberately downsizing the population of cases and the scope to which a hypothesis is supposed to apply.

In Chapter 6, the discussion turns to the role of theory and temporality for *process tracing.* As regards the nexus between theory and process tracing, I claim that one can theorize what I call a *realized process* (or sequence of intermediate steps) and *anticipated processes*. The first section of the chapter introduces the two forms of theorizing about processes and supplements them with empirical examples. The second and third sections are devoted to a comparative discussion of the pros and cons of case studies on realized and anticipated processes. They demonstrate that it is important for case study researchers to know the process of theoretical interest because inquiries into realized and anticipated processes have important ramifications for process tracing. Building on the first two sections, the third part expands on the role of *time* and *temporality* for process tracing.

This summary of Chapter 6 hints at the fact that it is reserved for the elaboration of issues that are unique or at least affine to the logic of causal inference via process tracing. For this reason, the chapter excludes several issues that are at the heart of the current literature on process tracing. Many of the issues subsumed under process tracing in the present debate extend to cross-case comparisons are therefore postponed until Chapters 7 and 8, which take a more general perspective.

In this spirit, Chapter 7 considers three elements of causal inference that hold regardless of the four dimensions. Section 7.1 first takes an epistemological perspective and discusses the handling of different types of sources with a focus on the *source coverage problem* and potential presence of *source coverage bias*. The next section ties the problem of generating cross-case inferences, introduced in Chapter 4, to process tracing because it is frequently argued that the latter diminishes the former or eliminates it altogether. Principled arguments and an empirical example show that it is an empirical question of whether process tracing can decrease the extent of indeterminacy. Finally, the chapter deals with *counterfactual reasoning,* which can be of great value when some indeterminacy remains after the empirical analysis. Criteria of good counterfactuals are presented in combination with empirical illustrations in Section 7.3.

The discussion of causal inference continues in Chapter 8 with a consideration of *frequentism* and *Bayesianism* in hypothesis-testing case studies. The chapter starts with a review of an established typology of four different types of hypothesis tests that is based on the distinction between the uniqueness and the certainty of observable implications that are related to a hypothesis, the latter capturing Bayesian causal reasoning. Section 8.2 first reconsiders the dimension of uniqueness and contends that it is currently misinterpreted. In order to remedy this shortcoming, it is proposed to a third dimension 'contradiction' that captures whether two or more propositions yield mutually exclusive predictions. In Section 8.3, frequentist and Bayesian causal inferences are systematically compared with respect to the tasks that they entail for different stages of the research process. The last two sections introduce moderately formalized Bayesian causal inference with an elaboration of *Bayes' theorem* and *Bayes factor*. The discussion shows that some intuitively plausible arguments on Bayesian causal inference deserve qualification once one takes a formalized perspective.

Chapter 9 turns to questions of *external validity* and *generalization*. Section 9.1 closes the circle with respect to Chapter 3 by linking different types of case studies to the scope of generalization of causal inferences. The next two sections consider strategies for the improvement of generalization. Section 9.2 explains how the generalization of causal inferences can be expanded and downsized via the disciplined modification of scope conditions. The final section picks up the criticism that it is only possible to generalize on the basis of case studies under very demanding assumptions, if at all. In response to this criticism, it is explained how the transfer of causes into scope conditions can be used to delineate intelligible and small populations that allow for generalization. Moreover, I introduce the strategy of *layered generalization* for the systematic and stepwise expansion of populations in case study research. This strategy transforms actual scope conditions into potential causes in order to discern their relevance for the outcome under scrutiny.

Chapter 10 concludes with a larger picture of the previous chapters and highlights the fact that the case study method can contribute to causal inferences on causal effects and mechanisms if it is implemented in a disciplined and reflective manner.

2

Case, Case Study, and Causation: Core Concepts and Fundamentals

Case studies are performed in a wide array of disciplines such as political science, sociology, history, and, to a lesser degree, economics. The philosophical foundations of the case study method vary across disciplines as well as within each discipline (Jackson 2010). The general notion of small-n research does not preclude any philosophical grounding, making it particularly necessary to clarify the foundations on which my discussion of the case study method rests. The chapter therefore is not about the philosophical foundations of *the* case study method but more narrowly about case studies interested in inferences about empirical regularities.[1]

This chapter starts with a clarification of what a case study and a qualitative case study are and what the *n* in small-n research refers to. In Section 2.2, the distinction between the cross-case and the within-case level is tied to the now established distinction between data set observations and causal process observations. The defining features of the two types of observations are introduced, particularly in respect to the fact that they permit researchers to generate inferences about causal effects and causal mechanisms, respectively.

Taking the definitional clarifications as the basis, Section 2.3 details the distinction between causal effects, causal mechanisms, and causal processes as the objects of causal inferences in qualitative case studies. Section 2.4 then reflects on the differentiation between case studies interested in the causes of effects and the effects of causes. The two styles of causal analysis have recently received increased attention and need to be included in this discussion to further clarify the scope of the arguments made in the following chapters.

The final section focuses specifically on causal effects and notes the distinction between covariation and set relations as two ways of thinking about causation on the cross-case level. Different forms of correlational and set-relational causal effects are introduced in order to explain in the subsequent chapters how they influence the conduct of small-n research and the production of causal inferences.

2.1 Definitions and clarifications

Case, case study, and the *n*-question

The definition of 'case' that underlies this book follows many existing conceptualizations of it (for example, Gerring 2004; Levy 2008). A case is defined as *a bounded empirical phenomenon that is an instance of a population of similar empirical phenomena*. The two attributes of the definition – 'bounded empirical phenomenon' and 'instance of a population of similar phenomena' – require elaboration. The latter attribute refers to the inherent meaning of a case as being a case of something of which there are more empirical instances that together form the population of interest (Gerring 2004; Ragin 1992). The membership of a case within a population implies that making generalizations is deemed to be feasible because all cases in the population are similar. 'Similar' does not mean that cases are perfectly identical but that they are assumed to be similar with respect to a specific research question (Sartori 1991).

The relevant criterion is *causal homogeneity*, which signifies that a cause–effect relationship is, on average, expected to hold true for the cases within the population (Collier et al. 2004c, 29). The understanding of what similar cases are, and thus the shape of the population, may change throughout the research process (Ragin 2000, chap. 2). Such changes can be due to the need to redefine concepts (Adcock and Collier 2001), to add or remove scope conditions (Walker and Cohen 1985), or to exclude individual cases from the population (Mahoney and Goertz 2004). For these reasons, one should always stay open-minded to a change of the population and an understanding of causal homogeneity during and after the empirical analysis. Although the delineations of the population and cases are always preliminary to some degree (Ragin 1992; 1997, 30), it is nevertheless mandatory that one has a specific population in mind at a given stage of the research process because the proper implementation of a case study and causal inference are tightly linked to the composition of the population (see Chapters 3 and 9).

The second attribute of the definition of 'case' is as a bounded empirical phenomenon. This attribute is ambiguous in order that it can be open to the broad range of bounded phenomena that can be examined empirically. The general principle is that a given set of boundaries should fully circumscribe a case. This means that a case is an empirical entity that is *exhaustively* delineated by a certain number of boundaries.[2] This criterion can be concretized because every case has a *temporal* and a *substantive* bound.

The central role of these two bounds can be clarified with an empirical example from welfare state research. Suppose that you are interested in the determinants of radical welfare state retrenchment and select New Zealand in 1991 as a case of radical reform (see Starke 2008). The temporal end point of this case could be the decision of the national parliament as the last step of the political decision-making process. Since there is some leeway in the implementation of political decisions, one may also opt for an extended period of analysis and include the implementation stage into the case study. The temporal bound on the case is completed by the specification of the case study's starting point. Again, it is up to the researcher to justify this decision because there rarely is a natural point to start with. One can take a narrow focus and examine retrenchment from the first time a reform was discussed in the political arena. Alternatively, one may emphasize distant causes and historical causation (Pierson 2004, 95; Stinchcombe 1968, chap. 3) and trace how decisions made decades ago resulted in radical welfare state retrenchment at a considerably later point in time. But it is evident that even case studies emphasizing historical causes must also impose a temporal bound and opt for a point in time at which the empirical analysis begins.

The second type of bound that every case has is substantive. In the given example, 'radical welfare state retrenchment' signifies the substantive limits of the case. This substantive boundary has three elements. The case study is about a radical cutback of the welfare state and excludes nonradical retrenchment. In addition, it is a case study of retrenchment instead of expansionist and status-quo oriented welfare state policies. Finally, the case is substantively limited to the welfare state. In practice, this requires delineating the welfare state from other policy fields that one could examine as well. This example shows that a case always has a substantive boundary because one is concerned with a specific empirical manifestation of the conceptualization of the outcome – such as radical welfare state retrenchment – that is specified at the outset of a case study. One does not simply observe a reform or policy making but radical retrenchment in the domain of the welfare state, thereby substantively delimiting the case to a specific policy field and form of policy making.

The temporal and substantive limits must be complemented by at least one other bound in order to fully circumscribe a case. The boundary or boundaries that one additionally needs to invoke depend on the given research question. The example of radical welfare state retrenchment exemplifies this because the case additionally has a spatial and, probably, an institutional bound. The welfare state can be understood as an institution, implying the need to determine the elements of this institution that are covered by the case study. Is the case study only concerned with formal rules? Or does it also extend to informal procedures? These are some of the questions that one needs to address in delineating the institutional boundary of a case. The meaning of a spatial bound is more apparent because the radical reform of the welfare state is committed by New Zealand. If radical retrenchment in Germany had been selected, the spatial limits of the case would be territorial boundaries of Germany. Many cases in the social sciences have a spatial bound, and sometimes it is even made a defining feature of cases (Gerring 2007a, 19). However, this would overstate the importance of spatial boundaries because a case does not always have a spatial limit. This is particularly true of the field of international relations and research on international institutions. Consider a case study on the effects of judicialization of dispute settlement procedures in GATT and the WTO on the compliance of member states with the procedures (Zangl 2008). Since the case study concerns the judicialization procedures of the GATT and the WTO as two international institutions, the notion of a spatial boundary of the case is futile here.

The discussion of the multiple boundaries describing a case implies that every case can be understood as a *multidimensional* empirical phenomenon. A case is multidimensional because one imposes boundaries on a specific dimension and the bounds on all dimensions together constitute the case. The previous discussion of the 1991 case of radical welfare state retrenchment in New Zealand implicitly referred to bounds on four dimensions: a temporal dimension, demarcating the period of analysis; a substantive dimension that includes three subdimensions ('radical' 'retrenchment' of 'the welfare state'); a spatial dimension (New Zealand); and an institutional dimension (if one conceptualizes the welfare state as an institution worthy to be delineated). It will be shown that the understanding of cases as multidimensional and bounded phenomena contributes to the rigor of a case study, particularly during the case selection and generalization stage.

In concluding the discussion of bounds and dimensions, one should be aware of the differences between the boundaries and dimensions of a case and those of a population. A population comprises all positive and negative cases for the causal relationship of interest (Mahoney and Goertz 2004). If cases of radical welfare state retrenchment are positive cases because radical retrenchment is the event of theoretical interest, instances of nonradical welfare state reform could be taken as relevant negative cases.[3] Scope conditions delineate the boundaries within which a causal relationship is expected to hold and that must be met by positive and negative cases alike (Walker and Cohen 1985).[4] In the welfare state example, OECD membership could be a spatial scope condition indicating that the generated causal inferences are limited to countries having reached a certain level of economic and democratic development. The post–Cold War period, a viable temporal scope condition, would highlight the fact that one does not aim to infer anything about welfare state reforms before the end of the Cold War.

With regard to the relation between the bounds of a case and the population, the boundaries of a case are very likely to be smaller than the boundaries of the population on one dimension at least (otherwise one would be dealing with a population of size one). In the given example, the spatial boundary of the population is the OECD world, whereas the spatial limits on the specific case are the boundaries of New Zealand. It is exactly this gap between the bounds of a case and the population that creates problems of case selection, comparisons, and generalization.

I conceptualize a case study as the *empirical analysis of a small sample of bounded phenomena that are instances of a population of similar phenomena*. This definition does not impose any limits on what data and techniques of data collection and data analysis are used. Although this book is exclusively concerned with qualitative case studies ('qualitative' being defined below), the definition of a case study should not preclude the possibility that it can be qualitative and quantitative (Gerring 2007a, 10–11). To give an example of a quantitative case study, imagine you are interested in the effects of political communication on electoral success in US presidential elections. In a quantitative case study, one would analyze the campaigns of the Republican and Democratic candidates and try to link the success or failure of each candidate to his or her communication strategy. A quantitative text analysis of the hundreds or even thousands of press statements of the two candidates could be used to classify communication strategies as being either aggressive or conciliatory (for example). The number of cases is two – the successful candidate's campaign and the unsuccessful candidate's campaign – whereas the number of within-case observations – the statements made subject to a quantitative analysis – is quite large.[5]

The exclusive use of statistical tools for causal inference is what I take as the defining feature of a quantitative case study. It is not the use of numbers and figures that qualifies a case study as quantitative (Maxwell 2010, 476–7) but the exclusive analysis of data with statistical means. Consequently, a case study is designated as qualitative when the mode of causal inference involves different or at least additional elements of qualitative assessment.[6] In contrast to genuine quantitative case studies, the pool of collected evidence is more diverse and includes primary and secondary sources as well as interviews yielding observations that cannot be made subject to a quantitative analysis (Collier et al. 2004b).[7] In light of this definitional exercise, there is no unique case study method because a wide range of quantitative and qualitative tools can be used. However, since this book is exclusively concerned with qualitative case studies, it should be understood that the term 'case study method' and 'case study' refer only to qualitative small-n research in the following.

The last definition to address in this section concerns the meaning of the famous *n* in small-n and large-n research. For some time at least, qualitative and quantitative researchers seemed to have a different understanding of what *n* captures (Gerring 2004; Goldthorpe 1997a). Quantitative researchers often referred and sometimes still refer to *n* as the number of cases (Beck 2006; King et al. 1994). Since the number of cases is small in qualitative case studies, it seems straightforward to argue that they suffer from inferential problems because it seems difficult to make strong inferences about regularities with a small number of cases. However, since Campbell's famous article on within-case analysis (Campbell 1975), case study researchers emphasize that the number of cases may be small but that the number of observations (number of interviews, number of primary sources, and so on) can be quite large. If one prefers to use n for signifying the number of observations, it is then evident that n can be fairly large in case studies. In order to avoid misunderstandings of what I understand n to be, I consistently use n to denote the number of cases. According to this definition, it is possible to equate the conventional qualitative case study with small-n research and conventional quantitative research with large-n analyses.

Levels of analysis and types of observations

This section now turns to the relationship between the two levels of analysis and corresponding types of observations in small-n research. In recent years, it has become common to denote observations on the cross-case level as data set observations (DSOs) (Collier et al. 2004b), which implies that cross-case observations and DSOs are the same. DSOs are standardized observations that are comparable across cases. DSOs can be conceived of as observations that one collects and organizes in a data set and are amenable to a quantitative analysis. As elaborated in Chapter 1, however, DSOs are also integral to case studies because they form the basis for choosing cases for process tracing and the generation of cross-case inferences. The observations that one gathers on the within-case level are called causal process observations (CPOs) (Collier et al. 2004b).[8] In contrast to DSOs, CPOs are not standardized and are not necessarily comparable within and across cases, which is the salient difference between the two types of observations.

The difference between the two levels of analysis and types of observations is exemplified by an analysis performed by Ziblatt (2009). On the cross-case level, Ziblatt is interested in the effects of landholding inequality on the occurrence of electoral fraud on the district level in transition countries.[9] The hypothesis that landholding inequality has an effect on the frequency of electoral fraud can be assessed with data on the distribution of land and the frequency of electoral fraud in districts. Since one uses the same or comparable measures for the cause and the outcome across electoral districts, one is dealing with data set observations on the cross-case level with the opportunity to examine the causal effect of landholding inequality.

Figure 2.1 supplements the cross-case argument with a stylized process that includes two elements called *intervening steps*. In short (see also Section 2.2 on mechanisms), the argument is that, if a country undergoes a democratic transition process, landowners feel threatened because democracy undermines the political influence that they have wielded in nondemocratic societies owing to their wealth and social status. The perceived threat is greater, the more land a person owns, that is, the greater the landholding inequality, the more that person has to lose from a successful transition to democracy. This part of the explanation is captured by the intervening step

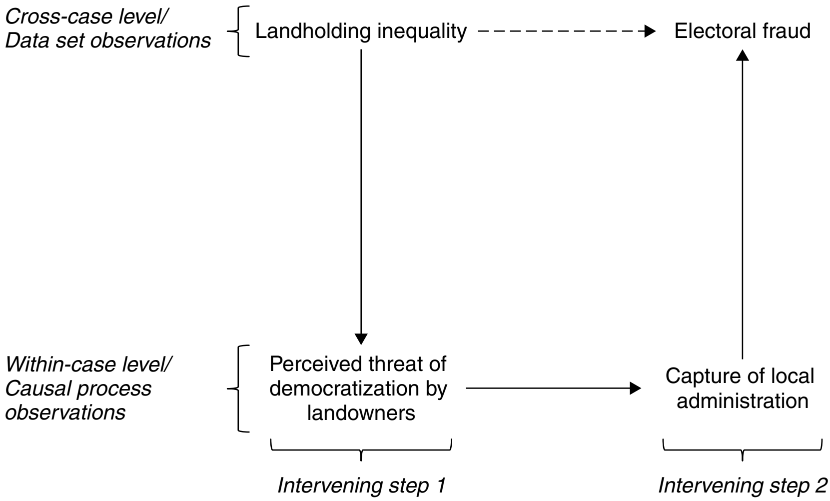

Figure 2.1 Levels of analysis and types of observations

one in Figure 2.1. As a response to the perceived threat, landowners interfere in the staffing of the local administration that is in charge of the conduct of elections. This part of the argument is captured by the second intervening step. In turn, the deployment of people who are committed to the landowner's interest accounts for electoral fraud on the district level because the officials selected by the landowner are prepared to manipulate elections in the latter's favor.

Assuming that one performs a hypothesis-testing case study and theorizes the two intervening steps presented in Figure 2.1, the task of process tracing is to gather confirming observations for each of the steps. There are two reasons why most of these observations are not comparable. First, the observations are related to two different empirical phenomena, namely, the perception of threat and the intervention in the staffing process of local administrations. Second, for each of the intervening steps, information is gathered from different sources that, although related to the same step, yields noncomparable pieces of evidence. For instance, one could derive information about the perceived threat of democratization from the records of landowners or articles published in contemporary newspapers. Information about the second intervening step could also be gathered from landowners' records and from staff records of local administrations. An indication of the influence of landowners on the staffing process would be an increase in the number of local landed nobility, that is, people from the landowner's staff in the local administration (Ziblatt 2009, 14). These insights are based on causal process observations because data derived from official staff records are not comparable to statements in diaries, which are in turn not directly comparable to newspaper reports. In contrast to data set observations, however, noncomparability is not an issue here. On the contrary, gathering and evaluating evidence on the within-case level is similar to assembling a jigsaw puzzle; every piece is more or less different from every other piece, but when put together, they deliver a full picture of the phenomenon of interest, which consists of the causal process and the constitutive intervening steps in Figure 2.1.

This example shows that DSOs and CPOs can be neatly distinguished in a specific case study. On a general level, in contrast, the distinction between the two types of observations is fluid (Seawright and Collier 2004, 277–8). DSOs can be collected for the usual suspects, such as a country's political regime or gross domestic product, but this is not necessarily true. In the electoral fraud example, the analysis of staff records in one electoral district allowed one to gather CPOs. This would be different if one were able to examine staff records in multiple districts and to code for each district whether the share of local notables increased over time or not. The resulting data would be standardized and comparable across districts; thus, staff records can be used for the collection of both DSOs and CPOs. It therefore holds that what is a CPO in one study could be a DSO in another.

2.2 Causal effects and causal mechanisms

Causation and regularities

Before the discussion specifically turns to causal effects and causal mechanisms, it is first necessary to deal with ontological issues that are important for my elaboration of the case study method. The following discussion of small-n research and methods is based on the premise that there are general patterns in the social and political world.[10] This means that case studies serve to test hypotheses and generate inferences such as 'The stronger exporter lobby groups are in a country, the more liberal the country's trade policy', and 'International negotiations about climate protocols fail when the growth rate of the world economy is low because of the opposition of industrial lobby groups'.[11]

Concerning the *epistemological* side, that is how to infer causation, Hume's (2003 [1740]) conception of causation as constant conjunction has been very influential (Brady 2008; Jackson 2010, chap. 3; Kurki 2007, chap. 2).[12] Without going into the details here, Hume argued that causation is ultimately unobservable and that causal relationships cannot be anything but inferred from observation. According to Hume, three criteria must be met for causal inference on the basis of empirical observations: the occurrence of the cause is temporally prior to the occurrence of the effect; the cause and effect are spatially and temporally contiguous; and the outcome must regularly follow the cause.

All three elements of Hume's account have received ample attention in the past and are only briefly reconsidered here (see Brady 2008). The first element – that the cause must precede the outcome – is uncontroversial (leaving aside here conceptions of backward causation arguing that the effect can precede the cause in time).[13] The second element is spatio-temporal contiguity and means that the cause and the outcome are proximate in temporal and spatial terms. The imputation of a causal link is more convincing the smaller the spatial distance is between cause and effect and the smaller the temporal distance is between the cause and the outcome.[14] However, there is nothing inherent that speaks against the analysis of distant cross-case causes. Above all, the distance between the cause and outcome should be determined by theory instead of subsuming theory under the requirement of spatio-temporal contiguity (Hall 2003). As Hume himself notes (in other words), in a cross-case analysis the requirement of spatio-temporal contiguity can be met by specifying the intervening steps linking cause to outcome. The rationale is straightforward because the spatio-temporal distance between the intervening steps that constitute the process is necessarily smaller than the distance between the cross-case cause and the outcome. (In this view, the criterion of spatio-temporal contiguity overlaps with mechanismic theories of causation, see Waskan 2011.) The third element – constant conjunction between cause and effect – denotes an invariant regularity that is without exceptions. As this is an excessively demanding criterion, I follow the criterion of causes as raising the probability of occurrence for the outcome (this is not meant to disparage alternative epistemologies). While I cannot go into the details here, this is a widely held epistemology; although certainly not without its problems (like any theory of causation), it is suitable for causal inference in case studies (Hitchcock 1995; Northcott 2010).

The question of whether the social and political world is governed by general cause–effect relationships is a long-standing one. On the one hand, there are promising signs for empirical regularities. For example, democracies almost never fight each other (George and Bennett 2005, chap. 2), and the share of cabinet posts of a party almost perfectly corresponds with its share of parliamentary seats (Gamson 1961; Warwick and Druckman 2006). Scholars denying the existence or detectability of regularities can come up with equally compelling examples and arguments (Friedrichs and Kratochwil 2009; Steinmetz 2004; Thomas 2010). The question concerning the existence of detectable regularities is an ontological one, rendering it futile to argue about who is right and who is wrong (Hay 2006). The perspective adopted in this book therefore is not a degradation of other views on causation.[15]

Returning to the distinction between the cross-case and within-case level and their centrality for causal explanations, it is evident that the premise of regularities extends to both levels.[16] On the cross-case level, one expects a systematic causal effect, while on the within-case level, there should be evidence for recurrent processes and mechanisms (Mayntz 2004). With respect to the electoral fraud example given above and the hypothesis that a capture mechanism ties landholding inequality to electoral fraud (Ziblatt 2009), two things are implied. On the cross-case level, a positive association between the degree of landholding inequality and the occurrence of electoral fraud should be discernible across multiple electoral districts. On the within-case level, one should observe that landowners try to capture the local administration, and they are more successful the more unequally the land is distributed within a district.

In order to know what a causal effect and a causal mechanism are and what to search for in a case study, the two terms are discussed on a general level in the next two sections. The following section gives a general discussion of causal effects, leaving a more detailed treatment of different varieties of correlational and set-relational effects for Section 2.4.

Causal effects

In recent years, the distinction between causal effects and causal mechanisms constituted an increasingly important dimension in the social sciences (Gerring 2008; Little 2010, 1998; Mayntz 2004; Tilly 2001). There is a consensus that the two concepts refer to different levels of analysis; causal effects have to be examined on the cross-case level, while causal mechanisms are operative on the within-case level (Stinchcombe 1991).

As regards causal effects, there is some disagreement about what qualifies as such. I propose that a purported cause can be assigned a causal effect if one observes a theoretically intelligible and systematic cause–effect relationship. This definition is deliberately broad and subsumes covariational and set-relational cross-case associations. Pending a more detailed discussion of covariational and set-relations causation below, a correlation captures the change in the outcome as a result of a change in the cause. If one is interested in the effect of the gross domestic product (GDP) on per capita illiteracy rates, the effect is correlational if the illiteracy rate decreases as the GDP increases. In contrast, set relations emphasize cause–effect relationships between an invariant cause and an outcome (sufficiency) or an invariant outcome and a condition (necessity). An illustrative set relation is the observation that whenever there is a high GDP per capita (invariant condition), the illiteracy rate of a country is low (invariant outcome). This implies that a high GPD per capita is a cause of low illiteracy, which is not the same as saying that the higher GDP per capita is, the lower the illiteracy rate (see Section 2.4).

It is occasionally argued that the search for set relations does not allow one to say anything about causal effects (for example, King and Powell 2008, 16). This assertion is based on a covariational, that is, difference-making understanding of causation that does not fit squarely with set-relational thinking. Put simply and assuming that the cause is dichotomous, the causal effect of a variable is the difference between the scores of the outcome when the variable takes different scores (Morgan and Winship 2007). If one defines a causal effect as such, it is straightforward to reject set relations as a different manifestation of a causal effect because they are characterized by invariant causes or outcomes.

However, when one aims to discern causal effects, more broadly understood as regular cross-case associations, there is no inherent reason to favor correlations over set relations, and vice versa. Instead, the choice between a correlational or set-relational causal effect should be based on theory. Elaborated theory is necessarily lacking in hypothesis-building case studies, and that is exactly the reason why they are pursued (it creates special problems that will become apparent in later chapters). It is only when a theory does not allow one to derive a decidedly correlational or set-relational hypothesis that a researcher is free to adopt a covariational or set-relational viewpoint. This is necessarily an arbitrary ontological decision, meaning that there is no ground for dismissing either of the two conceptions of cross-case relationships as inherently deficient.

Causal mechanisms

Patterns of associations on the cross-case level provide no information about whether and how cause and effect are related to each other via a causal process and causal mechanism (Abbott 1998; Bunge 1997; Freedman 1999).[17] This deficiency of cross-case research is a major reason for the search for causal mechanisms (Checkel 2008; Collier et al. 2004b; Hedström and Swedberg 1996; Weber 2007).[18] While there is some disagreement about what constitutes a causal effect, there is even more ambiguity as regards the proper definition of causal mechanisms (Gerring 2008; Hedström and Ylikoski 2010; Mahoney 2001).[19] Some scholars argue that mechanisms are observable intervening variables that connect the cause and outcome to each other (Mayntz 2004). For others, mechanisms are fundamentally unobservable entities with causal power (Demetriou 2009; Kurki 2007). Similarly, some find that mechanisms work (nearly) deterministically and independently of context (Little 1991), whereas others argue that mechanisms depend on context (Falleti and Lynch 2009) or allow for the presence of probabilistic mechanisms (Bunge 1997, 419). Arguably, the aspect about which there is the broadest consensus is that mechanisms are operative on a lower level than causal effects; that is, they are located at the within-case level (Stinchcombe 1991). However, some also ascribe a role to mechanisms in genuine macro analyses (Mahoney 2003b). This discussion of the various dimensions along which causal mechanisms are discussed exemplifies the current definitional ambiguity and shows that reference to causal mechanisms can be more confusing than illuminating (Gerring 2008, 178).

On the one hand, more definitional clarity about causal mechanisms is desirable. On the other hand, one can wonder about the extent to which the heated debate about mechanisms, which mainly takes place on the ontological dimension (what a mechanism *is*), matters for the case study method and empirical small-n research. The claim that mechanisms play out on a lower level than does the causal effect means that one should search for mechanisms via process tracing on the within-case level. With the exception of some quantitative research that takes a different view on mechanisms (Imai et al. 2011), this can be considered a truism of qualitative research. The recently emphasized argument that mechanisms possess the capacity to produce or prevent change of the outcome (Bunge 1997; Kurki 2007) is immune to direct observation in process tracing. Instead, the claim that a mechanism possesses causal power must be inferred from the observation of a spatio-temporally ordered chain of events (Waskan 2011), therefore following the conventional procedure of tying observations to concepts in order to generate inferences.

The ontological question of whether mechanisms operate deterministically or probabilistically has ramifications similar to studies that are concerned with causal effects (Lieberson 1991). Unless one is able to search for causal mechanisms in all cases of the population of interest, one is forced to make the (probably) contestable assumption that the mechanism discerned in the cases under scrutiny is operative in all or at least most cases in the population.

One could continue with additional dimensions addressed in the literature on mechanisms, but the previous discussion suffices to point to an important aspect. On one hand, different conceptions of mechanisms naturally have different implications for the realization of case studies and the small-n method. On the other hand, these ramifications pertain to established issues of empirical research, such as the subsumption of observations under a previously specified concept (Adcock and Collier 2001), which simply is the conceptualization of a specific mechanism in case studies on causal mechanisms. As long as one ensures that the realization of a case study and the generation of causal inferences match the selected definition of causal mechanism, the tasks that one confronts in the analysis of mechanisms mirror the ordinary requirements for theory-guided empirical research.

A second aspect that puts the debate about mechanisms into perspective is the theoretical concern with the process connecting cause and effect and the elements of the process that are taken for the generation of causal inferences about a mechanism. One can distinguish three different perspectives on mechanismic explanations. First, the entire process connecting cause to effect is taken as evidence for a single mechanism. Second, the process is decomposed into three parts, and mechanisms are examined on the basis of Coleman's macro-micro-macro view on macro phenomena (1998). In the third view on mechanisms, each sequence of intervening steps that jointly constitute the process is conceived of as a manifestation of a mechanism.

The three conceptions and implications for the case study method can be exemplified with an application to Ziblatt's analysis of electoral fraud in transition democracies (2009). As explained above, on the cross-case level, electoral fraud is more likely with an increasing degree of inequality among landowners, which is taken as a measure of the distribution of wealth, power, and prestige. Owners of large shares of land have much to lose from democratic rule in their country, which is why the occurrence of electoral fraud is expected to get more likely, the more unequal the distribution of land. On the within-case level, Ziblatt (2009, 14) theorizes a *capture mechanism*, denoting that landowners capture the local administration in charge of guaranteeing free and fair elections. Capture can take place either by exerting influence on local officials or by recruiting people from the landowner's staff for the local administration.

Figure 2.2 contains a simple visualization of this line of reasoning. The line with the arrow leading from the macro cause to the outcome is dashed in order to denote that there is an association but that the causal link between the two is established by a causal mechanism. A solid line runs from the cause to the mechanism, which is in turn causally related to the outcome and thus connected to it by a solid line, too.[20] Figure 2.2 exemplifies that the explanation can be referred to as a *single-mechanism explanation* because capture of the local administration is stipulated as the only mechanism underlying the macro relationship.

At this point, it is useful to introduce Machamer, Darden, and Craver's distinction between *entities* and *activities* that jointly carry the explanatory burden implicit to a mechanism (2000). In the social sciences, entities in mechanismic explanations can be countries, organizations, individuals, and so on. These entities engage in activities that ensure productive continuity between the cause and the outcome and explain why the former has an effect on the latter (Machamer et al. 2000, 3). In Figure 2.2, landowners are the entities that influence the staffing of the local administration, which is implicit to the first arrow in the figure. The second arrow then assumes

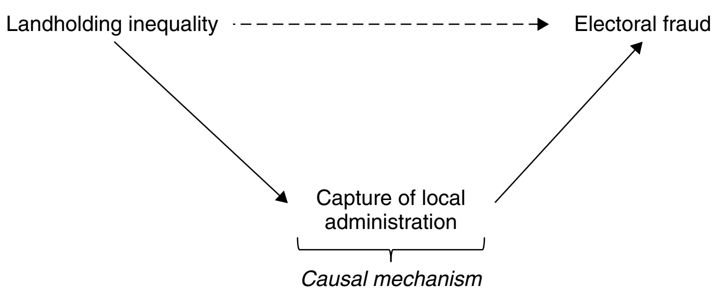

Figure 2.2 Single-mechanism explanation

that local officials are the entities that feel loyal to landowners and therefore manipulate elections in their favor. Thinking in terms of entities and activities is not uncommon in the social sciences, where these two terms are more properly referred to as actors and their behavior.[21] However, activities are not necessarily equivalent to an actor's actions and behavior (see below), speaking for the more general terminology of entities and activities. In addition, the explicit usage of this pair of terms tends to offer greater insight than the reference to a mechanism alone.

While it is perfectly legitimate to theorize entities and activities and subsume them under a single mechanism, it would be equally possible to take a somewhat more nuanced explanatory perspective on cross-case phenomena. Following James Coleman (1998, chap. 1), an explanation of a cross-case association can be decomposed into three steps. In the first step, the cross-case cause is translated into structural implications for actors on the within-case level. In step two, these implications prompt a certain type of actor behavior that aggregates across individuals to a cross-case phenomenon in step three.

Hedström and Swedberg (1996, 1998) and Hedström and Ylikoski (2010) take Coleman's conception of explanations of cross-case associations as the basis and argue that each of the steps should be supplemented with a causal mechanism. Since three mechanisms are involved, I refer to this view as the *triple-mechanism* conception of cross-case phenomena. Figure 2.3 exemplifies the triple-mechanism perspective by adding one step to the explanation in Figure 2.2.

The additional step is located between the macro cause and the capture of public officials and is part of what is referred to as the situational mechanism.

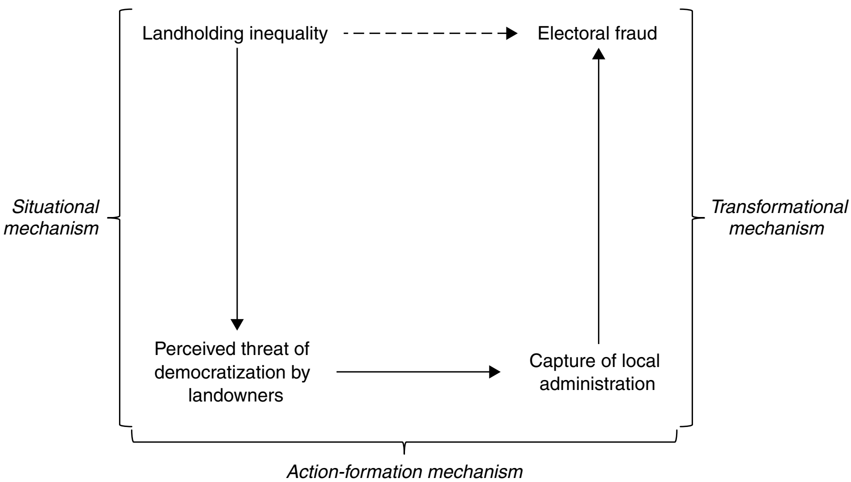

Figure 2.3 Triple-mechanism explanation

In this example, the situational mechanism entails that landowner feel threatened by democratization if their power is due to an unequal distribution of land. On the within-case level, the action-formation mechanism captures that landowners employ a counterstrategy and capture the local administration. Finally, the transformational mechanism links capture at the local level to the occurrence of electoral fraud.

Compared with the single-mechanism perspective, a triple-mechanism explanation offers a more detailed account of a cross-case relationship by decomposing it into three components. The richer account of the macro outcome is also attributable to the need for specifying more entities and activities. For example, for the first step, one has to explain why landholding inequality results in landowners perceiving democratization as a threat. The entity involved in this step are landowners that evaluate the consequences of democratization and conclude that it is a threat for their power base.

Triple-mechanism explanations add detail to the explanation of cross-case phenomena, but one can think of more fine-grained mechanismic accounts. For presentational purposes, I limit the empirical example to adding one additional step to the process. In contrast to Figure 2.3, Figure 2.4 now explicitly details that capture leads to electoral fraud because of the loyalty of captured officials to the local landowner. The inclusion of one additional step to the process converts it to what I call a *multiple-mechanism* explanation because each of the intervening steps is taken as evidence for an individual mechanism (Waldner 2012), each of which entails its own actors and activities.[22] In Figure 2.4, the link between landholding inequality and perceived threat is one mechanism (however it is labeled), the tie between perceived threat and capture a second mechanism, the connection between capture and loyalty the third mechanism, and the connection between loyalty and electoral fraud the fourth one.

The previous discussion of three different conceptions of mechanismic explanations conveys three important insights for causal inference on mechanisms. First, given that the same cross-case phenomenon can be underpinned with three different kinds of mechanismic explanations, is one of the three superior to the other two? One might be inclined to say that greater theoretical precision and leverage is always better, which would yield the multiple-mechanisms perspective as the champion. However, one should recall that mechanisms are specified for explanatory purposes and discrimination of competing explanations on the within-case level.

In fact, Ziblatt (2009, 14) notes that a traditional social power mechanism is also compatible with the cross-case association between landholding inequality and fraud. According to this mechanism, landowners as the patrons exert control over their clients in a specific area. This could be achieved by the deference of the clients and/or the patron influencing them not to cast their votes for candidates that are likely to undermine the landowners' power. Capture of the local administration is not implied

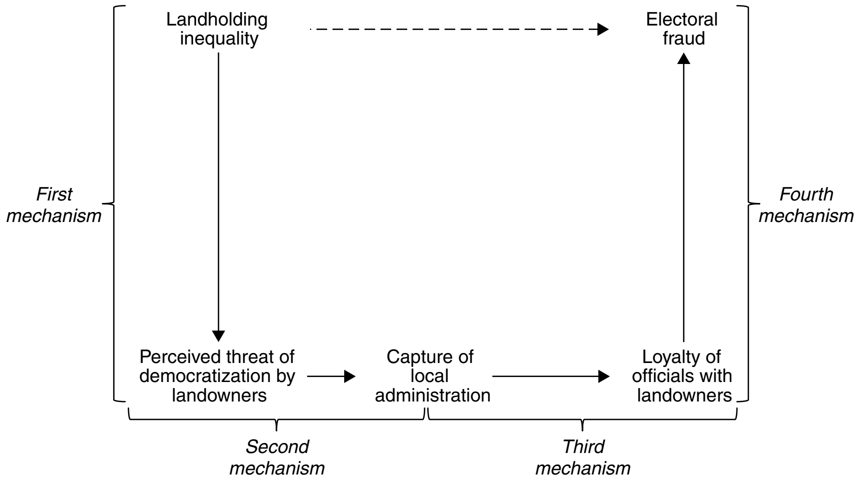

Figure 2.4 Multiple-mechanism explanation

by the traditional social power mechanism, which means that capture and the manipulation of elections by captured officials suffices to confirm the capture mechanism. From a theoretical point of view, one therefore cannot criticize that the capture explanation is of the single-mechanism type because it allows one to discriminate between this account and the social power explanation.[23]

Second, the legitimate quest for theoretical precision and leverage is independent of the number of mechanisms that one specifies. In the previous examples, theoretical specificity increased with the number of mechanisms per explanation because more mechanisms simply imply more entities and activities and a finer account of productive continuity. While viable, it is equally possible to specify the two processes in Figures 2.3 and 2.4 and to attach a single mechanism to them. In fact, the analysis of a process involving multiple intervening steps is one key advantage of qualitative case studies because of the opportunity to discriminate between competing hypotheses (Campbell 1975; George and Bennett 2005, chap. 10; Hall 2008). However, nothing mandates that each intervening step and entity and activity is treated as an individual mechanism. The level of detail of a causal process is therefore independent of the number of mechanisms that one infers to be operative.

The different degree of detail that each of the three processes conveys can be captured with Machamer, Darden, and Craver's notions of *mechanism schema* and hierarchies among schemata (2000). A visualization of a process is called a mechanism schema when it abstracts from some entities and activities that ensure productive continuity in fact. For instance,

Figure 2.2 is a mechanism schema that abstracts from the perception of threat on behalf of landowners and therefore is more coarse-grained than the schema in Figure 2.3. The schema in Figure 2.3 is in turn more general than the schema in Figure 2.4. It follows that the three schemata are not contradictory but can be brought into a hierarchy of mechanismic explanations according to their degree of abstraction.

If we now consider the possibility that the three processes in Figures 2.2, 2.3, and 2.4 are schemata of a single mechanism with a different degree of abstraction, it becomes apparent that the three conceptions of mechanismic explanations are very close to one another. That which is a separate mechanism in the triple- and multiple-mechanism explanation is taken as one intervening step and corresponding activity and entity in a single-mechanism explanation. The upshot is that three case study researchers can theorize and find evidence for exactly the same causal process but differ by attaching one, three, or multiple mechanisms to it, conditional on the preferred account of mechanismic explanation.

If all three variants of mechanismic explanations can deal with the same process and entail the same level of detail, what value is there in distinguishing the types of mechanismic accounts at all? Apart from the fact that some scholars make ontological and thus irresolvable claims for one account or the other, different mechanismic explanations entail different theoretical foci (Craver 2006). A single-mechanism explanation highlights one particular factor that links the cause to the outcome. In his analysis, Ziblatt theorizes a capture mechanism and supplements it with arguments as to why it takes place and how it accounts for the effect of landholding inequality on electoral fraud. These supplementary arguments on entities and activities are essential but are likely to be less well perceived and remembered by the reader when compared with the capture mechanism. Consequently, a triple-mechanism explanation is more appropriate when one aims to direct attention to how the cross-case constellation shapes attitudes on the within-case level, how these translate into consequences on the within-case level, and how they are aggregated and prompt effects on the cross-case level. A multiple-mechanisms explanation of a cross-case outcome distributes attention even more broadly than the other types of explanation. Unless one has a strong philosophical predisposition for one of three kinds of mechanismic explanations, one can therefore choose between them on the basis of the theoretical elements that one aims to put to the forefront.

The third implication of the previous discussion of mechanismic explanations is that the three varieties entail very similar implications for the case study method. Three researchers might theorize the same process between cause and effect, such as the one in Figure 2.4, but assign one, three, or more mechanisms to it. In the process-tracing part of the case study, each of the three researchers would be obliged to gather causal process observations for each of the expected intervening steps. They would use the same sources, handle them with the same care, and subsume observations under the same intervening steps. The only difference would be theoretical and rest in the fact that causal inferences are generated for one, three, or more mechanisms. This is an important ramification because it determines what elements of the explanation garner more or less attention, but it is a matter that is unrelated to the case study method serving to build, test, or modify one of the three types of mechanismic explanations.

Drawing together all the previous arguments, I conclude that the various definitions and conceptions of mechanisms and mechanismic explanations do not entail special implications for the case study method beyond the truism that process tracing is the proper technique for general inferences about causal mechanisms. Since the focus of this book is on the case study method, I refrain from adopting a specific definition of causal mechanism here. This might come as a surprise, but ontological debates about what a mechanism is tend to lose sight of the implications of mechanismic analyses for small-n methods (Kurki 2007, 233–4). Discussions about the proper definition of mechanisms suggest more disagreement than actually exists once one examines what a specific definition of mechanisms implies for case study research.[24] In order to avoid the impression that the arguments in the following chapters are accustomed to a specific definition of causal mechanism, I therefore decide against adherence to one particular conceptualization and generally refer to causal mechanisms where appropriate in this book.[25]

2.3 Causes of effects v. effects of causes

Another salient dimension that has to be considered distinguishes between research interested in the causes of effects (CoE) and the effects of causes (EoC), 'effects' here being synonymous with outcomes (Goldthorpe 2000, chap. 1; Mahoney and Goertz 2006; Morton and Williams 2010, 33–41). The two terms refer to what can be called the *causal perspective* that one takes in an empirical analysis. On a general level, a CoE study is centered on the outcome and seeks to discern the relevant causes. In contrast, an EoC analysis is centered on a cause and asks whether it has a (specific) effect on a given outcome. In other words, CoE research is backward looking, while EoC studies are forward looking (taking the direction of causation as the benchmark for defining 'forward' and 'backward'). These commonly held definitions (Mahoney and Goertz 2006) suggest that the two causal perspectives are only about causal effects. This impression would be misleading, though, because there is nothing in the CoE and EoC perspective that inherently prohibits the analysis of causal mechanisms underlying a causal effect.

A more detailed treatment of the two causal perspectives is warranted in order to clarify the scope of my discussion of the case study method. The two terms are frequently invoked in the literature so as to highlight,

Table 2.1 Causal perspectives and theory-centered research goals

		Causal perspective	
		Causes of effects	*Effects of causes*
Theory-centered research goals	*Hypothesis building*	Exploratory case study centered on outcome	Exploratory case study centered on cause
	Hypothesis testing	Test of multiple, complementary hypotheses	Test of single hypothesis
	Hypothesis modifying	Exploratory case study centered on puzzle	–

for instance, the distinguishing features of Comparative Historical Analysis and quantitative research (Mahoney and Terrie 2008) or case-oriented and variable-oriented research (Ragin 1997). The adoption of a regularities perspective might give rise to the belief that my book is only about EoC research. In the remainder of this section, it is shown that this impression would be erroneous because EoC and CoE analyses are covered as long as they are theory centered. In order to achieve this, the distinction between CoE and EoC research is intersected with the distinction between the three types of theory-centered case studies introduced in Chapter 1 (Table 2.1) This intersection produces six combinations of perspectives and research goals, whereas, as I explain below, one of the six does not signify a feasible variant of case study.

Arguably, the most common goal of an EoC analysis is to build or test a hypothesis that stipulates a specific outcome for a given cause. Because of the concern with causes, EoC designs are also called X-centered (Gerring 2001, 137).[26] As an example of hypothesis-testing EoC research, consider the hypothesis that the lobbying efforts of well-organized exporters cause a government to negotiate a liberalizing trade agreement from which the exporters expect to benefit (Dür 2010). The purpose of this study is limited to making an inference about the effect of exporter lobbying. Additional potential causes – such as the state of the domestic economy or the ideology of the government lobbied by the exporters – only need to be taken into consideration if it is necessary to control for their impact in order to generate that inference (see Chapter 4).

In an EoC perspective, it is equally possible to engage in genuine hypothesis building. The phenomenon of primary interest is a cause, and the primary question concerns the nature of that cause's impact. The relevance of a cause can be evaluated only with respect to a given outcome, meaning that it must be specified in an EoC analysis before exploratory process tracing is done.

This then allows one to assess whether the cause is tied to the outcome and to build a hypothesis on the causal mechanism and process.

An example of a hypothesis-building EoC analysis is Kerwer and Teusch's case study on the Europeanization of policy-making activities and the implications for the European nation states (2001). Kerwer and Teusch do an exploratory case study seeking to answer the question of whether Europeanization has an influence on traditional styles of national policy making (such as interventionist and market-oriented policy making). Although this outcome might suggest itself for analysis, a range of other outcomes are also conceivable if one is interested in the effects of Europeanization, for example, the effect of Europeanization on citizen satisfaction with democracy or the European Union. Since the outcome is only of secondary importance in this case study, it is best described as a hypothesis-building study taking an EoC perspective. Hypothesis-modifying case studies are incompatible with an EoC perspective because they are genuinely exploratory and seek to determine the effects of a puzzling outcome.

In contrast to EoC research, a CoE perspective is compatible with all three theory-centered research goals that are on offer. Case studies that seek to build or modify hypotheses have a strong affinity with a CoE perspective because of their exploratory nature. Both types of theory-centered case studies focus on an outcome for which no hypotheses exist or for which sufficiently well-performing propositions are lacking. Since CoE studies are Y-centered (Gerring 2001, 137), their goal is to discern all factors that potentially qualify as causes for the outcome of interest and that need to be subjected to a test in a subsequent hypothesis-testing case study.

For example, a hypothesis-building case study on the determinants of trade liberalization would start with the question 'How can we explain liberalizing trade agreements?' and gather evidence via exploratory process tracing. A hypothesis-modifying case study would start with a puzzle such as 'Why do countries negotiate liberalizing trade agreements when the domestic economy performs poorly?' (which presumes that a bad state of the economy militates trade policy in the direction of protectionism). Both examples are of the CoE type because they center on the outcome and seek to explore potential causes.

The link between exploratory case studies and a CoE perspective is obvious, but this causal perspective is also compatible with small-n research that tests hypotheses. In this instance, one formulates *multiple complementary* hypotheses, each of which is concerned with the effect of a cause on different outcomes that are all closely related to the phenomenon of key interest. On the level of individual hypotheses, one is therefore taking an EoC perspective. However, the multitude of complementary hypotheses that focus on different outcomes implies taking a CoE perspective because they yield a comprehensive picture of the phenomenon under scrutiny.

Such a CoE case study can take two forms that are best illustrated with the previous trade liberalization example. First, one can theorize that the outcome is characterized by equifinality, meaning that multiple causes produce the same outcome (Bennett and Elman 2006, 457). In addition to exporter lobbying, factors capable of causing liberalization are political actors who believe in the welfare effects of reciprocal liberalization or political actors that want to reap the security externalities flowing from liberal trade (Bhagwati 2002). A hypothesis-testing case study therefore is not confined to showing that one factor can cause the outcome but that multiple causes can have the same effect

The second route to a comprehensive test of a hypothesis calls for the analysis of the outcome from different but complementary angles, which is known as testing multiple observable implications that one derives from a theory (Lave and March 1975). The goal is to understand the occurrence of the outcome as well as related aspects such as the timing of its occurrence, the lag between the occurrence of the cause and the outcome, the behavior of different actors in the process leading to the outcome, and so on. The more hypotheses can be confirmed empirically, the more credible the focal hypothesis becomes and the more comprehensive our knowledge about the phenomenon of interest.

As regards the trade example, one could theorize that the lobbying of exporters will prompt counterlobbying by national producers that would be confronted with higher levels of imports if a liberalizing agreement comes into existence (Pahre 2008). Moreover, one could hypothesize that the government, which is responsive to the demands of well-organized groups, compensates domestic producers for the reduction of tariffs, for example, by imposing new regulations favoring national companies and disadvantaging importers. These examples show that each hypothesis has a different outcome; the occurrence of counterlobbying in the first example and the compensation of national producers in the second. But all hypotheses are tied to an explanation of the same overarching phenomenon, namely, the negotiation of liberalizing trade agreements.

The CoE research that is explicitly not covered in this book consists of case-centered case studies. Case-centered studies seek to give a complete explanation of an empirical phenomenon that is substantively important *without* drawing lessons that extend beyond the examined case.[27] The fact that no generalization of causal inferences is intended allows a researcher to incorporate case-specific factors into the explanation that do not extend to other cases *if* the case under analysis is conceived of as an instance of similar phenomena. In comparison with theory-centered CoE analyses, case-centered CoE research thus produces richer explanations that are confined to the examined cases.

The elaboration of theory-centered CoE and EoC case studies now forms the background against which it is possible to delineate the scope of my arguments with respect to approaches and terms commonly invoked in the case study literature. The approaches and terminology I have in mind concern Comparative Historical Analysis (CHA) (Mahoney and Rueschemeyer 2003b) and research anchored in the notions of *causal complexity* and *causal heterogeneity* (Ragin 1987, chap. 2), which is in turn is related to the distinction between *variable-oriented* and *case-oriented research* (Ragin 1987, chaps 3–4; Ragin 1997).[28]

In retrospect, King, Keohane, and Verba's book *Designing Social Inquiry* (1994) represents the impetus for a broad and sophisticated debate about the characteristics of CHA (Mahoney 2004, 2003a; Mahoney and Rueschemeyer 2003a; Rueschemeyer 2003; Skocpol 2003).[29] In short, the key characteristics of CHA are a theoretical anchorage in historical institutionalism (Thelen and Steinmo 1992); the formulation of comprehensive explanations for big and substantively important phenomena, such as the evolution of the nation-state; an emphasis of historically specific (highly context-dependent) causal relationships, which then implies the analysis of relatively few and homogenous cases; comparative case study designs; and a central role of process tracing and temporality in the form of concepts such as path dependence.

Because of the particular engagement with *Designing Social Inquiry*, the usual points of reference for this debate are (mainstream) quantitative methods (Goldstone 1997; Mahoney 2005; Mahoney and Terrie 2008; Mahoney and Villegas 2007; Skocpol 2003). While this contrast proved valuable and highlighted several important features of CHA, it equally holds that CHA and its methods fit squarely into the case study method as it underlies this book. CHA qualifies as CoE research because of the aspiration to formulate comprehensive explanations. At the same time, CHA is theory centered because cases serve to generate explanations that hold within a population of causally homogenous cases (Mahoney and Rueschemeyer 2003a). In practice, it might be that case studies in the field of CHA examine smaller populations than usual in order to do justice to the idea of historical specificity, that most of them prefer the set-relational perspective on causal effects, that they particularly emphasize temporality, and so on (Mahoney 2004). These certainly are important matters of practice and are warranted in order to achieve a fit between historical institutionalism and CHA's inventory of methods (Hall 2003). From the viewpoint of small-n methods, however, CHA's instruments are not qualitatively different from the tool kit available to case study researchers outside of the field of CHA. Because my goal is to provide an integrative discussion of the case study method, I refer to CHA specifically only if the issue at hand is closely tied to the methods that characterize CHA (which holds true for generalization, for example, see Chapter 9).

Two other established distinctions in the social sciences that originate from a struggle with quantitative research contrast *variable-oriented* and *case-oriented* research and emphasize, in relation with the latter, the idea of *causal complexity* and *causal heterogeneity*. The distinction between variable-oriented and case-oriented research mirrors the differentiation between correlational and set-relational research (see below). Among other things, variable-oriented research is usually taken as equivalent to quantitative research interested in marginal effects estimated on the basis of a sample of cases. In contrast, case-oriented research describes cases as wholes in terms of configurations of conditions that lead to the outcome (Ragin 1987, chaps 2–4). In this salient respect at least, this contrast between variable-oriented and case-oriented research is not reflected in my book. As I define it, theory-centered research uses cases instrumentally for the improvement of theory instead of conceiving of theory as instrumental for a comprehensive explanation of cases, which is the characteristic of genuine case-oriented analyses. If case studies are theory centered, they can be variable centered because it is possible to examine correlational causal relationships involving an independent and a dependent variable (see below). However, theory-centered case studies are not necessarily variable centered because they can also concern set relations and causal mechanisms. Because of these ambiguities and in order to clearly denote the scope of the present book, my top-level distinction is only between case-centered and theory-centered small-n research, the latter described by the four dimensions elaborated in Chapter 1.

The contrast of variable-oriented and case-oriented research led Ragin to develop a synthetic strategy, which is better known as Qualitative Comparative Analysis (QCA) (Ragin 1987). Two key elements of QCA are *causal heterogeneity* and *causal complexity*. Causal complexity means that outcomes are produced by configurations of conditions and different configurations can lead to the same outcome (see Section 2.4). The latter aspect is also explained by the notion of causal heterogeneity, which is equivalent to equifinality and the claim that multiple configurations of conditions bring about the same outcome. Both elements, causal heterogeneity and complexity, were originally introduced in order to adhere to the idea of historical specificity, meaning that cases must be analyzed within the historical context in which they are embedded.

Causal heterogeneity is occasionally pitted against the conception of causal homogeneity that underlies quantitative research. The latter is said to unduly favor generality over specificity, which is supported by the fact that quantitative analyses often cover samples or populations that include dozens or hundreds of cases such as all the countries in the world over an extended period of time (Goldstone 2003; Skocpol 2003).

This short discussion reiterates that (mainstream) quantitative research is the point of reference in the debate about QCA and historical specificity. A transfer of the meaning of causal complexity and heterogeneity to the realm of case studies reveals that the former are accommodated by the framework that underlies the present book. Similar to CHA, case studies invoking the notions of causal complexity and causal heterogeneity take a CoE perspective, as the aim is to formulate comprehensive explanations for a given phenomenon. The path to achieving this goal is seen in set-relational causation (Mahoney 2004), the discussion of which forms an essential part of book.[30]

The emphasis of historical specificity manifests itself in the analysis of configurations of conditions and, again as in CHA, the specification of multiple scope conditions that delimit the population to a small and rather homogeneous set of cases (Ragin 2000, chap. 2). Most empirical research involves scope conditions (Bunce 2000; Lieberson 1997, 375; Paige 1999), and the number of boundary statements that one specifies depends on the individual belief about the causal homogeneity of cases in question. This is a matter of practice and not of principle (Ragin 1997), though, and nothing in the preceding and following discussion of the case study method precludes a strong belief in historical specificity.

Finally, the notion of causal heterogeneity might seem to be at odds with the idea of general causation and the generalization of causal inferences on the basis of qualitative case studies. However, this impression would be erroneous. QCA and causal-heterogeneity based research more generally presuppose the analysis of a population of cases (Ragin 2000; chaps 2 and 7; Ragin 1997). The dialogue between ideas and evidence implies that the shape of the population is always preliminary (Ragin 2006a, 635–6), but that there is a population underlying the analysis. Against this backdrop, causal heterogeneity means that, *within* the confines of a population of causally homogenous cases, there are different ways in which the outcome can come about.

The argument that causal heterogeneity and causal homogeneity are not contradictory can be substantiated with a case study on the delay in the transposition of EU directives into national law (the outcome) (see, for example, Steunenberg and Kaeding 2009). In order to keep the example simple, assume you find that a delay is due only to a misfit of the content of the directive and current national law or the opposition of strong lobby groups that benefit from the current law. The two reasons for a delayed transposition clearly are a manifestation of causal heterogeneity.

At the same time, this example points to two elements of causal homogeneity. The first facet of causal homogeneity manifests itself in countries that share the same determinant of a delayed transposition. For example, if Germany, France, Portugal, and Spain are all marked by a misfit and delayed transpositions, the process-tracing insights derived from the analysis of one such country are generalized to the three other countries characterized by the same determinant and outcome (known as contingent generalization, see George and Bennett 2005, 110–13). Secondly, the notion of causal homogeneity extends beyond the cases that share the cause and outcome of interest. Suppose that Germany failed to transpose a directive because of a misfit and Denmark did not transpose a directive in time because of the resistance of lobby groups. Causal homogeneity now means that *if* Germany had been characterized by strong lobby groups instead of a misfit, it too would not have transposed the directive. Similarly, one assumes that Denmark would have failed equally to transpose in time *if* there were a misfit but no opposition by lobby groups. This counterfactual line of reasoning follows from the circumstance that Germany and Denmark are assumed to be causally homogeneous.

Both manifestations of causal homogeneity entail that it does not qualify as the antonym of causal heterogeneity. Instead, causal heterogeneity is better understood as a second-order notion of how outcomes come about because heterogeneity occurs within a set of homogenous cases. It follows that the idea of causal heterogeneity and equifinality are in line with the case study method as it is elaborated in this book. Thus, the discussion of the case study method in the following chapters is also integrative in that it includes small-n research that takes a CoE and EoC perspective in causal analyses and places a premium on causal heterogeneity and causal complexity.

2.4 Types of causal effects: correlations v. set relations

In the past, a lot of the debate on methods centered on the qualitative versus quantitative and large-n versus small-n divide (for example, Ebbinghaus 2005; Mahoney and Goertz 2006; Prakash and Klotz 2007). With this debate still ongoing (Beck 2010; Collier et al. 2010), a salient divide that cuts across these dimensions is between covariational and set-relational cause–effect relationships. Reference to the quantitative versus qualitative and large-n versus small-n divides fails to capture the fact that there are advocates of a covariational framework among case study methodologists (Gerring 2007a). Similarly, the large-n camp is divided into groups of researchers who favor statistical techniques and those who favor QCA, the latter of which is built on the assumption of set-relational causation (see Achen 2005; Ragin 2008).[31]

The remainder of the chapter is devoted to a discussion of the features of correlational and set-relational cross-case relationships that are relevant for the subsequent treatment of the case study method. The next two sections focus on inherent elements of the two types of causal effects and not on how they are handled in practice. This is a point worthy of emphasis because debates about correlational and set-relational causation and methods often center on how they are practiced and not their inherent characteristics. Methods and ontological commitments should not be judged on the basis of how they are practiced but on what they can achieve if properly implemented.[32] In this spirit, I first consider covariational causation and then turn to set relations.

Correlational causation

The core characteristic of a covariational conception of causation can be described as *symmetry*. For a *positive* correlation, symmetry means that an increase in the independent variable prompts an increase in the dependent variable and that a decrease in the independent variable coincides with a decrease of the dependent variable. Correspondingly, a *negative* correlation is symmetric because an increase in the independent variable is followed by a decline in the outcome, and vice versa. Gamson's law (1961) is a good example for a strong positive correlation in the political world. Gamson's law states that there is a linear, one-to-one relationship between a government party's seat share in parliament and its share of cabinet posts. This means that if a government party's seat share increases from 15 to 30 percent, for example, its share of minister posts increases from 15 to 30 percent as well.[33] At the same time, a party that experiences a loss of parliamentary seats from 30 to 15 percent would suffer a corresponding decrease in its cabinet share. The causal effect of the seat share on the cabinet share is symmetric because the outcome increases and declines as the cause increases and declines.

Correlational causation is often equated with linear-additive causality (Ragin 2008, chap. 6). However, neither linearity nor additivity are inherent features of the covariational view on causation but instead refer to its implementation in regression analysis (Beck and Jackman 1998). Additivity captures the assumption that the causal effects of independent variables are independent of each other and that these effects add up. While additive causal effects are mostly associated with regression analysis, in Chapter 5, I show that a specific variant of comparative case study permits examination of them in qualitative small-n research.

Additivity may be a characteristic of empirical correlational research (Mahoney and Goertz 2006), but, in principle, it is of course possible to theorize and to try to infer the presence of an interaction effect. An interaction effect is present when the causal effect of one independent variable hinges on the scores that another independent variable takes. An empirical example involving a discussion of independent and interaction effects in case studies can be found in Prontera's analysis of policy change in the electricity sector of France and Italy (2010). Prontera is interested in the effects of Europeanization on the liberalization (or nonliberalization) of the electricity sector in the two countries. A central factor in the analysis of Europeanization is the misfit between national institutions and policies and those demanded by legislative action of the European Union. In Prontera's case study, a misfit pertains to France and Italy because both maintained state-controlled electricity sectors, whereas the European Union mandated a market-oriented electricity policy.

According to one line of reasoning in Europeanization research, a high level of misfit should hinder the country's compliance with the EU demands because of high adaptation costs. A low level of misfit, on the other hand, should be congruent with compliance. This means that the extent of misfit is expected to correlate with compliance behavior. As regards the electricity policy of France and Italy, one therefore should not observe compliance. The first three columns in Table 2.2 show that the expectation of

Table 2.2 Interaction effects and correlational causation

Case	**Electricity policy (Y)**	**Misfit**	**Institutionalization**
Italy I	State-oriented	Low	Low
Italy II	Market-oriented	High	Low
France I	State-oriented	Low	High
France II	State-oriented	High	High

an independent effect of misfit is not confirmed. Both countries share no misfit prior to EU action because a state-oriented policy was feasible due to a lack of any constraint by the EU (Italy I and France I). After EU legislation came into existence, the level of misfit was high (Italy II and France II). However, only France continued to pursue a state-oriented policy, whereas Italy switched to a market-oriented electricity policy.

In light of this evidence, Prontera argues that misfit does not have an independent effect but depends on the degree to which the national institutions and policies are institutionalized. A change in the misfit influences the national institutions and policies only when their level of institutionalization is low. This holds true for Italy but not for France, because its electricity policy was highly institutionalized. In contrast to misfit alone, an interaction between the level of misfit and the level of institutionalization therefore accounts for the differing trajectories of electricity policy in France and Italy.

Additive causation is one part of linear-additive causality, with linearity being the other one. The link between correlation and linearity is attributable to the common practice of regression analysis to model the causal effect of independent variables as linear (Beck and Jackman 1998). As is discussed in Chapter 5 in more detail, this modeling practice is not relevant for correlational case studies. Linearity refers to the specification of the functional form of a causal effect, requiring the analysis of a sufficiently large number of observations. Case studies emphasize the depth of the within-case analysis at the expense of breadth in terms of the number of cases (Gerring 2004, 347). This, in turn, makes it futile to infer a specific functional form from a handful of cross-case observations. Covariational case studies confine themselves to the detection of a positive or negative correlation among bicategorical or multicategorical causes and outcomes and do well in leaving the test for specific functional forms to a large-n study (see Chapter 5).

Although one does not aspire to analyze the functional form of the causal effect, it is important to note that continuous data can be used in cross-case comparisons for a different purpose. The use of continuous data is related to two different measurement strategies that can be employed in covariational case studies. One can establish a correlation between an independent and a dependent variable, each of which measures either *differences in degree* or *differences in kind*. Table 2.2 depicts correlations between variables taking scores that denote differences in kind. Misfit and institutionalization are either high or low, and electricity policy is either market oriented or state oriented. The alternative is to establish a correlation by measuring differences in degree. In this instance, one country has a higher or lower level of misfit and a higher or lower degree of institutionalization, and the country pursues a more or less state-oriented electricity policy.[34]

Both measurement strategies are feasible in correlational case studies, but they differ fundamentally with respect to the criterion used for the assignment of scores to cases. When the scores of a variable differ in kind, one uses theory and conceptual knowledge to assign values to a case. With respect to misfit, for example, one must specify a benchmark separating high from low levels of misfit.[35] No theory is needed in the analysis of cases that differ in degree because one country simply has a higher level of misfit than another country, a higher degree of institutionalization, and so on.

A look at the empirical small-n literature shows that both measurement strategies are applied. Jakobsen's analysis of the effects of Europeanization and globalization on liberalization in the Danish telecommunications and electricity sector is based on the measurement of differences in degree (2010). He hypothesizes that the degree of liberalization in these sectors is larger the higher the extent of external pressure in terms of globalization and Europeanization. In his case studies, Lange (2009) measures the cause and outcome in terms of differences of kind. He is interested in explaining the level of economic development of former British colonies and distinguishes between a low, medium, and high level of development. His hypothesis is that the extent of development hinges on whether the country was subject to direct or indirect British rule back when the state was a British colony, therefore establishing differences in kind on the independent and the dependent variable. Lieberman's study of AIDS politics is a mix of the two measurement strategies (2009). His goal is to explain differences in degree in the state response to the spread of AIDS. Lieberman hypothesizes that the response of policy makers to AIDS is less aggressive in countries with ethnically divided societies compared with states that have homogenous societies. A difference in kind – ethnically homogenous versus divided societies – is hypothesized to produce differences in degree as regards the level of policy response to AIDS.[36]

These measurement strategies are equally viable and the decision between them should depend on theory. Whenever the theory under analysis is not specific enough to select one of the two measurement strategies, it is up to the case study researcher to choose between them when formulating a hypothesis. This decision has important consequences for covariational case studies to be detailed in later chapters. Without going into the details here, a causal relationship between differences in kind does not necessarily entail a relationship between variables that are measured in terms of differences in degree. In a similar vein, a correlation between differences in kind does not automatically imply a causal relationship between differences in degree. This means that there may be no correlation between differences in kind, while one observes covariance if one measures differences in degree, and vice versa. In making covariational causal inferences, one thus should be particularly careful when the case study indicates that the independent and dependent variable do not covary. All that one should infer is that there is no evidence for a correlation *given* the selected measurement strategy. Whether there is no correlational relationship at all in place can be determined only after having performed a case study that is built on the respective other measurement strategy.

Set-relational causation

Set-relational causation differs from correlational causation in many respects. First of all and as the name suggests, set relations establish relationships between sets in which cases are either members or nonmembers (Ragin 2008, chaps 5–6).[37] Assume one aims to explain why some welfare states are large, making 'large welfare state' the set of interest (also called the outcome set). Presuming that one has some measure for the size of a welfare state – for example, welfare state spending as a share of the GDP – it is possible to assign those states that meet the requirements of a large welfare state to the corresponding set. These cases are members of the set and are also said to be in the set. On the other hand, all countries that do not have a large welfare state do not belong to the set; they are nonmembers and out of the set 'large welfare state'. In other words, they are members of the *negation* of the set of interest, which is a nonlarge welfare state in this example.

The differences between sets and variables are important to understand because they have profound implications for the way in which a case study is conducted. If one specifies the set 'open economy' as a cause for a large welfare state, there is only interest in countries that qualify as members of the set open economy. Variables that correspond to the sets 'open economy' and 'large welfare state' but are not fully equivalent are the level of economic openness and the size of the welfare state. If this is the independent variable in a correlational case study, one scores *all* countries with respect to their degree of economic openness and size of the welfare state. This further entails that a correlational design includes countries with any level of economic openness and relates them to the respective size of the welfare state; the empirical analysis of a set-based case study, meanwhile, only covers states that have an open economy.

The exclusive focus on cases that are in a specific set has three interrelated ramifications that distinguish set-relational case studies from covariational case studies. First, set-relational causation is based on *invariant* cause–effect relationships. This becomes manifest in the set-relational proposition that countries with open economies have a high level of spending. This implies that one is interested only in open economies, and no claim is made about closed economies and small welfare states. Second, invariance implies the idea of *asymmetric causation*, which contrasts with symmetric causality inherent in covariational analyses. The precise manifestation of asymmetry depends on the variant of set relation and is exemplified in the following sections dealing with specific set relations.

Third, instead of speaking of independent variables (or covariates) and dependent variables, as is done in correlational designs, set-relational research is about how a *condition* (X) is related to an *outcome* (Y) (Schneider and Wagemann 2010, 4). This may strike one as a definitional subtlety, but it is important to stay true to the proper terminology of set-relational and correlational research. The term 'condition' signals that causal inferences are about patterns of invariance, that is, for example, whether the welfare state is always (or at least mostly) large in a country that maintains an open economy. In order to emphasize the difference between correlational and set-relational research, I use the effect-specific vocabulary when talking about covariational and set-relational causation. Whenever an argument equally extends to both causal effects, I speak of causes and effects or outcomes because these are neutral terms.

The following sections deal with different varieties of set relations. The discussion begins with sufficiency and necessity as the two cornerstones of set relations. Equifinality, conjunctural causation, as well as INUS and SUIN conditions are more complex types that can be derived from them and are discussed after sufficiency and necessity.

Sufficiency

A condition is sufficient for an outcome when the presence of the condition coincides with the presence of the outcome. In terms of Boolean logic that underlies the analysis of set relations (Ragin 1987, chaps 6–7), sufficiency can be expressed as $X \rightarrow Y$. The arrow signifies the set relation that is in place between X and Y and should not be interpreted as the direction in which the causal relationship works. The arrow means that the set at which the arrow starts – X in the case of sufficiency – is a subset of the set to which it points – Y in a pattern of sufficiency.

A classic example for a pattern of sufficiency is the democratic peace phenomenon (George and Bennett 2005, chap. 2). X denotes the set democratic dyad, in which any pair of countries is either a member or not depending on whether both states in the dyad are democratic. Y represents the set peaceful dyad (or simply peace), which is either present or absent. The argument and empirical finding that two democracies do not fight each other is one of sufficiency: *if* a dyad is democratic (X is present), one observes peace (Y is present). The democratic peace example also demonstrates the asymmetry of set relations. The claim that a democratic dyad maintains peaceful relations does not imply the argument that nondemocratic dyads are always at war. In fact, many nondemocratic dyads are at peace with each other.

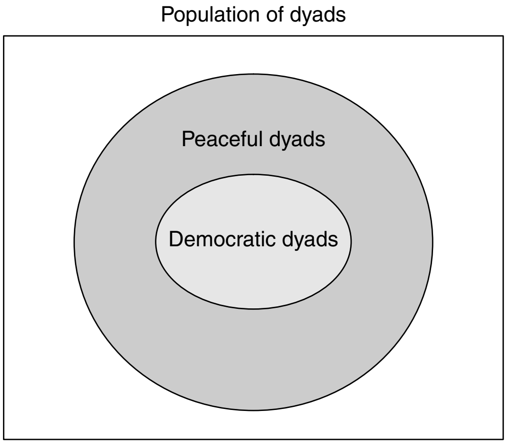

Figure 2.5 Venn diagram for sufficiency

The asymmetry that is inherent in a pattern of sufficiency can be nicely visualized by Venn diagrams, commonly used in combination with set relations. The Venn diagram in Figure 2.5 depicts the democratic peace phenomenon. The rectangle represents all cases in the population. The inner circle shaded in light gray represents the set of cases that are democratic dyads. All cases that are located outside this circle are nondemocratic dyads. The dark-gray outer circle contains all instances of peaceful dyads.

The Venn diagram demonstrates that the set of democratic dyads is a subset of the set of peaceful dyads, which shows that all democratic pairs of countries are at peace with each other. Moreover, the diagram visualizes the asymmetry inherent in sufficient causation because all cases that are outside of the set 'democratic dyad' and in the set 'peace' are instances of peaceful nondemocratic dyads. Thus, the claim that a democratic dyad is sufficient for peace does not mean that a change from a democratic to a nondemocratic dyad alters the outcome from peace to war.

Necessity

A condition is necessary when the outcome occurs *only if* the condition is present. Formally, this means $X \leftarrow Y$. The arrow that signifies the subset relation now points in the direction of the condition because the set of cases with Y present is a subset of the set of cases with the necessary condition in place.[38] The meaning of a necessary condition can be clarified by transforming the democratic peace phenomenon into a nondemocratic war phenomenon. This is achieved by taking war as the outcome and nondemocratic dyad as a necessary condition. Whenever one observes war, it is

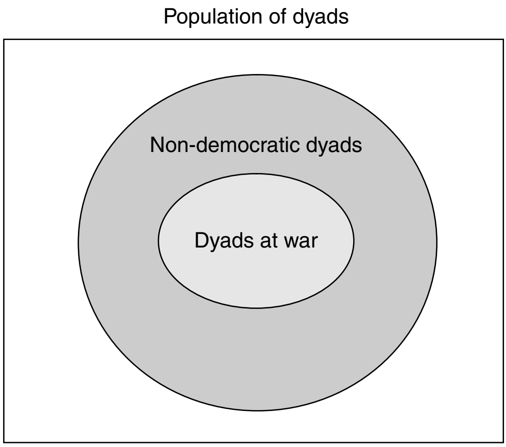

Figure 2.6 Venn diagram for necessity

certain that it is waged by a nondemocratic dyad because we know that two democracies do not fight each other (Goertz and Starr 2003a, 5). The observation that wars are fought between nondemocratic dyads complies with the definition of necessity because a war occurs only if a nondemocratic pair of countries is involved.

The Venn diagram in Figure 2.6 visualizes the argument that a nondemocratic dyad is necessary for war. The circle shaded in dark gray captures the presence of the necessary condition 'nondemocratic dyad'. The light-gray circle signifies the presence of the outcome 'war'. It is a subset of the larger circle and thus points to a pattern of necessity. Furthermore, the Venn diagram highlights the asymmetry characteristic for necessity. The fact that all wars are fought between nondemocratic dyads does not mean that all instances of peace include democratic dyads because the set of peaceful dyads includes democratic and nondemocratic countries alike.[39]

Necessity and sufficiency

A condition is necessary and sufficient *if and only if* the outcome is present in the presence of the condition. When a pair of democratic countries is the condition and the outcome is peace, a democratic dyad would be necessary and sufficient for peace if all such dyads were to maintain peaceful relations (sufficiency) and if all instances of peace were to involve democratic dyads (necessity). Formally, a necessary and sufficient condition is presented as $X \leftrightarrow Y$, where the double-headed arrow denotes that the condition and the outcome form perfectly overlapping sets. This feature of a necessary and sufficient condition also becomes apparent in the Venn diagram presented in Figure 2.7.

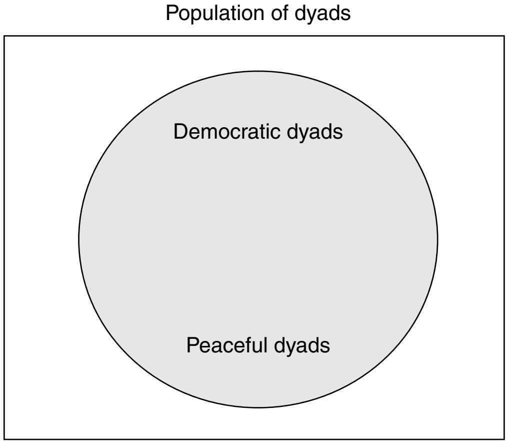

Figure 2.7 Venn diagram for necessity and sufficiency

The Venn diagram highlights that the queen of all set relations qualifies as symmetric because the absence and presence of the outcome depend only on the absence and presence of the condition. When discussing necessary and sufficient conditions, the neat distinction between correlations as instances of symmetric causation and set relations as instances of asymmetric causation breaks down. However, there is a salient difference between a necessary and sufficient condition, on one hand, and an independent variable that is measured in terms of differences in kind, on the other.

Gamson's law (1961) illustrates this difference. Gamson's law states that there is a one-to-one correspondence between a party's share of cabinet posts – the outcome – and the party's share of seats in the parliament – the cause. In a set-relational study, a high seat share is the condition and a low seat share is its negation, and a high cabinet share is the outcome and a low cabinet share is the negated outcome. If a high seat share is necessary and sufficient for a high cabinet share, the corresponding inference reads 'The cabinet share of a party is low when the seat share is low and high when the seat share is high'. The inference implies symmetry because the cabinet share always changes as the seat share changes. A correlational inference relying on differences in kind for the independent and dependent variable would be 'The cabinet share of a party changes from low to high as the party's seat share changes from low to high, and vice versa'. This statement stipulates a symmetric relationship but says nothing about the cabinet share when the seat share is low or high. The correlational inference applies only to a *change* of the share of cabinet posts once the seat share alters. A necessary and sufficient condition therefore implies symmetry, but a genuine correlation relying on differences in kind does not entail anything about necessity and/or sufficiency.

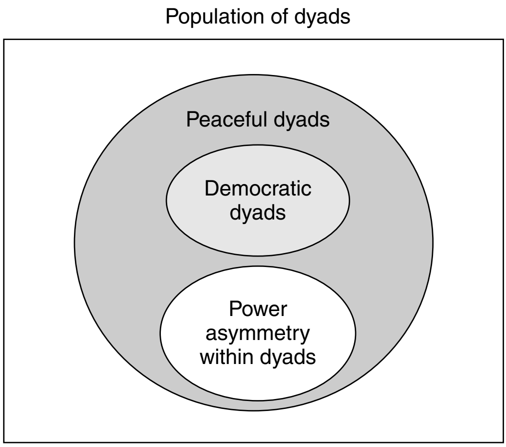

Figure 2.8 Venn diagram for equifinality

Equifinality

Equifinality denotes a constellation where two or more conditions (or combinations of conditions, see below) are individually sufficient for the same outcome. Expanding on the democratic peace example, imagine that a power asymmetry leads to peace within dyads, too. The two conditions 'democratic dyad' and 'power asymmetry' are both individually sufficient for peace because each condition can bring about the outcome. Since the outcome can occur when a democratic dyad is present and power asymmetry absent and vice versa, it follows for equifinality that neither of the conditions is individually necessary.

Formally, equifinality can be written as $X1 + X2 \rightarrow Y$ where X1 and X2 represent two different conditions, such as a democratic dyad and a power asymmetry. The + represents the OR operator in logic and denotes that the presence of either X1 or X2 suffices to bring about the outcome.[40] This characteristic of equifinality is presented by the Venn diagram for the democratic peace example (Figure 2.8). The set of cases with peace present now includes two subsets representing the presence of a democratic dyad and the presence of a power asymmetry.[41]

Conjunctural causation

Conjunctural causation, also known to as configurational causation, means that two or more conditions produce the outcome only if they are simultaneously present. As an illustration, assume one aims to explain welfare state retrenchment. Welfare state retrenchment is due to configurational causation if it results when an economic crisis occurs under the watch of a

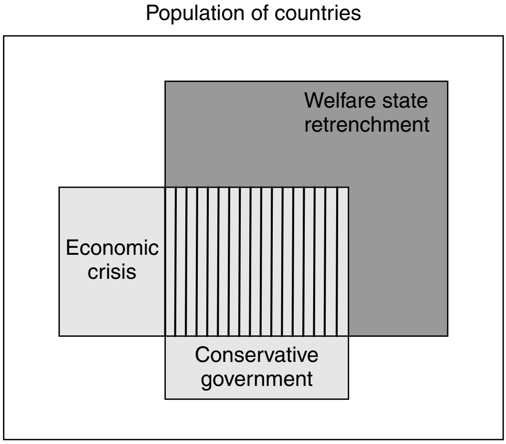

Figure 2.9 Venn diagram for conjunctural causation

conservative government that seizes the opportunity to cut spending. The two conditions economic crisis and conservative government are individually necessary parts of a conjunction and jointly sufficient for the outcome. A crisis does not produce retrenchment on its own, in the same way, a conservative government would have no effect in the absence of a crisis. In terms of logic, configurational causation is presented as X1 * X2 → Y. The * represents the AND operator in logic, denoting that the conditions must occur together in order to produce the outcome.

Figure 2.9 depicts a Venn diagram for configurational causality. For ease of presentation, a set now takes the form of a rectangle. The conditions economic crisis and conservative government are not complete subsets of the outcome set, welfare state retrenchment, which reflects the fact that neither of the two conditions is individually sufficient. However, the shaded area that denotes the conjunction of both conditions is a full subset of the outcome and therefore qualifies the interaction of the two conditions as sufficient.

INUS conditions

INUS causes are in place when at least one conjunction and equifinality come together. In this instance, at least two conditions are *i*nsufficient, but *n*ecessary elements of a conjunction that is *u*nnecessary, but *s*ufficient for the outcome (Mackie 1965). Modifying the previous example, assume that welfare state retrenchment occurs if a conservative government goes along with an economic crisis, or if labor unions are weak in times of a high public deficit because the latter creates pressure for reduced spending that weak

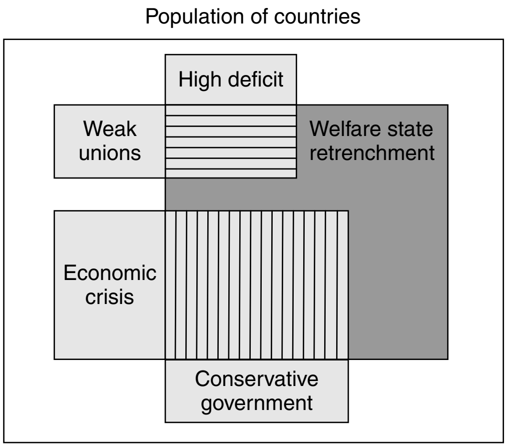

Figure 2.10 Venn diagram for INUS conditions

unions cannot prevent. Thus, welfare state retrenchment is observable when one of two conjunctions is present. The first conjunction includes the conditions 'conservative government' and 'economic crisis', whereas the second conjunction is constituted by the conditions 'weak labor unions' and 'high deficit'. Each of the four conditions is an INUS condition in this example. They are individually insufficient because they require the presence of a second condition in order to take effect. At the same time, they are necessary parts of a conjunction because it is only a configuration that can produce the outcome. While the first two letters in INUS capture the properties of single conditions, the last two letters refer to the properties of a configuration. Each conjunction is unnecessary because the phenomenon is due to equifinality, which further implies that each of them is individually sufficient.

A formula that captures the previous example in abstract terms is $X1 * X2 + X3 * X4 \rightarrow Y$, which highlights the fact that INUS causes require the combination of equifinality and conjunctural causation. Figure 2.10 is a visualization of the previous empirical example. Compared with the diagram for conjunctural causation in Figure 2.9, the Venn diagram for INUS causation simply includes an additional conjunction that is sufficient for the outcome.

SUIN conditions

Recently, SUIN conditions have been proposed as the last missing piece in the set-relational toolbox (Mahoney et al. 2009, 143). An elaboration of SUIN causes requires a move from the level of conditions to the level of attributes or dimensions of individual conditions.[42] A cause is SUIN if it is a *s*ufficient but unnecessary attribute of a condition that is *i*nsufficient

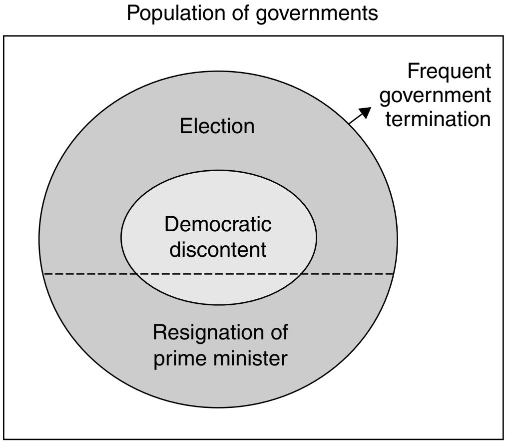

Figure 2.11 Venn diagram for SUIN conditions

but *n*ecessary for the outcome. Take the case of government termination as an example of SUIN causes and assume that 'frequent government termination' is a necessary condition for the outcome 'public discontent with democracy'.[43] Government termination is defined as given when elections are held or when the prime minister resigns. In other words, one speaks of government termination if either of the two attributes holds empirically. In this example, the resignation of the prime minister and an election are SUIN causes because their individual presence is sufficient for counting a government as terminated. Each attribute is not necessary, though, because there are two ways in which termination can result. The first two letters in SUIN therefore refer to conceptual issues and the subsumption of empirical events under the attributes of concepts. The final two letters in SUIN capture the causal quality of the condition that is defined by the attributes. In the hypothetical example, government termination is an insufficient but necessary condition, as it generates the outcome 'public discontent with democracy' only in combination with at least one additional condition.

The presentation of SUIN causes in Venn diagrams is straightforward (Mahoney et al. 2009, 128). Continuing with the example of government termination, the outer circle in Figure 2.11 captures the necessary condition. The visualization of SUIN causes requires the set 'frequent government termination' to be split into two parts. The dashed line separates all cases of government termination into governments that were brought to an end by an election (upper part) and the resignation of the prime minister (lower part). The dashed line runs through the set of the outcome that is a subset of the condition in order to signify that the condition is necessary. Thus, a subset of all cases of frequent government termination resulting from an election are instances of public democratic discontent, while another group of cases in which the outcome is present witnessed the resignation of the prime minister.

2.5 Conclusion

This chapter forms the basis for the subsequent chapters dealing with specific aspects of the case study method. It is worth repeating that I do not argue that everyone should adopt the ontological and epistemological positions discussed in this chapter. Instead, the goal of this book is to discuss the case study method *given that* one believes in empirical regularities. Besides the clarification of some key terms, a major message of this chapter is that this premise extends to the cross-case and within-case level and that there are two different perspectives on what a cross-case regularity is. The correlational and set-relational conception of causal effects rest on different premises as regards the nature of cross-case patterns. The elaboration of the menu of set relations further highlights the variety of cross-case patterns that count as a set relation. The following chapters on case selection and cross-case comparisons will particularly show that the symmetric nature of correlations and asymmetry and variety of set relations has far-reaching ramifications for the case study method and the realization of empirical case studies.

3
Types of Case Studies and Case Selection

The goal of generating statements about empirical regularities implies that a population of cases exists to which these statements extend. Unless one is able to examine all cases in the population, something that might be possible when the population is small, one faces the pertinent challenge of case selection. The systematic choice of cases is crucial because the case is not interesting in itself (at least not in the first place), but for learning something about the population of cases from which it is drawn. One central goal of this chapter is to show that different types of cases lend themselves to this purpose.[1] The types of cases are characterized by different features, each type having different implications for causal inference and calling for different case selection strategies. Figure 3.1 gives a snapshot of the types of case studies and basic selection strategies and additionally demonstrates that distinguishing between the three dimensions introduced in Chapter 1 – research goals, levels of analysis, and variants of causal effects – is central for case studies and case selection.[2]

On the top level, the discussion follows the distinction between the three theory-centered research goals.[3] In each of the three sections dealing with the generation, testing, and refinement of a hypothesis, I first focus on the cross-case level and turn to the within-case level afterward. For each level of analysis, it is further necessary to differentiate between correlational and set-relational propositions, with the latter in turn being subdivided into statements of necessity and sufficiency.[4] For hypothesis-testing case studies, I further describe case selection for the test of a single hypothesis and then extend the discussion to comparative tests of multiple propositions and the incorporation of interaction effects in the case selection process.

An additional source of complexity to consider here is that one can follow two different case selection strategies that are tied to different types of case studies. The first strategy follows the idea of *distribution-based* selection because cases are chosen with respect to their location in a distribution of cases spanned by the cause or causes and/or the outcome of theoretical interest. Depending on the research goal, the three distribution-related types

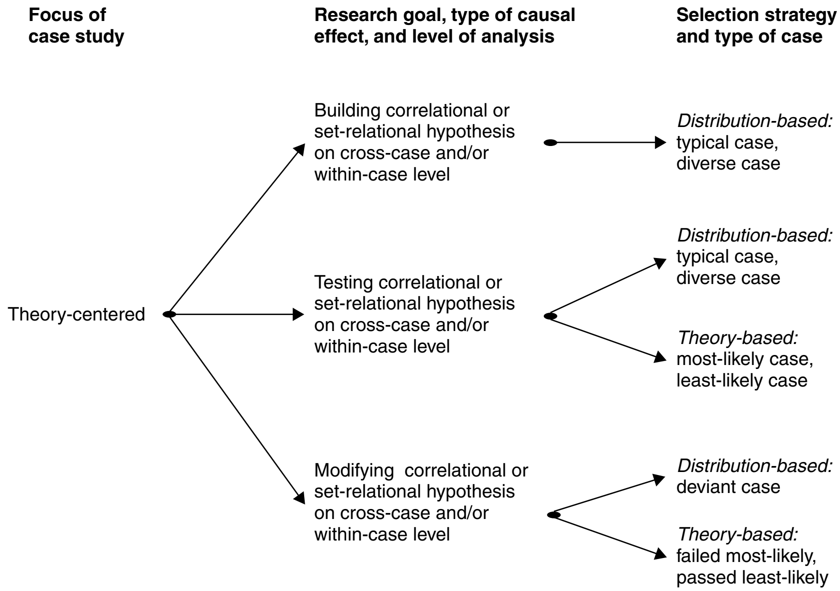

Figure 3.1 Types of case studies and case selection strategies

of case studies are the *typical*, *diverse*, and *deviant* case study. The alternative strategy is *theory-based* case selection, where cases are chosen with an eye on their implications for theory. Again depending on the research goal, this selection strategy includes the *most-likely* and *least-likely* case study and the *failed most-likely* and *passed least-likely* case study.[5] In Chapter 8, it is detailed that distribution-based and theory-based case selection are tied to frequentist and Bayesian modes of causal inferences in tests of hypotheses. In order to maintain the broader scope of this chapter, I postpone a detailed discussion of frequentism and Bayesianism until Chapter 8, continuing with the tenets of distribution-based and theory-based case selection in the present chapter.[6]

Before starting the discussion with hypothesis-building case studies, it is necessary to clarify two additional issues. The first one is related to the question whether one chooses cases from *populations* or *samples*, while the second issue considers the role of different *measurement levels* for case selection. On the first point, I will always speak of case selection on the basis of a population, as opposed to sample-based case selection. The selection principles are the same for case studies that choose their cases from a sample or a population. However, it will become clear that systematic case selection requires knowledge about all cases in the population. When cases are selected on the basis of a sample, an additional element of uncertainty is introduced because one only knows a subset of the population about which causal inferences are made. There are well-known statistical tools for generalizing from samples to the population, but these techniques can hardly be invoked in qualitative case studies because they are not concerned with estimating the size and significance of a causal effect. Moreover, the samples with which we are dealing in macro-research are rarely random samples that would allow for the application of the conventional tools in the first place (Ebbinghaus 2005; Kittel 2006). The bottom line is that one can pick cases from a sample but will fare better by properly delineating a population of cases and by selecting cases from this population (Mahoney and Goertz 2004; Ragin 2000, chap. 2).

The second general issue touches on the link between case selection for the analysis of correlational and set-relational patterns (either as an end or a means) and the measurement level of the outcome and cause(s) forming the basis for the choice of cases. A short discussion of this topic is in order here because the actual choice of a specific type of case hinges on the measurement level in combination with the type of cross-case relationship. Table 3.1 summarizes the interplay of the two aspects.

Qualitative case studies do not estimate causal effects in the way quantitative research does, but continuous variables can play a role for case selection nonetheless. Continuous measurement can form the basis for case selection when one is interested in a *correlation between differences in degree*. As explained in Section 2.4, differences in degree are established by comparing two cases, one of which has a higher score on a variable than the other. For instance, Luxembourg has a higher gross domestic product (GDP) per capita than the Netherlands. From a qualitative point of view, both states would qualify as having a high GDP per capita, but this is irrelevant if the correlation of interest is only about differences in degree and how one case scores relative to another.

Differences in kind are central when a correlation includes nominal and ordinal variables because nominal and ordinal measurement establishes qualitative differences and similarities between cases. The territorial

Table 3.1 Measurement levels and types of causal effects

		Cross-case relationship	
		Correlation	*Set relation*
Measurement level	*Continuous*	Differences in degree	Necessity & sufficiency (continuum within a set)
	Ordinal	Differences in kind	Necessity & sufficiency

organization of the state can be conceived of as a nominal variable when it distinguishes between unitary and federal countries. Each federal (or unitary) state is qualitatively similar to every other federal (or unitary) state, and qualitatively different from every other unitary (or federal) country. The same holds true for ordinal variables, the only difference to nominal variables being that the categories and cases can be ranked. In an analysis of party systems in democracies, one could distinguish between two-party systems and multiparty systems. When the size of the party system in terms of the number of parties – two v. more than two – is the guiding criterion, countries can be ranked according to whether they belong to the class of two-party or multiparty systems. The following discussion of types of case studies demonstrates that the measurement level must be taken into account for case selection in correlational analyses. Since the choice of cases is a little bit more demanding when the underlying measures are continuous, the focus is always on continuous variables first and extended to the simpler scenario of categorical measurement afterward.

Set-relational case studies have a strong affinity with nominal and ordinal measurement because a set is equivalent to one category of a nominal or ordinal variable. For instance, the variable territorial organization of the state can assume the scores 'federal' and 'unitary'. In a case study on necessity or sufficiency, one takes one of the two categories as a set and checks for the presence of set-relational causation. In this view, continuous measurement seems to be at odds with a set-relational analysis because the latter requires the specification of sets. If one variable is continuous, such as GDP per capita, the variable must be calibrated in order to obtain a set such as 'rich country', meaning that continuous data seems to be irrelevant (Ragin 2008, chaps 4–5). However, as indicated by the top-right cell of Table 3.1, the following discussion will show that continuous measurement can play a specific role even in set-relational analyses. This argument is best elaborated in the context of a concrete example and thus is postponed until the following section.

In concluding this discussion, it is to note that the proper measurement level is easy to determine in case studies that test and modify hypotheses. The reason is that the hypothesis under scrutiny, if precisely formulated, conveys information about the measurement level that is entailed. This argument extends to case studies forming a within-case proposition because a cross-case association involving specific measurement scales for the cause and the outcome builds the empirical means for the choice of cases. The following section shows that this is different for the generation of cross-case hypotheses on the determinants of an outcome. Here, one is fundamentally uncertain about whether the causal effect is set-relational or correlational and, concerning the latter, whether it is about a correlation between differences in degree or kind. This makes case selection for genuinely exploratory research on the cross-case level particularly demanding in this respect.

3.1 Types and selection for the formation of hypotheses

The choice of cases for the formation of new hypotheses permits only the *distribution-based* choice of cases. Theory-based selection is not available because this strategy requires theoretical expectations about what one should observe (in tests of hypotheses) or should have observed in the empirical analysis (in modifications of hypotheses) with a certain degree of confidence. Such expectations cannot be derived in genuinely exploratory case studies because they seek to lay the basis for this in the first place. Consequently, this section exclusively centers on the typical case and diverse case as the only two available types in hypothesis-building small-n analyses.

Cross-case level

The formation of cross-case hypotheses might seem to offer a great deal of leeway in the case selection stage because one is performing a genuine exploratory case study that does not draw on an existing body of research.[7] However, the impression that 'anything goes' is incorrect because the degree to which one can make a step forward in the development of theory depends on how cases are selected. In general, though not necessarily (see Section 2.3), hypothesis-building case studies are motivated by the desire to explain a puzzling phenomenon (Mahoney and Goertz 2006). If this research interest leads one to select cases that all share the phenomenon in question in order to search for commonalities, the study is limited to the identification of necessary conditions (Braumoeller and Goertz 2000). Such designs, which can be labeled *no-variance-on-Y designs*, are therefore a good starting point only if the goal is to formulate hypotheses on necessary conditions.

No-variance-on-Y outcome designs face limits if the case study seeks to discern the determinants of an outcome. Besides of the pertinent problem of not knowing whether the latter is attributable to correlational or sufficient causation, the selection of cases that share the outcome is suboptimal. In a correlational perspective, a no-variance-on-Y design allows one to eliminate those variables only as individual causes that vary across the selected cases. Although the elimination of variables reduces the set of potential causes, it seems more straightforward to search for variables that covary with the outcome. In this view, a *variance-on-Y design* is more appropriate because it permits the identification of potential independent variables that could be made subject to a subsequent hypothesis test (in ways described below).[8]

An additional reason speaking for a variance-on-Y comparison is the possibility that the outcome is due to sufficient conditions. A no-variance-on-Y design is inappropriate for this purpose if theory is weak in the sense that one does not know whether the presence or absence of a condition is sufficient for the outcome (see Section 5.2). If one lacks this knowledge, as is likely to be the case in genuine exploratory small-n research, a no-variance-on-Y analysis does not permit it to refute any cause at all as individually sufficient.[9] In order to have some inferential leverage in a hypothesis-building comparison, it is mandatory to create a variance-on-Y design via case selection. The variance on the outcome allows one to reject all invariant causes as individually sufficient because the outcome should always be present when the condition is present. Correspondingly, all varying causes qualify as sufficient because one does not observe the presence of a condition in combination with the absence of the outcome.[10] For these reasons, exploratory comparisons aiming at the determinants of an outcome fare better with a variance-on-Y design than a no-variance-on-Y analysis.

The plea for variance-on-Y designs seems to reproduce well-known arguments for case studies that establish variance on the outcome made by Geddes (1990) and King, Keohane, and Verba (1994, 128–42). However, there are important differences between the two lines of reasoning. The traditional argument for variance on the outcome is based on concerns about the introduction of a selection bias undermining the correct assessment of a correlational causal effect.[11] But a selection bias cannot exist in the first place if there is no variance on the outcome because there is no basis for determining a correlation of any kind (King et al. 1994, 129).[12] Consequently, Geddes and King, Keohane, and Verba argue for variance on the outcome in order to be able to test correlational hypotheses. Besides that I am concerned with the formation of hypotheses, I recommend to achieve variance on the outcome in order to be open for the possibility that the cause–effect relationship is correlational *or* involves a sufficient condition.

Typical case study

In the preceding discussion of case selection, nothing has been said about what type of case to choose in particular. In principle, one can choose between a *typical case study* and a *diverse case study*. A case counts as typical when it is representative in the sense of being able to generalize insights from it to similar cases in the population. This definition does not say anything about the extent to which causal inferences are generalizable. In principle, one can declare any case to be typical, the question always being what the degree of representativeness is in terms of the number of similar cases. Since generalization is the underlying goal of a typical case study, there is a rationale for choosing a case that achieves the *maximum* degree of representativeness given a specific research question, that is, the case that is similar to the largest possible number of other cases in the population.

It is interesting to note that the limitation of generalization to similar cases is in discord with the notion of causal homogeneity, for this would mean that a case would be typical for the entire population. This understanding of causal homogeneity would make case selection and generalization very easy because every case in the population would be equally suitable. For two reasons, however, this is not the common understanding of similarity in typical case studies. First, it is likely that some cases do not conform to the cross-case (or within-case) pattern displayed by the typical cases (see below). Second, depending on how the cases are distributed in the population, some cases are deemed to be too dissimilar in comparison with the typical case in question. A discussion of case selection by Haggard and Kaufman (2008, 19) is exemplary for these concerns. Haggard and Kaufman are interested in the link between economic crisis and reforms on the trajectory of social policy in new democracies. They consider choosing typical cases because of their potential to generalize to comparable cases, but decide against it because of worries that the selected cases may not be fully representative for all the countries in the population.

Although it creates some friction with the meaning of causal homogeneity, there are legitimate reasons to take a cautious approach to generalization and put some limits on the degree to which a typical case is representative. The delineation of cases for which a typical case is or is not representative lacks a hard and fast rule, and so it is up to the researcher to make transparent what he understands to be similar and dissimilar cases in his typical case study.

It was explained before that the formation of hypotheses on the determinants of an outcome should follow a variance-on-Y design. This highlights that the choice of typical cases and the specification of similar cases entirely depend on the distribution of cases across the outcome. Because of the salient role of how the cases are distributed, it now becomes apparent why the choice of typical cases follows the broader strategy of *distribution-based case selection*. The elaboration of distribution-based selection requires us to follow the distinction between case studies searching for necessary conditions, on one hand, and the determinants of an outcome, on the other.

In necessary condition case studies, the choice of typical cases and delineation of similar cases is straightforward because one is interested in the presence of an outcome that is measured in categorical terms. All cases that are instances of the outcome are qualitatively identical to the selected typical case. It follows that every case that displays the outcome is representative of every other case that also has the outcome in place (the same holds true for the negation of the outcome). For example, if we believe that strong unions are a necessary condition for a large welfare state, we would generalize this insight to all countries with high spending because they are qualitatively similar to the examined case.

The identification of a typical case tends to be more protracted when one wants to learn something about potential determinants of an outcome. Building on what was said in the introduction on levels of measurement, the focus is first on an outcome that is measured *continuously*. When the outcome is measured continuously, one must take into account that the causal relationship could involve a correlation between differences in degree and differences in kind and that the outcome could be brought about by a sufficient condition.

In light of these possibilities and the need to select cases with variance on the outcome, the choice of typical cases should consider three criteria. First, with an eye on discerning potential sufficient conditions, one case must be a member of the outcome and one case a nonmember (implying that one has to decide about a threshold that divides the continuous variables into two sets). Second, as regards the possibility of a correlation between differences in kind, the case selection principle is the same as for sufficiency because one should choose one case having the outcome present and one case having the outcome absent. Since we are talking about qualitative similarities between cases that belong to the same set, it again holds that any case that belongs to a certain category is typical for any other case from the same category.

The third case selection criterion aims at the possibility that the outcome is due to a correlation between differences in degree. In order to discern such a causal relationship, the difference of the selected cases' score on the outcome should be large. A large difference on the outcome refutes the criticism that it is due to chance and thus not worthy of an explanation. In addition, a huge difference on the outcome promotes the search for an independent variable as the two cases are likely to display a discernible difference here as well. However, a huge difference does not suffice for the choice of typical cases. Since cases do not belong to categories anymore and are no longer qualitatively similar, it becomes essential to identify the group of cases for which a typical case can be claimed to be similar on the continuous outcome. Recalling that generalizability is the rationale for a typical case study, the goal must be to choose two cases that differ widely on the outcome and that are similar to the maximum number of cases with respect to the scores on the outcome.

As an illustration of the three criteria, suppose you are interested in the determinants of the average of welfare state spending of 21 OECD countries in the period between 1997 and 2001 (Figure 3.2). In a correlational view that focuses on differences in degree, the choice of cases is based on the continuous outcome. With respect to a correlation between differences in kind and sufficient causation, it is necessary to transform the continuous outcome into a categorical one. For the data in Figure 3.2, we assume that the threshold separating large from small welfare states lies at 30 percent of the social expenditure. All countries with higher spending belong to the category of countries with large welfare states, whereas countries with lower spending levels are instances of small welfare states.

The application of the three above-mentioned case selection criteria to the distribution of cases in Figure 3.2 indicates Spain and Sweden as a suitable pair for analysis; Sweden and Spain are instances of a large and small welfare

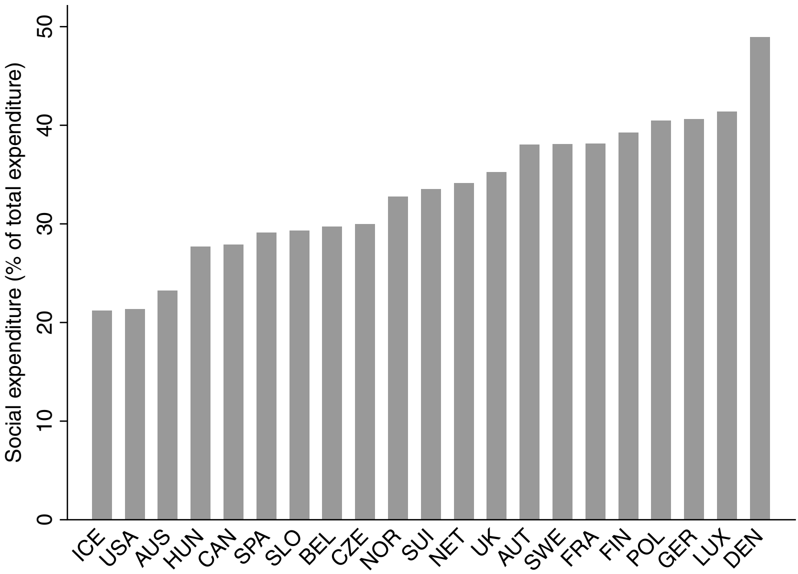

Figure 3.2 Case selection on the basis of continuous outcome

state; Sweden's spending is considerably higher than that of Spain (about ten percentage points); it can be argued that Spain's spending level is similar to those of five other countries (Hungary, Canada, Slovenia, Belgium, and the Czech Republic) and that Sweden is similar to six other countries as regards the level of spending (Austria, France, Finland, Poland, Germany, and Luxembourg). The choice of Spain and Sweden thus makes it possible to perform an exploratory analysis on whether the outcome is attributable to a correlation between differences in kind, differences in degree, or sufficient causation (leaving it for Chapter 4 to discuss whether this is an easy endeavor).

Although Sweden and Spain cover a good proportion of the population, there are some cases that are dissimilar to both. For example, one would refrain from generalizing the insights to Australia, Iceland, the United States, and Denmark, countries that clearly stand apart from the rest. It has been indicated above that the notion of a typical case and maximum generalizability is flexible and depends on what one wants to find out. If one has no idea about factors that could produce the outcome, the recommendation is to follow the strategy in the empirical example and choose two cases that achieve maximum generalizability as regards the entire population. Once solid knowledge about the majority of cases in the population has been gathered, it is worthwhile to expand the perspective and look at cases not covered by the generalization of inferences so far.

In the empirical example, this particularly pertains to the three countries with the lowest levels of spending. To find out whether the inferences about countries from the middle of the distribution extend to the low-spending countries, one would have to select one country with moderate levels of expenditure, such as Sweden, and one of the three states from the lower end of the distribution, such as the United States. Compared with the analysis of Spain and Sweden, the scope of generalization is smaller because the United States is similar only to Iceland and Australia. However, it is still a typical case study with respect to the goal of finding out whether the inferences that hold true for countries with moderate spending levels travel to countries with low spending. In total, the empirical example demonstrates that the engagement with one research question – what accounts for social expenditure? – may demand multiple typical case studies, depending on how the cases are distributed across the outcome and the feasibility of generalization.

Diverse case study

The diverse case study is an alternative to the typical case study for the generation of cross-case hypotheses. On a general level, the ideal diverse case study involves two cases that span the entire range of scores on the cause and/or the outcome, depending on the research goal and the level of analysis (Seawright and Gerring 2008, 300–1). The importance of covering a range of scores on a cause and/or outcome implies that diverse case studies are feasible when they are measured continuously or multi-ordinally.[13] Much as for a typical case study, the main rationale for a diverse case study is tied to a specific strategy for the generalization of causal inferences. The selection of the two extreme cases is based on the premise that the insights derived from diverse cases can be generalized to all cases located between them. The larger the range that the two diverse cases span, the more contestable this assumption becomes because the cases become more and more dissimilar. However, when one does not share this concern and believes in generalizability from the diverse cases to the more moderate ones, one enjoys the advantage that diverse cases tend to be easier to determine than typical cases.

Applying the diverse case study to the specific goal of generating cross-case propositions and the empirical example, there is only difference between this type of case study and the analysis of typical cases. With respect to the possibility of a correlation between *differences in degree*, one now chooses the country with the lowest and the one with the highest level of social expenditure: Iceland and Denmark, respectively. Provided that one could gather within-case evidence for a correlation between differences in degree, the corresponding inferences are generalized to the other 19 countries in the population.

The hunt for diverse cases on a continuous outcome is unrelated to the selection strategy for the possibility of sufficient causation and a correlation

Table 3.2 Types and selection for formation of cross-case hypotheses

Basis of case selection	**Case label**	**Case selection principle**	
		Correlation	*Set relation*
			Necessity: No variance on Y
Distribution	Typical	*Difference in degree, kind, and sufficiency:* Two cases from different categories with sufficient variance on Y and maximum similarity to other cases	
	Diverse	*Difference in degree, kind, and sufficiency:* Two cases from different categories with maximum variance on Y	

between differences in kind because here cases are chosen on the basis of categorical measures. This means that one is still choosing typical cases when selecting one case from each of the two categories on which the outcome is measured. As regards the formation of cross-case hypotheses, it is thus not entirely correct to speak of *the* diverse case study. Because of the need to simultaneously check for both the presence of sufficiency and for a correlation between differences in kind and differences in degree, the features of a typical case study are blended with those of a diverse case study.

The arguments on case selection for the generation of cross-case hypotheses are summarized in Table 3.2. In this table and all the following ones, the column 'case selection principle' distinguishes between the choice of cases for necessary-condition and sufficient-condition case studies. If no such distinction is made, the two variants follow the same logic of case selection. Similarly, the tables distinguish between correlations between differences in degree and differences in kind where necessary.

Within-case level

For the formation of within-case hypotheses, again, one only has the choice between the typical and the diverse case study. However, case selection follows different principles compared with the generation of cross-case propositions because one now has a cross-case basis from which to choose cases for process tracing.

Typical case study

Generally, a typical case study has to meet two criteria. First, the typical case displays the scores on the cause *and* the outcome that are in line with

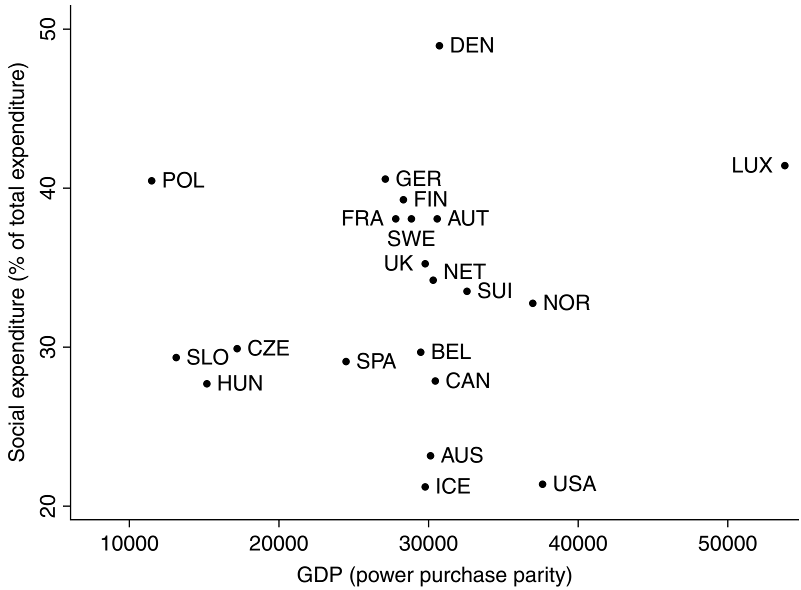

Figure 3.3 Case selection on the basis of two continuous variables

the cross-case relationship, for which a complementary within-case hypothesis should be formulated. This is a major difference from the generation of cross-case hypotheses, where case selection is limited to the outcome.[14] The second criterion has already been introduced above and pertains to the choice of cases that achieve the maximum degree of generalizability. Following what has already been said about the use of continuous and categorical measures, this is more demanding for case studies seeking to explain a correlation between differences in degree.

In a case study that aims to explain a correlation between *differences in degree*, a typical case should achieve the maximum degree of generalizability in the two-dimensional space that is spanned by the independent and the dependent variable. In order to illustrate the choice of typical cases in a two-dimensional space, suppose that you are interested in the relationship between the wealth of a country – the GDP per capita measured in terms of power purchasing parity – and its social expenditure, which is the same outcome as in the previous section. The distribution of cases in the two-dimensional space is presented in Figure 3.3. Further assume that the correlation between GDP and social expenditure is taken to be positive and that exploratory process tracing serves to formulate a hypothesis explaining this pattern.

In order to meet the first criterion for systematic case selection, the two cases must establish a positive correlation because this is the cross-case pattern that one aims to underpin with a within-case hypothesis. Given this constraint, the second criterion requires that the two cases should be similar to the largest possible number of other cases. A look at the distribution in Figure 3.3 indicates that a suitable typical case study could include Hungary and the United Kingdom.

The two countries establish a positive correlation because the United Kingdom has a higher GDP and a higher level of social expenditure than Hungary. At the same time, this pair of countries maximizes the number of countries to which the insights could be generalized. Hungary can be argued to be similar to the Czech Republic and Slovenia, while the United Kingdom is closely located to the Netherlands, Switzerland, Austria, France, Sweden, and Finland (perhaps even Germany). In total, this typical case study would allow generalization to eight out of 21 countries. This may not strike one as much, but Figure 3.3 shows that there are no better cases available for a typical case study.

Case selection for exploratory within-case analyses is more straightforward when the causes and the outcome are measured *categorically*. This applies to the measurement of differences in kind in covariational case studies as well as set-relational research and is due to the fact that cases from the same category are typical for all other cases from the same category. Assume that one is interested in the same causal relationship as in the previous example – the effect of wealth on social expenditure – but that one now distinguishes only between a low and high GDP (benchmark is US$20,000 per capita) and low and high social expenditure (benchmark is 30 percent of social expenditure relative to the GDP). Presuming that one is interested in a *correlation between differences in kind*, the basis for process tracing is laid by selecting a country with a low GDP and level of expenditure, on the one hand, and a high GDP and level of expenditure, on the other. Owing to the circumstance that all cases belonging to the same category on the cause and outcome are qualitatively similar, one can choose any pair of cases that establishes a positive correlation for process tracing.

Case studies anchored in the *set-relational view* on causation follow the same two principles to the extent that one should select cases with knowledge of their scores on the cause and the outcome. Given that set-relational analyses rely on categorical measurement, a maximum degree of similarity is automatically achieved by intentionally selecting cases that are in accord with the set relation of interest. Typical case studies on sufficiency and necessity actually aim for the same cross-case pattern because the best possible context for process tracing is achieved by choosing cases that have the condition *and* the outcome present.

However, the focus of the within-case analysis depends on the type of set relation of interest. For necessity, the question is to find out why the absence of the purported necessary cause would produce the absence of the outcome as well (Goertz and Levy 2007). For instance, Pahre (2001, 2008, chap. 12) hypothesizes that most favored nation (MFN) treatment, which mandates nondiscrimination between importers, is a necessary condition for a country's decision to hold its bilateral trade negotiations simultaneously (clustered) instead of sequentially. A within-case analysis of this hypothesis would have to search for process tracing evidence, for example, statements of chief negotiators explaining why the bargains would not have been clustered if cooperation had not been based on MFN treatment.

In a case study on sufficiency, in contrast, it must be assessed why the presence of the cause prompts the outcome. Simplifying his argument, Dür (2007b, 463) hypothesizes that the mobilization of exporters makes the government of the country in which the exporters are located accept a liberalizing trade agreement with a foreign country. The underlying determinants of this pattern can be discerned in exploratory process tracing by collecting statements of officials and politicians indicating why they have to engage in reciprocal liberalization whenever they are confronted with the lobbying of exporters. Although the perspective of process tracing is different in the two typical case studies, they have a uniform principle of case selection: choosing cases with the condition and the outcome in place.

Diverse case study

The diverse case study is the second available type of case study for the formation of within-case hypotheses. In a diverse case study, the selected cases span the maximum range of scores on the underlying cause (or causes) and the outcome. In contrast to the cross-case level, the diverse case study is suitable for the formation of hypotheses substantiating a correlational and a set-relational pattern alike. More precisely, the following discussion shows that the function of a diverse case study is slightly different for correlations between differences in degree, on the one hand, and correlations between differences in kind and set-relational research, on the other.

When exploratory process tracing serves to find an explanation for a *correlation between differences in degree*, the ideal diverse cases span the maximum range of scores on the independent *and* the dependent variable.[15] The rationale for the diverse case study again is that the generated insights can be generalized to all other cases that are embraced by the two extreme cases. Although it is conceivable that two cases establish the required correlation and occupy the extreme ends of the distribution on two variables, this is not necessarily so. Taking the distribution of cases in Figure 3.3 as the basis for case selection, it can be easily seen that the ideal diverse case study cannot be realized. Iceland and Denmark are the diverse cases for the dependent variable but have almost identical scores on the independent variable. Luxembourg and Poland, on the other hand, are the diverse cases on the independent variable and display very similar scores on the outcome.

In the empirical example, one confronts a trade-off between the maximization of the range of scores on the cause and the outcome. A feasible diverse case study that spans almost the entire range of scores on the GDP includes Luxembourg and Slovenia because the GDP of Slovenia and Poland is almost identical. However, six countries spend less on social matters than Slovenia, and one country spends more than Luxembourg. With respect to social expenditure, this case study therefore suffers from less than optimal generalizability because some cases in the population are not embraced by the two selected cases.

The notion of a diverse case study does not seem to be compatible with the measurement of *differences in kind* because diverse cases, as defined above, can be selected only for multi-ordinal measures. However, a diverse case study can also play a role in research relying on differences in kind but for slightly different reasons than in case studies invoking differences in degree. A diverse case study is feasible under two conditions: first, a continuous variable needs to be transformed into a categorical measure (see Ragin 2008, chaps 4–5) and, second, one wants to determine the validity of the thresholds that create the categories.[16]

The two requirements can be illustrated with a set-relational case study on sufficiency, noting that all arguments extend inquiries into necessity and research on correlations between differences in kind. Imagine one is interested in whether a high GDP per capita is sufficient for a large welfare state. Countries with a GDP of US$20,000 per capita or higher are taken as rich countries and countries with 30 percent social expenditure or more as large welfare states. In a strict set-relational perspective, any country that is in the set 'rich country' could be chosen to test for the presence of high spending. However, one might be uncertain about whether US$20,000 of GDP per capita is an appropriate cutoff point.

A diverse case study can be realized if one wants to make the thresholds subject to an empirical analysis. The threshold can be checked for its validity by selecting two cases that belong to the same category but that differ widely in their scores on the underlying continuous variable. One case should assume the highest score on the variable and thus be the highest-ranking member of the corresponding set. The other case should take the lowest possible score that still qualifies as a just-so member of the set. In the given empirical example, this would mean choosing two countries that are full members and just-so members of the sets 'rich country' and 'large welfare state' alike. If exploratory process tracing in the full and just-so members shows that the same mechanisms and processes underlies the link between a high GDP and high social expenditure, one can be more certain than previously that the selected countries are members of the same sets and that a high GDP is sufficient for a large welfare state.[17]

In order to illustrate the diverse case study with an empirical example, the two diverse countries within the set 'rich country' are Luxembourg (about US$54,000 per capita) and Greece (slightly more than US$20,000 per capita). For the set 'large welfare state', the diverse countries are Denmark (about 49 percent social expenditure) and Norway (about 33 percent social expenditure). The fact that the diverse cases within each set are not identical underscores the problem of achieving the ideal diverse case study when the two cases must occupy extreme positions on the outcome and one cause (or multiple causes). In this example, this problem is manageable as regards Luxembourg because it has the second highest level of expenditure and thus is close to being an extreme case in both sets. However, Greece does not belong to the set 'large welfare state' at all and ceases to be an object for process tracing. Norway, on the other hand, ranks third in the set 'rich country', forcing the case study researcher to look for another case that necessarily fails to be extreme in either of the two sets.

The example seems to indicate that the diverse case study in set-relational research differs from its correlational counterpart. The latter is linked to a specific generalization strategy, while the former invokes diverse cases for checking the validity of thresholds. Indirectly, however, the empirical analysis of the cutoff points is related to generalization because the thresholds decide what cases belong to a certain category and are covered by the generalization of causal inferences. In this view, diverse case studies follow a similar rationale in set-relational and correlational case studies.

In total, it is interesting to note an inverse relationship between the initial confidence in the generalizability of typical case studies and diverse studies anchored in the correlational and set-relational framework. In covariational research, typical case studies limit the extension of causal inferences to similar cases, whereas the diverse case variant makes a sweeping generalization to all other cases. In a set-relational framework, on the other hand, a typical case is expected to be representative of all other cases that belong to the set. Here, it is the diverse case study that takes a more cautionary approach because it initially does not fully trust the set memberships of cases and selects two members of the set of interest on the basis of the underlying variable instead. Table 3.3 concludes the section by summarizing the key features of hypothesis-building case studies on the within-case level.

3.2 Types and selection for the test of hypotheses

In a test of a hypothesis – called working hypothesis in the following – it is good practice to think about a *rival hypothesis*. The competing hypothesis can stipulate a cause other than the working proposition or specify an interaction effect of which the cause in the working hypothesis is a constitutive element. In the following sections on tests of cross-case and within-case hypotheses, the focus is first on the simplest scenario of a test of a single hypothesis including one cause. The discussion then turns to the slightly more complicated analyses that take interaction effects and rival causes into account.

Table 3.3 Types and selection for formation of within-case hypotheses

Basis of case selection	**Case label**	**Case selection principle**	
		Correlation	*Set relation*
Distribution	Typical	*Difference in degree:* expected covariance with maximum similarity to other cases *Difference in kind:* expected covariance and one case from each category	Case with X and Y present
	Diverse	*Difference in degree:* expected covariance and cases are most extreme on X and Y *Difference in kind:* expected covariance and one case from the two most extreme categories	Cases are just so members and full members of X and Y

Cross-case level

The major difference between types of case studies and case selection for the formation and test of cross-case hypotheses lies in the stronger theoretical basis that one necessarily has in confirmatory analyses (see Section 1.3). The formulation of theoretical expectations prior to the empirical analysis renders the choice of cases more straightforward and opens the opportunity for the theory-based choice of cases and realization of a broader repertoire of types of case studies.

Typical case study

In distribution-based selection, a case generally qualifies as typical when it displays the theorized cross-case scores on the cause *or* the outcome. The one-sided choice of cases is, of course, mandatory because the rationale for a cross-case test is to test for specific scores on the outcome given the scores on the cause, or vice versa. In correlational case studies, the choice of a case that displays the variation of interest either on the cause or the outcome allows one to test for the predicted *co*variation with the outcome or the cause (King et al. 1994, 140–42).

Imagine that the aim is to test the proposition that increasing economic openness leads to a decrease in welfare state spending (using the same data as in the example above, see Figure 3.4). This proposition establishes a correlation between differences in degree and is known as the efficiency hypothesis

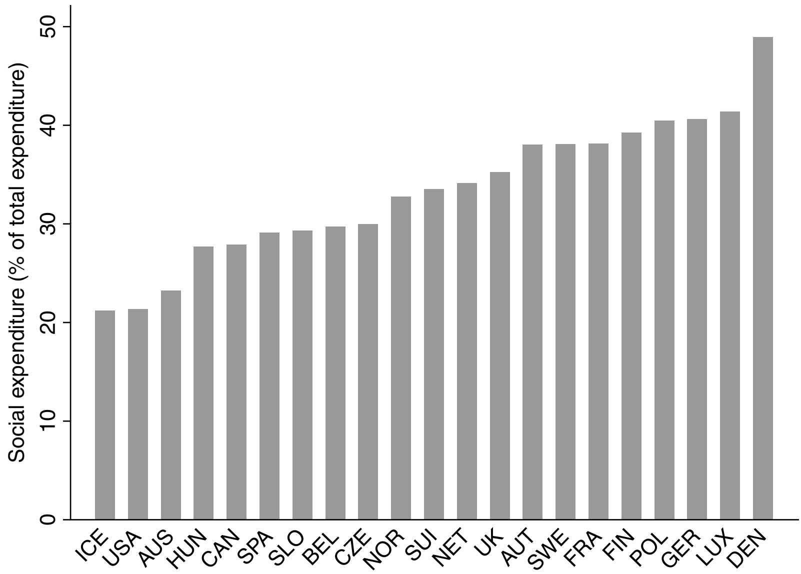

Figure 3.4 Case selection on the basis of a continuous variable

(Obinger et al. 2010, chap. 1). The formulation of this proposition makes the choice of typical cases easier compared with the exploratory formation of a cross-case hypothesis. The reason is that one needs only to be sure that the two cases differ strongly and achieve a maximum degree of generalizability on the continuous variable underlying case selection (social expenditure in this example). There is no need anymore to also take into account that the causal relationship might be a correlation between differences in kind or a set relation. Following the principles laid out above, it could be argued that Belgium and France are the best typical case study available. To give another example, Spain and Sweden would be suitable as well.

Given the distribution of cases in Figure 3.4, one could generalize the result of the test to 11 other countries. If one is prepared to argue that the spending levels of Norway, Switzerland, the Netherlands, and the United Kingdom are also similar to the expenditure of Belgium or France, it would be feasible to generalize to 15 out of 19 countries, as Iceland, the United States, Australia, and Denmark would be the only countries that stand apart.

The choice of typical cases is more straightforward when the causes are measured categorically in a test of a covariational hypothesis that includes differences in *kind* and set-relational analyses. The basic logic of case selection on the basis of categorical causes and outcomes has been discussed in detail in the preceding section, so I leave it with a short note here. The identification of a typical case is easy when cases belong to categories because all cases from the same category are qualitatively identical. Imagine the efficiency hypothesis transformed into the statement that when economic openness changes from low to high levels, spending changes from high to low levels. Any pair of countries that establishes a correlation between differences in kind on the independent or the dependent variable is suitable for testing this proposition. The choice of typical cases is even more straightforward for set-relational research. Here, we only need to ensure that we select one case that is a member of condition in analyses of sufficiency so as to test for the presence of the outcome, or that is a member of the outcome in a case study on necessity in order to test for the presence of the condition.

Typical case study and rival hypotheses

Case selection for tests of hypotheses is not as easy as it may have seemed in the previous paragraphs because competing hypotheses have been ignored so far. In general, a rival hypothesis can state an interaction effect, as opposed to an individual effect of a cause, or stipulate an entirely different cause than the one covered by the working hypothesis. Though these two issues can be distinguished analytically, the following discussion shows that they have similar implications in practice. For ease of presentation, the focus is therefore on interacting causes and only refers to rival causes where appropriate.

I return to the example of economic openness and welfare state spending in order to demonstrate that interaction effects should be taken into account in the case selection process. Elaborating the efficiency hypothesis stating a correlation between *differences in degree*, one could claim that the unemployment rate is important because it moderates the influence of openness on welfare state spending.[18] When one ignores the unemployment rate in case selection, it may be that the countries chosen with regard to spending only no longer achieve the optimal degree of representativeness. Figure 3.5, which plots a country's level of social expenditure against its unemployment rate, underscores this point.

In the previous example of unidimensional case selection centered on the outcome, Belgium and France constituted a suitable typical case study for testing the effect of globalization. Figure 3.5 shows that this pair of countries is not the best choice anymore once unemployment is in the picture. Two reasons account for this. First, France and Belgium have a different unemployment rate and a different level of spending in the two-dimensional distribution. As is discussed in Chapter 4 in detail, this creates inferential problems. When globalization correlates with the outcome as expected, it becomes impossible to distinguish between the effect of openness and unemployment on social spending. Second, France and Belgium remain typical for other cases, but generalizability is no longer maximum. Belgium forms a cluster with three other countries, but France is located

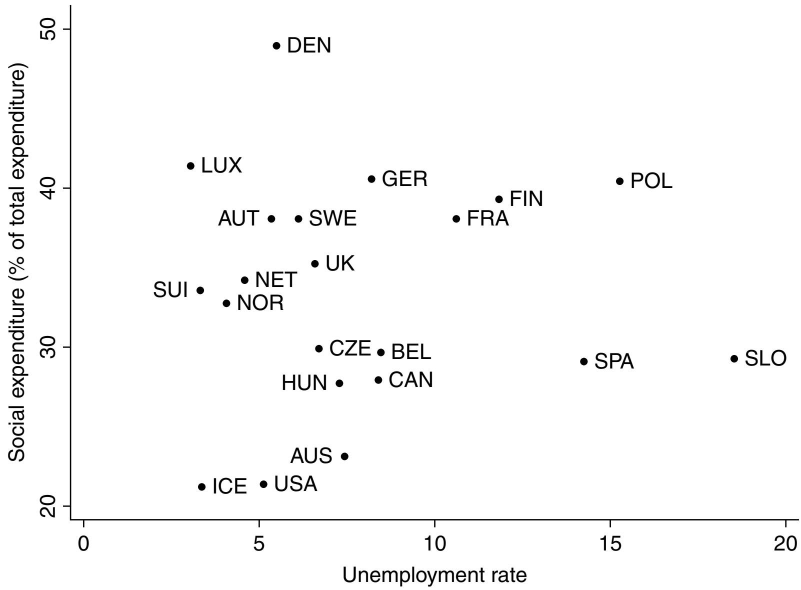

Figure 3.5 Case selection on the basis of two continuous variables for test of an interaction effect

only close to Finland. In one-dimensional case selection, the two countries could be argued to cover 11 or even 15 other states, whereas generalization is limited to four countries in two-dimensional space. A diminished opportunity for generalization is not surprising because the more causes are added to the analysis, the more likely it is that countries that are similar in lower-dimensional space turn out to be dissimilar on at least one cause in a higher-dimensional space.

A look at Figure 3.5 shows that the salient criteria, the control for third variables and the maximization of generalizability, are better achieved by a pair of countries other than France and Belgium. For example, it would be possible to choose the United States and the Netherlands. They have similar levels of unemployment and very dissimilar levels of social spending.[19] The United States is closely situated to Iceland and Australia, while the Netherlands is situated near Austria, Norway, Switzerland, and the United Kingdom. The gain in generalizability is not large, but it is larger and apparently maximal, given the distribution of cases in two-dimensional space.

The arguments on case selection on the basis of continuous variables can be fully extended to case studies involving categorical outcomes and causes. The two criteria to be regarded in the analysis of rival causes and interactions are easier to meet owing to the nature of categorical measurement. The choice of typical cases and identification of similar cases is easy when the

Table 3.4 Case selection on the basis of a 2 × 2 table

		Regime type	
		Unitary	*Federal*
Social spending (Y)	*Low*	9	3
	High	5	4

case study involves categorical measures because all cases belonging to the same category are qualitatively identical. For a similar reason, the control of third causes is more straightforward. Instead of finding two cases that assume similar causes on a continuous factor, it suffices to identify two cases that belong to the same category on the cause one needs to control for.

These arguments can be illustrated by taking a qualitative perspective on social spending and only distinguishing between a high and a low level (the benchmark is set at 30 percent). Theoretically, we are still interested in a correlational relationship but one involving *differences in kind*. In addition, we now take into account that the regime type in terms of federalism and unitarism might condition the effect of economic openness. The argument could be that decision-making processes in federal countries involve many veto players that render it difficult to reduce spending in response to globalization pressure. The correlation between higher openness and lower spending should therefore only hold for unitary countries. Adding the regime type as a genuine categorical variable to the analysis, the basis for case selection is a 2 x 2 table (Table 3.4).

A test of the efficiency hypothesis against unitary countries now requires the choice of one unitary country with high spending and one unitary country with low spending. The goal of the analysis then is to determine whether the two countries qualify as closed and open, respectively. Which of the nine low-spending and five high-spending countries is chosen is irrelevant because each country is representative for all other unitary countries that populate the same cell. The insights generated in this typical case study would not be extended to federal countries because they are qualitatively different as regards the regime type. Consequently, an additional typical case study is in order for federal countries.

Case selection for set-relational case studies follows similar lines in the face of rival accounts, meaning that there is, of course, a need to control for competing causes. Suppose we are interested in the *necessary conditions* of high spending and hypothesize that a high level of economic openness is necessary. Another, competing hypothesis states that federalism is necessary for high social spending. In order to test these propositions, one could simply select a country with high social spending and test whether the purported conditions are in place. However, it may turn out that the selected state is

Table 3.5 Choice of cases for test of sufficiency hypothesis on the basis of a 2 × 2 table

		Regime type	
		Unitary	*Federal*
Economic	*Low*	7	2
openness	*High*	7	5

federal and open, making it impossible to discriminate between the two hypotheses. This situation can be circumvented if we include the regime type of the country in the picture in the case selection stage. The distribution of cases in Table 3.4 would tell us that federalism is not necessary for high spending because five unitary states maintain high spending. A test for the necessity of economic openness should be based particularly on one of these five countries in order to fully focus on the link between openness and spending.

Case selection for *sufficiency* is most powerful when it follows the same principle of control. Imagine now that high economic openness and federalism are hypothesized to be individually sufficient for high spending. If case selection is concerned only about the choice of an open country, one may end with a country that is federal and has a high level of spending. This situation can be avoided by selecting an open country that is unitary. Table 3.5, which cross-tabulates the level of economic openness and the regime type for the 21 countries, provides the basis for systematic case selection. The table shows that 7 countries are unitary and maintain high spending, rendering them valuable for a test for the sufficiency of high openness.

Diverse case studies

The basic features of diverse case studies and their implementation for case selection for process tracing have been discussed already. Taking this discussion as the basis, the elaboration of diverse case studies for a test of cross-case hypotheses is simple. In correlational studies relying on *differences in degree*, the idea is to pick two cases that span the maximum range of cross-case scores for the cause or the outcome, meaning that all other cases are located between the two extreme cases. If a hypothesis can be confirmed, the assumption is that the analysis of any other pair of possible cases would support the proposition as well. In case studies that test correlational or set-relational hypotheses involving *differences in kind*, one can implement a diverse case study in ways described above with the additional goal of checking the validity of the cutoff points imposed on a continuous variable. The only difference between hypothesis-testing and hypothesis-building case studies is that the former are centered on a cause *or* the outcome, the implication being that the selected cases need to be diverse only on one factor as opposed to two.

Diverse case study and rival hypotheses

In terms of the assessment of rival hypotheses, the typical and diverse case study face similar challenges and share the same solution. Consequently, the discussion of tests of competing hypotheses in diverse case studies can be kept to a minimum. If one wants to make the working hypothesis subject to a test via a diverse case study, one needs to make sure that other potential causes take similar scores in both selected cases. Whether the other causes are hypothesized to form an interaction with the cause of main interest or are stand-alone causes does not matter in this respect.

Returning to an example from the previous section, imagine you hypothesize that the degree of welfare state spending depends on the level of economic openness in interaction with the unemployment rate. Figure 3.6 reproduces Figure 3.4 by plotting social expenditure against the unemployment rate. In a diverse case study focusing on the effects of globalization, we are seeking two countries with the maximum difference in welfare spending and the minimum difference in unemployment rates.[20] Denmark and the United States meet these two criteria; the pair of countries would be

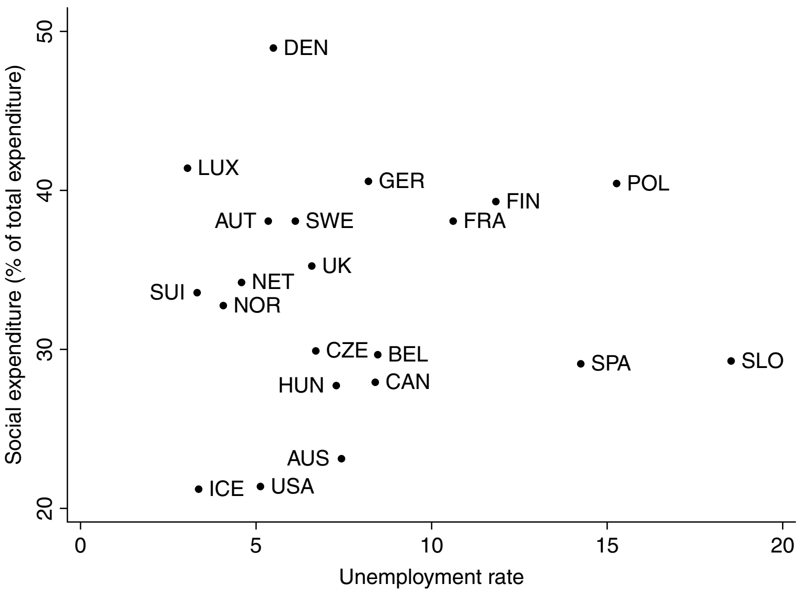

Figure 3.6 Case selection on the basis of two continuous variables for test of an interaction effect

different if the focus was on social spending alone (Denmark and Iceland would then be the choice, though these two countries also do not differ much in unemployment rate). This example again underscores the importance of taking all relevant causes into account in case selection for comparative tests of hypotheses.

Theory-based types and selection: crucial, most-likely, and least-likely cases

The theory-guided choice of cases is the alternative to distribution-based case selection when tests or modifications of hypotheses are the goal.[21] Three types of case studies and corresponding case selection principles can be distinguished for hypothesis-testing research: crucial, most-likely, and least-likely case studies (Eckstein 1975; Lijphart 1971, 692). Crucial case studies require the identification of cases for which a hypothesis is almost certain to be confirmed or disconfirmed. A positive or negative test result provides evidence beyond reasonable doubt because alternative reasons for the result, such as chance effects or measurement error, can be denied. Thus, it is possible to perform a crucial case study if a hypothesis makes precise, invariant predictions that can be measured with causes and outcomes that entail negligible measurement error.

The high demands that a crucial case study must meet explain why it has little applicability in the social sciences (Eckstein 1975). There is a widespread belief that the social world is governed by probabilistic causal relationships and that hypotheses should be framed accordingly (Goldthorpe 2000; Lieberson 1991; Zuckerman 1997). Moreover, current social science theories and measures are imprecise (Freedman 1991; Lieberson and Horwich 2008), making it always possible to attribute the test result to these factors and salvage a hypothesis from (dis)confirmation.

Most-likely and least-likely case studies are more frequently implemented in the social sciences because they are relaxed variants of the crucial case study. A most-likely case has a relatively high probability of confirming the proposition under scrutiny, while a least-likely case goes hand in hand with a comparatively low probability.[22] For example, if one is interested in the effects of globalization on welfare state spending, it is most likely to detect an effect in countries that exhibit high levels of trade and direct foreign investment. This example indicates that every most-likely case can be framed as a least-likely case, and vice versa (Eckstein 1975, 118–19). A country that is most likely (least likely) to respond to the pressures of globalization is least likely (most likely) not to show a reaction. It is common practice to use the labels in relation with the confirmation of a proposition, meaning that cases are expected to be most likely and least likely to confirm a hypothesis.[23]

Most-likely and least-likely case studies are frequently performed in empirical research (for example, Grieco 1990; Zangl 2008) and receive favorable treatment in the methods literature because of the opportunity to update our confidence in a proposition through the analysis of a few cases or even a single case (Abell 2009a, 2001, 2009b, 2004; Bennett 2008; Dion 1998; McKeown 1999; Rueschemeyer 2003). The tight link between a hypothesis and cases rests in the formulation of an expectation of how likely it is to find a specific observable implication confirmed in a given case. Depending on whether a very likely or unlikely implication receives empirical support, we can adjust our trust in the validity of a hypothesis accordingly. What 'accordingly' means is best discussed in the context of a general treatment of Bayesian causal inference and thus is postponed until Chapter 8. The conventional argument behind the choice of most-likely and least-likely cases (which is qualified in Chapter 8) rests on the intuitively plausible goal of generating surprising insights. This means that the confirmation of a most-likely test is less insightful than a passed least-likely test and that a failed most-likely case is more valuable for learning something about a hypothesis than a nonsuccessful least-likely test.

In order to know how surprising empirical evidence is, it is necessary to formulate the probability of finding a hypothesis confirmed by empirical evidence *conditional* on the assumption that the hypothesis is correct and conditional on features of the selected case (to be detailed below). Most-likely and least-likely case studies are usually implemented in an informal manner, that is, without providing a specific conditional likelihood for finding supportive evidence. From the viewpoint of rigorous small-n research and pending the discussion in Chapter 8, however, there are three ways in which case studies would benefit from a more formalized approach. (For ease of presentation, I refer only to most-likely cases in the following paragraphs).

First, one may argue that it is impossible to quantify the probabilities because the qualitative aspects of a case cannot be reduced to a single figure.[24] It is certainly true that the translation of qualitative features of a case into probabilities is a difficult enterprise. Nevertheless, this step merely brings to the forefront what is left implicit when most-likely and least-likely cases are picked on an informal basis. Formalization does not change anything about most-likely and least-likely case selection. On the contrary, one makes it more transparent because one has to explain why exactly one believes that a case is a most-likely case and how likely 'most likely' is.

Second and relatedly, the mere notion of a most-likely case is unspecific because 'most likely' is a relative criterion. How could one argue that a case is most likely of all cases in the population to support a hypothesis when one does not know the conditional probabilities for all cases? Moreover, the case with the highest probability of all cases in the population truly qualifies as a most-likely case, but the likelihood of confirming the proposition may still be relatively low in absolute terms. If the probability of the most-likely case is, say, 60 percent, reference to a most-likely case tends to be misleading because 'most likely' may suggest a higher likelihood in the eyes of the reader. This is an essential aspect because the implications of a passed or failed hypothesis test depend, among other things, on the size of the conditional probability (see Chapter 8 and Dion 1998). In order to prevent the reader of a case study from mistaking the label 'most likely' for a much higher conditional probability, it makes sense to spell it out.

Third, in the realm of comparative hypothesis testing, one may exaggerate differences between the explanatory power of rival propositions. Grieco's case study of international economic policy making illustrates this point. Grieco (1990) tests the explanatory power of liberal institutionalism and neorealism against the implementation of agreements, so-called codes, on the reduction of nontariff barriers to trade that were negotiated at the GATT Tokyo Round in the 1970s. Grieco asserts that the implementation of the codes is a most-likely case for liberal institutionalism because this theory should particularly apply to international economic cooperation. At the same time, this case is claimed to be a least-likely case for neorealism, the main domain of which is international security affairs.

These arguments cannot be disputed, but they may exaggerate the difference in the *ex ante* conditional probabilities regarding the explanatory power of neorealism and liberal institutionalism. The notion of 'code implementation' as a most-likely and least-likely case suggests differences in kind between liberal institutionalism and neorealism that may be more properly described as differences in degree. For example, it is more interesting to compare two hypotheses that have conditional likelihoods of 90 percent (liberal institutionalism) and 20 percent (neorealism) than those with 60 and 50 percent. Informally seen, both constellations allow it to declare the codes to be most likely to confirm liberal institutionalism and least likely to support realism. However, formalization of the conditional probabilities highlights that the prior confidence in the hypotheses diverges much more strongly in the first constellation than in the second. For these three reasons, there is added value in formalized theory-based case selection. Case study researchers who prefer to select most-likely and least-likely cases on an informal basis should take the previous arguments into account as well in order to avoid some potential pitfalls of this case selection strategy.

Presuming now that one engages in formalized theory-based case selection, it is necessary to specify the probability of gathering supportive observations for the working hypothesis conditional on the selected case and the assumption that the hypothesis is true. As an illustration of this likelihood and means to quantify it, let's refer to the democratic peace hypothesis as H_{DP}. It states that two democratic countries are at peace, which is the outcome of interest. The conditional probability of interest then reads $p(E|H_{DP}\ \&\ case)$, meaning that the probability of collecting confirming evidence – E – is conditional on the assumption that H_{DP} is correct and in light of theory-relevant features of the chosen case (see below).[25]

As regards the democratic peace example, the theoretical state-of-the-art implies that if one selects two established democracies that are known for being averse to war and with heads of state that are deeply committed to democratic norms, one can safely argue that $p(E|H_{DP}$ *& case*) is about .99 or so. This means that given the assumption that the democratic peace hypothesis is true and what we know about the selected cases (established democracies and so on), there is only a very slight chance these two countries will be at war with each other.

The picture changes when one selects two countries that are coming to the end of a transition from authoritarianism to democracy and are known to have fought wars with each other in the past. Although the states qualify as democratic, one may argue that their recent transition to a democratic regime renders them politically unstable and thus with an increased probability of going to war (Mansfield and Snyder 2005), in particular, because they previously fought each other. These features of the case suggest a lower conditional probability compared with a dyad with two established democracies that have been always peaceful.

The democratic peace example illustrates how conditional probabilities can be derived in empirical research. In the given example, the cross-case cause 'democratic dyad' does not influence the conditional likelihood in itself because the analysis of a democratic pair of countries is a prerequisite for a meaningful empirical analysis. More generally, the cross-case cause is not relevant here because it is categorical. This differs when one measures *differences in degree* because the difference between the cases' score on the cause affects the conditional probability. Constantelos's analysis of lobbying behavior in federal countries, in particular, the United States and Canada, is illustrative in this respect (2010). The main hypothesis is that lobby groups and companies adjust their lobbying behavior to the relative importance of the national and regional level of the political system. The more salient one level is relative to the other in terms of policy making authority, the more this level should be the target of lobbying activities. Constantelos compares the United States and Canada, which differ only slightly in terms of the relative importance of the national and regional level. Because that difference is small, there is very little likelihood of observing a significant difference between the lobbying activities in the two countries.[26] In terms of theory-based case selection, the United States and Canada thus constitute a least-likely case study. Considering that Constantelos gathers supportive empirical evidence, one can be confident that lobbying behavior also adapts to the relative authority of the subnational and national level when their relative importance is wider.

The measurement level in terms of continuous and categorical measurement is one issue that needs to be taken into account in the specification of conditional probabilities. As demonstrated by the democratic peace example, a second issue concerns the influence of factors that moderate the effect of the cause of main interest. For example, the likelihood that two democracies do not fight each other hinges on the age of the democracy. Members of the political elite in democracies that are in the midst or at the end of a transition process are not as committed to norms as the elite in an established democracy because the latter have more experience with democratic procedures and practices. Another factor that could theoretically moderate the effect of democracy is a border dispute because a long-standing dispute increases the probability of war between two countries. Other possible factors are ethnic hostilities between the two states, the fact that one state came into existence through its secession from the other country, and so on.

The general prescription for the formulation of conditional probabilities therefore is to list every empirical feature of the case and reflect on whether and how a given element influences the conditional likelihood. It is important to note that case-specific features should moderate the effect of the main cause and not constitute a rival cause (which might not always be easy to determine). Most-likely and least-likely case studies always ask about the probability that a specific cause produces the outcome, that is, the analyses take an effects-of-causes perspective on the causal relationship. Another possibility would be to ask how likely it is to observe the outcome, which implies taking a causes-of-effects view. In this instance, it is not important how likely it is that a single cause (and moderating factors) can generate the outcome but how many causes capable of producing the outcome are given in a case. Although case studies can also consider how likely it is to observe the outcome anyway, the classic most-likely and least-likely case takes an effects-of-causes perspective.

Arguments on types of cases and case selection for a test of cross-case hypothesis are summarized in Table 3.6. For theory-based case selection, it should be understood that this table and those following include information only on the outcome and cause of main interest. The score that a case takes on the cause plays a role in the designation of cases as most likely or least likely, but the previous discussion shows that there is more to this because conditioning factors need to be taken into account as well. However, this is a general characteristic of most-likely and least-likely case studies and thus not specific to any version of theory-based selection. In order not to overcomplicate the tables, they do not include information on conditioning factors.

Within-case level

For *distribution-based* case selection, the types of case studies and corresponding selection principles are easily explained because they are identical to those that apply to hypothesis-building research. The difference between process tracing for the formation and testing of hypotheses lies in the nature of the within-case analysis, which is exploratory for the former research goal and confirmatory for the latter (see Chapter 1). Beyond that,

Table 3.6 Types and selection for test of cross-case hypothesis

Basis of case selection	Case label	Case selection principle: *Correlation*	Case selection principle: *Set relation*
Distribution	Typical	*Difference in degree:* two cases creating variance on X or Y and with maximum similarity to other cases *Difference in kind:* two covarying cases from different categories on X or Y	*Necessity:* Y present *Sufficiency:* X present
	Diverse	*Difference in degree:* two cases creating variance and being most extreme on X or Y *Difference in kind:* two cases from the most extreme categories on X or Y	*Necessity:* cases are just so members and full members of Y *Sufficiency:* cases are just so members and full members of X
Theory	Most-likely or least-likely	Variance on either X or Y	*Necessity:* Y present *Sufficiency:* X present

the same types of case studies and case selection rules apply. This means one chooses typical or diverse cases and that cases are selected with respect to their scores on the independent *and* the dependent variable in ways described above.[27]

In the context of hypothesis-testing analyses, intentional selection on the cause and the outcome is in discord with King, Keohane, and Verba's recommendation not to choose cases if one knows at the outset that the test will confirm the hypothesis of interests (1994, 142–6). This advice must be seen in light of their (implicit) preoccupation with hypothesis-testing on the cross-case level. For a within-case analysis, King, Keohane, and Verba's prescription is invalid because the selected cases should lay a solid foundation for process tracing. This is only achieved by choosing cases that meet the requirements in terms of the cross-case scores on the cause (or causes) and the outcome. While King, Keohane, and Verba's advice about purposeful case selection does not apply in the original sense, it can be transferred to the within-case level. If a case is selected with the knowledge that it passes the cross-case test, one of course should not know at the case selection stage that it will also confirm the within-case hypothesis.

Table 3.7 Types and selection for test of within-case hypothesis

Basis of case selection	Case label	Case selection principle: *Correlation*	Case selection principle: *Set relation*
Distribution	Typical	*Difference in degree:* two covarying cases with maximum similarity to other cases *Difference in kind:* two covarying cases from different categories on X and Y	X and Y present
	Diverse	*Difference in degree:* two covarying cases being most extreme on X and Y *Difference in kind:* two covarying cases from the most extreme categories on X and Y	Cases are just-so members and full members of X and Y
Theory	Most-likely or least-likely	Covariance of X or Y	X and Y present

In *theory-based* case selection, the logic of case selection for the test of within-case hypotheses is similar to that of cross-case studies. The hypothesis of interest refers to a causal mechanism or causal process, but the probability of finding confirming evidence for a within-case proposition is derived in the same way as described above and thus needs not be replicated here. The main arguments on case studies that test within-case hypotheses are depicted in Table 3.7.[28]

Building and testing hypotheses: on the same case or different cases?

After having discussed hypothesis-building and hypothesis-testing case studies, it is now possible to turn to one of the more widely perceived messages of King, Keohane, and Verba's *Designing Social Inquiry* (1994) and the responses it has received. They recommend using different data for the formation and testing of hypotheses (1994, 46). Otherwise, the proposition at hand is certain to find empirical confirmation because the hypothesis from which a derived from exploratory process tracing is tested against the very same data. This advice is embraced in principle by the George and Bennett, but they arguing that the same case can be used if one focuses on different observable implications (and thus data) in both stages of the research process (2005, 111–12).

The advantage of relying on the same case is practical because it saves resources compared with the analysis of different cases (which for example could be countries located on different continents). To illustrate that this strategy is viable in principle, assume that one is interested in the determinants of the democratic peace phenomenon. One hypothesizes on the basis of an exploratory case study that democratic norms account for peace because the political elite is found to be committed to the principle of peaceful conflict resolution. Moreover, the political elite trusts the elites in other democracies because of the expectation that all elites adhere to the same norms (Rosato 2003, 586).

One way to test the veracity of the norms-based argument is to search for an auxiliary outcome (Mahoney 2010, 129–31), that is, an effect other than democratic peace that should be in place if commitment to norms of peaceful conflict resolution is indeed the cause. One can claim that norms of peaceful dispute settlement and trust should also be at play in trade disputes and prevent the outbreak of a costly commercial war. One can test this proposition by analyzing the same countries that were used to generate the hypothesis that norms ensure peace between democracies. In the hypothesis-generating case study, the case is therefore an instance of peace between two democracies, while the test of the trade proposition focuses on the commercial relations of the same two democracies. This example shows that the trade-related hypothesis focuses on a different case (as defined in Section 2.1) and implies the search for different evidence. However, the case study would still be about the same countries, which can be taken as the unit of analysis in this example.[29] Regardless of this issue, the broader question is not whether one can use the same case for the formation and testing of a hypothesis but how appropriate this strategy is.

In light of the preceding discussion on case selection, it is impossible to give an unambiguous answer to this question. It is possible that the case that served as the basis for the formulation of a hypothesis is also appropriate for a test of hypotheses. However, there is no guarantee for this because the auxiliary hypothesis is derived only after exploratory process tracing has been performed. The previous discussion of case selection for tests of hypotheses demonstrates that the identification of suitable cases is performed against the backdrop of theory, which is either used to construct a distribution of cases or to derive conditional probabilities. In either scenario, intentional case selection is mandatory because one wants to choose a case that has, for example, a maximum number of similarities to other cases. If intentional case selection is not performed, it is simply a matter of pure chance how suitable the case is for an intelligible test of the hypothesis.

With regard to the democratic peace example, imagine that you would like to perform a least-likely test because of the hope that it will pass the test and generate unanticipated insights. However, if you also want to choose the same two countries that you chose for the formulation of the democratic norms explanation, your hands are tied as regards case selection. All that you can hope for is that the two countries exhibit features that qualify them as least-likely cases in light of the hypothesis that norms of peaceful conflict resolution also matter in trade conflicts. For instance, this would not be the case if the selected states are small because small countries have a higher likelihood of settling trade conflicts peacefully as they are likely to suffer greatly from trade wars (Conybeare 1987).

The bottom line is that the same country (or institution, policy field, and so on) can be used to generate a hypothesis and to test auxiliary observable implications, but this leaves one wholly uncertain about how appropriate the cases are for the test. In this respect, there are clear advantages in favor of starting case selection anew after having formulated additional propositions.

3.3 Types and selection for modification of hypotheses

Cross-case level

The goal of hypothesis-modifying research is to resolve anomalies or puzzles. This means that a hypothesis receives empirical support from a sufficiently large number of cases in the population but that some cases that should conform to the proposition fail to do so. Generally speaking, a cross-case puzzle is in place when cases do not have the score on the outcome (cause) that they should have owing to their cross-case score on the cause (outcome). In *distribution-based* case selection, this specifically means that a case is similar to a typical case on the cause (or the outcome) but not on the outcome (or the cause). These cases, known as *deviant cases*, represent an empirical puzzle and deserve closer scrutiny. The identification of a deviant case for the modification of a hypothesis requires the choice of cases on the basis of their cross-case scores on the cause *and* the outcome. Intentional selection on the cause and the outcome is indispensable. How else could one be able to modify a theory if one does not make sure that the selected case is suitable for this purpose in the first place?

Gamson's law (1961) is useful for illustrating the modification of hypotheses in a *correlational* framework. Gamson's law states that there is a one-to-one correspondence between the parliamentary seat share of a government party, which is the cause, and its share of cabinet posts, the outcome. In light of this proposition, which receives strong empirical confirmation (Warwick and Druckman 2006), a deviant case would be a party that has a disproportionately large or small cabinet share in comparison with its share of parliamentary seats. The aim of the case study then is to determine the reason for the unexpected cross-case score. The pursuit of this goal presupposes process tracing that aims to find the variable, which explains why this case is different from the typical cases.[30] Such a variable could be that landslide election victories give a party additional bargaining leverage in the coalition formation process and that Gamson's law should incorporate this variable. An additional, less widely known deviant case displays exactly the opposite cross-case scores on X and Y. Case selection for this form of deviant case is straightforward (and discussed under the rubric of an inverse correlation comparison in Chapter 4).

Set-relational case studies follow the same logic and call for an intentional choice of cases on the condition and the outcome. Taking the democratic peace phenomenon as the basis, an example of a deviant case in a set-relational case study on sufficiency would be two democracies that waged a war with each other. This case has the same cross-case score on the condition as a typical case but exhibits a striking outcome because one would expect peace. A second variant of deviant case is one having the outcome present in the absence of all purported sufficient conditions because of the expectation that the outcome is absent. Finally, a deviant case with respect to necessity is a case that displays the outcome but not the condition. If the outcome of interest is war, a deviant case would be a democratic dyad at war, since the democratic peace phenomenon implies that a nondemocratic pair of countries is necessary for the occurrence of a militarized conflict.

For *theory-based* case selection of anomalous cases, the relevant criterion is whether the empirical analysis produced surprising insights. This means that one should aim at *failed most-likely cases* and *passed least-likely cases* because they represent puzzles (Lijphart 1971, 692). Again, it holds that the choice of cases is intentional and based on a case's cross-case scores that deviate from the theoretically expected scores. An example of a failed *correlational* most-likely case in the context of Gamson's law is a party that has a cabinet share that is much higher or lower than its share of seats because it is most likely to have a share of parliamentary seats that corresponds to its share of cabinet posts. An example of a failed most-likely case in a *set-relational* setting would be two consolidated democracies with no history of war that fight each other because they are most likely to maintain peaceful relations.[31]

Within-case level

In *distribution-based* case selection, case studies that aim to modify a within-case hypothesis follow the same principles as their cross-case counterparts and should be based on the analysis of *deviant* cases. However, there is an important difference between the two variants, as there are two types of within-case studies. In the first scenario, a failed cross-case and within-case test go hand in hand. With regard to Gamson's law, presume a party has a share of cabinet posts that is much smaller than its seat share, which is a violation of the cross-case expectation. Further assume that the invocation of a proportionality rule by the negotiating parties is hypothesized to be the cause for a proportional allocation of cabinet shares. Given failure of the cross-case hypothesis, the within-case expectation equally lacks empirical resonance. The question that the exploratory case study needs to answer is why the parties behaved differently than expected, that is, why they did not rely on a proportionality rule. If this question can be answered via exploratory process tracing, one is likely to solve the cross-case puzzle as well because of the close link between the two levels. For example, process tracing could show that the party experienced a landslide loss of seats that weakened its bargaining position during the coalition formation negotiations. Gamson's law would then have to be amended by the condition 'no landslide loss of seat shares'.

The second variant of hypothesis-modifying research is characterized by a successful cross-case test and a failed within-case proposition. As regards Gamson's law, this implies that the seat share and cabinet share correspond to each other, but the correspondence cannot be attributed to the proportionality principle. Since the cross-case relationship nevertheless holds, the proportional allocation of cabinet posts seems to be the result of equifinal causal processes. Equifinality is likely to be in place on the within-case level because one knows already that some typical cases confirm Gamson's law owing to the proportionality rule. Otherwise, there would be no pattern from which the deviant case could diverge and no within-case hypothesis that could be modified because the modification of a hypothesis presumes that the proposition finds empirical support by a sufficiently large number of cases (see Section 1.3). Therefore, one major goal is to discern a second determinant of the correspondence between seat and cabinet share. In addition, one should try to determine the conditions under which one causal mechanism or the other accounts for the outcome (Falleti and Lynch 2009).

The different manifestations of deviance have important ramifications for case selection. Cases that are deviant on the cross-case and the within-case level are relatively easy to determine because one only needs to search for anomalous cases on the cross-case level. Case selection is more protracted when deviance occurs only on the within-case level because these cases are more difficult to identify. A simple look at the cross-case scores of cases does not suffice for informed case selection, rendering it necessary to analyze each case individually so as to discern whether they confirm the within-case hypothesis. Depending on how many cases are deviant on the within-case level, this may demand the collection and evaluation of considerable cross-case and within-case evidence before a case can be chosen for the actual empirical analysis. This requirement arguably makes this case study the most challenging variant because one cannot know at the outset how many cases one has to screen before one or more instances of this type are detected.[32]

If case selection is based on theory, *failed most-likely* and *passed least-likely* cases are puzzling and should be the object of case selection. In searching for anomalous cases, the caveat that I detailed in the previous paragraph extends

Table 3.8 Types and selection for modification of cross-case and within-case hypotheses

Level of analysis	**Basis of case selection**	**Case label**	**Case selection principle**	
			Correlation	*Set relation*
Cross-case	Distribution	Deviant	Cross-case scores differ from typical case on X or Y	*Necessity:* Y present and X absent
	Theory	Failed most-likely or passed least-likely	Theoretically unexpected cross-case scores on X or Y	*Sufficiency:* X present and Y absent or all X absent and Y present
Within-case	Distribution	Deviant	Causal mechanism/process operative in typical cases not in place	
	Theory	Failed most-likely or passed least-likely	Expected causal mechanism/process not in place	

to the theory-based variant. A case may fail the cross-case and within-case expectations, which makes it easy to identify and select for process tracing. However, it is equally conceivable that some cases exhibit the hypothesized cross-case pattern and yet fail to confirm the within-case proposition. The pervasiveness of this situation can be determined only by screening all cases that confirm the cross-case proposition with respect to their performance on the within-case level, which renders this case study variant as cumbersome as its distribution-based counterpart. Table 3.8 summarizes the main features of hypothesis-modifying case studies.

3.4 Conclusion

Over the last three decades, the case study literature gradually developed an inventory of types of case studies. One aim of this chapter was to place this inventory on a systematic basis by distinguishing between multiple dimensions that are relevant for the discussion of types of case studies and case selection. In combination, the dimensions allow one to locate all types of case studies in a unique scheme. The scheme highlights the fact that there are no simple recipes for case selection that hold true in all instances. Variance on the cause or the outcome is not always the best choice in correlational research, much as invariance on the cause and the outcome is not always the superior strategy for set-relational analyses. The only principle that extends to all types of case studies is that of intentional case selection. Although the intentional choice of cases has been seen with some suspicion in the past (Fearon and Laitin 2008), there is little basis for skepticism *if* a case study researcher makes explicit what criteria applied in the choice of the specific type of case.

The discussion of types of case studies further shows that different selection strategies and types of case studies are available for the testing and modification of hypotheses. This gives rise to the question of whether one selection strategy is preferable to the other. As mentioned in the beginning of this chapter, distribution-based and theory-based case selection are tied to a frequentist and Bayesian mode of causal inference. Since these are discussed in detail in Chapter 8 and Chapter 9, a comparative evaluation of the two modes of causal inference is postponed until then.

4
Forms and Problems of Comparisons

In principle, comparisons can be made on the cross-case level, the within-case level, and both simultaneously (see Section 1.4). A within-case comparison is appropriate for discerning whether the causal mechanism and causal processes are similar in the two (or more) cases at hand. In a cross-case comparison, on the other hand, one is determining the nature of the causal effect of a given cause. The two generic forms of comparisons are closely intertwined because it is useful to know whether the causal effect is underpinned by a mechanism and, in addition, what causal effect is produced by that mechanism. In this and the next chapter, the focus is on the cross-case level because of the special issues involved in these comparisons. The following chapters expand into process tracing and the analysis of causal mechanisms.[1]

Cross-case comparisons have been subject to heated debate for decades now. This chapter builds on this debate and elaborates the logic and problem of comparative case studies by explicit linkage of the three research goals with the correlational and set-relational view on cross-case relationships. The arguments extend to case studies for which the cross-case level is the theoretical end and the empirical means for process tracing. For presentational purposes, I limit the discussion to the purpose of generating cross-case inferences. This is not to deny that sound comparisons are vital for within-case analysis. On the contrary, if one wants to learn how a sufficient condition produces the outcome, one needs to be sure that the condition is sufficient in the first place and not part of a conjunction or an INUS condition. In light of the discussion of case selection in the previous chapter, all the points made in this chapter can be easily extended to case studies having the within-case level as theoretical end.[2]

Throughout the discussion of cross-case comparisons, an important distinction concerns the differentiation between *ideal*, or *optimal*, and *imperfect*, or *suboptimal*, comparisons. A design is optimal if the observable cross-case scores match the pattern that one wanted to create by the purposeful choice of cases. If one wants to compare two cases that differ only on the outcome and a single cause – the method of difference – the realized comparison is ideal–typical if one can find two cases that exactly meet these criteria. A comparison is imperfect if the observed cross-case scores deviate from the desired ones. In this example, an imperfect result would mean that the selected cases differ on the outcome and two or more causes.

Apart from this distinction, it is necessary to differentiate between *determinate* and *indeterminate* causal inferences. Since causal inference can pertain to conclusions about causal effects and causal mechanisms, one may encounter *effect-related* and *mechanism-related* indeterminacy in small-n research. Inferences on effects are determinate when only a single inference can be imputed in the cross-case pattern at hand. Correspondingly, a comparison is indeterminate as regards causal effects when more than one inference is feasible (Brady and Collier 2004, 236–7). The meaning of mechanism-related determinacy is then straightforward; causal inference on mechanisms is determinate when we are confident that only a single mechanism ties the cause to the outcome and indeterminate when more than one causal mechanism can account for the same causal effect.

Since causal effects and causal mechanisms refer to different elements of causal relationship, effect-related and mechanism-related indeterminacies can occur together. In its present state, the democratic peace phenomenon can be seen as an instance of their co-occurrence and is therefore useful for illustration. On the level of effects, some scholars dispute that the democratic nature of two countries accounts for peace between democracies and point to capitalism and economic exchange between countries as the actual causes of peace (Schneider and Gleditsch 2010). Leaving aside here whether the effect of democracy is correlational or set-relational, there is thus uncertainty about whether any effect at all can be ascribed to a democratic dyad. In addition, there is indeterminacy as regards the causal mechanisms because there are multiple elements of a democracy that are plausibly conducive to peace. For example, it could be because the political elite in democracies is committed to democratic norms and trust or because democratic institutions in some way constrain the elite from launching a war (Rosato 2003).

The formulation of an unambiguous causal explanation covering the cross-case and the within-case level therefore requires determinate inferences on causal effects and mechanisms alike. Seen from this perspective, both forms of indeterminacy are equally problematic. In the present chapter, the focus is on effect-related indeterminacy because this is the variant that one can confronts in cross-case comparisons. Mechanism-related indeterminacy manifests itself on the within-case level and must be addressed via process tracing, which is the reason that a discussion of this version of indeterminacy is postponed until Section 7.2.

The two pairs of terms – ideal–typical v. suboptimal and determinate v. indeterminate – should be kept separate. Under certain conditions, an ideal–typical comparison allows one to make only one causal inference. However, I show in this chapter that the analysis must meet additional, demanding requirements, and so it is likely that even the desired cross-case scores will allow only for indeterminate causal inferences. On the other hand, it is demonstrated below that suboptimal comparisons are always tied to indeterminacy because an ideal–typical comparison is necessary though not sufficient for unambiguous inferences.

For a long time now, the questions of determinacy v. indeterminacy and ideal v. suboptimal designs have been addressed with considerable intensity and different conclusions have been reached. On the one hand, Mill's famous method of difference and method of agreement have been perceived as ideal–typical comparisons for reasons that are detailed in this chapter (see Anckar 2008; De Meur and Berg-Schlosser 1994, 1996; DeFelice 1986; Frendreis 1983; Lijphart 1971). On the other hand, both methods have been criticized repeatedly, and their inferential utility has been severely questioned (Burawoy 1989; Goldthorpe 1997a, 4–5; Lieberson 1991).

In light of the ardent criticism and the repeated reference to Mill as himself being critical of the potential of his methods (Lieberson 1994, 1226), it may come as a surprise that this chapter spends some space on Mill's methods and cross-case comparisons more generally. Four reasons account for this decision. First, the method of agreement and method of difference are useful under specific circumstances. The conditions are restrictive, but they can be met and demand some discussion. Second, the two methods have been and still are discussed in the recent literature (Anckar 2008; Caramani 2010; de Vaus 2008; Lijphart 1975, 1971), and their logic is followed in empirical research (e.g., Trampusch 2010; Walter 2008).[3] Those researchers seeking to employ them should receive some guidance about their respective pros and cons and inferential potential.

Third, I concur with much of the criticism of Mill's methods. At the end of this chapter, however, I also aim to ask what the alternatives are. Depending on the research purpose and causal effect, it is advisable to construct cross-case comparisons other than the method of agreement or difference. However, these designs suffer from similar inferential restrictions. The problems with Mill's methods are not due to how method of agreement and difference works, but because they rely on a small number of cases. This implies that the shortcomings of Mill's methods are problems of small-n comparisons more generally. I highlight this point by discussing Mill's methods along with alternative designs. The general insight may lead one to conclude that small-n comparisons should be abandoned altogether (Lieberson 1994, 1236), which is not the perspective that I take because cross-case inferences can play a role in case studies. Moreover, the abandonment of cross-case analyses would deprive case study researchers of their basis for case selection for process tracing. Consequently, a rigorous discussion of cross-case comparisons is in order.

Fourth, as mentioned before, it has been repeatedly claimed that Mill argues against the application of his methods in observational research. It is of course interesting to know what the founder of Mill's methods thinks about their utility, but even if Mill would oppose their implementation, everyone is free to use his methods so as long as it is understood that there are provisos attached to each of his five methods. Besides, Mill is not fully against the application of his methods; he is, however, opposed to their implementation in an inductive fashion (which he calls the chemical method). Here, 'inductive' means an empirical analysis that is largely free of theory and infers causality from 'pure observation and [an] experiment' (Mill 1874, 613).[4] Mill claims that this is not possible in the social sciences because such experiments cannot be constructed and that, in addition, social and political phenomena are believed to be due to interaction effects and equifinality (610–13).[5] Consequently, Mill concludes at the end of a critical discussion of his methods that one should rely on the 'deductive method' (613), meaning that the use of his methods in observational research should be closely tied to theory (Zelditch 1971). It is thus misleading to present Mill as an opponent of *any* use of his methods. Equally important, his plea for the tight coupling of theory with his methods is fully in line with the discussion of theory-centered case studies in this book.

In this chapter, I start the discussion of comparisons by focusing on designs that include two cases and binarily measured causes and outcomes. Comparisons that include more cases and/or rely on multicategorical measurement are left for discussion in the next chapter. I start with pairwise comparisons with binary measurement because these are the simplest designs one can implement. Furthermore, they represent the comparison that is most often discussed in the small-n literature (Tarrow 2010). Taking this discussion and the diagnosis of pertinent inferential problems as the basis, the following chapter considers several instruments for the improvement of cross-case inferences.

Before the discussion of cross-case comparisons starts, a final note is in order on how the types of cases introduced in Chapter 3 are related to the types of comparisons elaborated in the following. This is an important matter because some discussions of case selection intermingle this topic with comparative designs (Klotz 2008; Odell 2004; Seawright and Gerring 2008). Types of case studies and comparisons are related, if only for the simple reason that one often chooses cases in order to compare them. Nevertheless, it is important to keep these two issues separate because every type of comparison can involve every type of possible case study. For comparisons that seek to build hypotheses, one can build on typical or diverse cases. Similarly, hypotheses-testing comparisons can include typical, diverse, most-likely, and least-likely cases. Comparisons for the purpose of modifying a hypothesis can draw on deviant cases and failed most-likely and passed least-likely cases.

The factors that shape the quality of a case as typical, diverse, and so on, enter into the cross-case comparison in the form of the cause of interest and the causes one additionally needs to take into account. Since the assignment of cases as a certain type requires knowledge of a distribution of cases or theory, it is not possible to designate the selected types of cases in the stylized representation of comparisons as done in this chapter. This is not a problem insofar as the generation of cross-case inferences does not depend on the type of case under analysis. The nature of the selected cases comes into play only in the course of relating the inferences to the theories under scrutiny and their generalization, a subject that will be discussed in Chapters 8 and 9. In the following sections, I therefore remain silent on what type of case study underlies the respective cross-case comparison.

4.1 Comparisons for the formation of hypotheses

Hypothesis-building comparisons develop propositions about the causes of an outcome that can be put to a test in a subsequent round of research. All causes that one discerns are preliminary and should be treated as potential causes until a proper test has been conducted. The way in which a comparison is constructed should be based on the existing state of knowledge about the determinants of the outcome. Comparisons are crafted differently, depending on whether insights from adjacent fields of research allow one to derive the expectation that the causal relationship is correlational or set-relational. When one has an expectation only about potential determinants of the outcome but not about the nature of the causal effect, it is recommended to start with a *variance-on-Y comparison*. As discussed in Chapter 3, factors that vary across the two cases qualify as potentially sufficient, while invariant factors are individually insufficient. In a similar vein, invariant factors can be ruled out as potential independent variables, while factors that covary with the outcome are carried over to a subsequent hypothesis test.[6]

If one can make a qualified hunch that the cause–effect relationship is covariational (or set-relational, see below), one can construct a comparison accordingly by comparing two cases for which one observes covariance between the dependent variable and the purported independent variable. The confidence that the causal relationship is correlational should derive from previous research that is related to one's own case study. Although the European Union (EU) is an unprecedented supranational organization, a study on the European parliament could draw on the extant body of research on national parliaments. Imagine that you are interested in the frequency with which issues from a certain policy field make it on the agenda of the EU parliament. A review of existing research on agenda setting in national parliaments suggests that the frequency of certain policy issues on the agenda increases with the number of corresponding lobby groups registered in the capital where the parliament resides. On the basis of this research, one can derive a similar expectation for one's own case study. Even though the European parliament and national parliaments may differ in respects related to agenda setting, one can nevertheless start with a hypothesis-building comparison that is guided by the expectation of a correlation. The case study then aims to lend more credence to the hunch that the frequency of issues on the agenda is positively correlated with the number of lobby groups via exploratory process tracing. In addition, the within-case analysis serves to discern other independent variables that may not have been thought of before.

Similar hunches can guide cross-case comparisons in set-relational case studies. Depending on the set relation of interest and the state of knowledge, set-relational case studies can form hypotheses through no-variance and variance designs alike. *No-variance-on-Y designs* are appropriate for the formation of hypotheses on necessary conditions (see Chapter 3 and Most and Starr 1989, 52). This is achieved by comparing two cases that share the outcome of interest and searching for conditions that are present in both cases.[7] All conditions that vary across the two cases can be ruled out as necessary, and the remaining candidates are made subject to a test in follow-up studies.

In analyses of sufficiency, one should follow a *no-variance-on-X* approach if one has a reasonably strong expectation that the occurrence of the outcome is attributable to sufficient conditions. As the name implies and as follows from the definition of sufficiency, the comparison involves two cases having the potential condition in common. The goal of such a comparison is not to demonstrate that the cross-case pattern conforms to the expectation of sufficiency because this is known from the outset. Instead, the rationale is to do exploratory process tracing in order to determine whether and how condition and outcome are related to each other. The arguments on different variants of hypothesis-building comparisons are summarized in Table 4.1.

One problem in hypothesis-building case studies (and comparisons more generally) is that the final pattern of cross-case scores is likely to be indeterminate (Frendreis 1983, 265).[8] In correlational small-n research, one confronts indeterminacy if multiple independent variables covary with the dependent variable.[9] A necessary-condition comparison is indeterminate when more than one condition is present when the outcome is present. A sufficiency design is indeterminate when the outcome is present in the presence of two or more potential conditions. In the ideal comparison, only one cause qualifies for the causal relationship, making subsequent tests easier. In practice, however, one is very likely to be disappointed (Lijphart 1971).

A more fundamental and hitherto largely neglected problem concerns the difficulty in adjudicating empirically between a correlational causal effect

Table 4.1 Comparisons for the formation of hypothesis

		Cross-case effect of interest	Comparison	Feature	Goal
Causes derivable from adjacent fields of research	*No*	Necessity	No-variance-on-Y	Cases share presence of Y	Identify invariant conditions
		Correlation/ sufficiency	Variance-on-Y	Cases vary on Y	Identify varying causes
	Yes	Correlation	Covariance	X and Y covary	1. Process tracing substantiating confidence in cross-case effect 2. Identify additional potential causes
		Necessity	No-variance-on-Y	Cases share presence of Y	
		Sufficiency	No-variance-on-X	Cases share presence of X	

and one of sufficiency in the first place. The previous discussion showed that this is a problem if there is no body of research from which one can derive the expectation that the outcome is attributable to a covariational or set-relational effect. The consequence is that one must rely on the observed cross-case pattern for discriminating between correlational causation and sufficiency. An exploratory case study by Trampusch (2010) exemplifies how demanding this is and additionally serves to illustrate the presence of indeterminacy in hypothesis-building studies. Trampusch compares two cases of self-preserving institutional change and two cases of transformative institutional change in vocational training systems in Austria, Germany, and Switzerland (contributing two cases to the analysis). On the basis of a literature review and carefully crafted exploratory process tracing, her final cross-case comparison focuses on three causes: the type of powerful companies in a country (small/medium v. large), the strength of the labor unions (weak v. strong), and the type of coalition between the state and powerful companies (liberalizing v. protectionist as regards institutional change). The resulting cross-case pattern is depicted in Table 4.2.

Trampusch interprets the cross-case pattern with the established logic that differences in the outcome must be explained with differences in a cause and that potential causes can be ruled out as causes if they vary across cases when the outcome is invariant (Gerring 2001, 210–18). In this view, one can infer that unions are individually irrelevant for the explanation of institutional change. The strength of unions is different between Switzerland I and Austria as well as between Switzerland II and Austria, whereas the

Table 4.2 Cross-case comparison on institutional change

Case	Institutional change (Y)	Powerful company	Type of coalition	Unions
Switzerland I	Self-preserving	Small/medium	Protectionist	Weak
Austria	Self-preserving	Small/medium	Protectionist	Strong
Switzerland II	Transformative	Large	Liberal	Weak
Germany	Transformative	Large	Liberal	Strong

type of institutional change is constant in both pairs of cases. Similarly, we observe variance on the outcome in a comparison of Switzerland I and II and of Austria and Germany, though the strength of labor unions is invariant. Consequently, we can reject the argument that the strength of unions correlates with the type of institutional change and that weak or strong unions are sufficient for the outcome.

The causes 'powerful company' and 'type of coalition', however, both correlate with the outcome, and so the design is indeterminate through the lenses of the covariational conception of causation. It may be that each cause alone can produce variance on the outcome or both together must vary so as to account for a different type of institutional chance. Both inferences cannot be evaluated on empirical ground because there is no case combining small and medium powerful enterprises with a liberalizing coalition, or large powerful enterprises with a protectionist coalition.

The problem that undermines causal inference here and in general derives from the size of the *property space* underlying the analysis and the share of the space that is covered empirically.[10] The property space captures the entirety of logical combinations of cross-case scores that the causes under scrutiny can assume. If all causes are measured dichotomously, the size of the property space is determined by 2^c, *c* being the number of causes. For Trampusch's analysis, this means that the property space includes eight combinations, only four of which are observed.[11] For instance, the property space includes the combination of powerful large companies, a protectionist coalition, and strong and weak unions, respectively. Without observing the outcome for these two combinations, one cannot determine whether the nature of the powerful company and the coalition, have an individual effect or form an interaction effect.

Furthermore, one can also impute set-relational inferences into the pattern. If one takes transformative change as the positive outcome, one can argue that large powerful enterprises and a liberal coalition are individually sufficient or jointly sufficient. The empirical cross-case evidence permits it neither to discriminate between these two set-relational inferences nor to adjudicate more generally between correlational and sufficient causation. In situations such as these, it is important to acknowledge and report these uncertainties in order to lay the best possible basis for subsequent tests of the various hypotheses that are compatible with the evidence at hand.

4.2 Comparisons for the test of hypotheses

The test of cross-case hypotheses has received ample attention in the past (Goldthorpe 1997a, 1997b; Lieberson 1991; Przeworski and Teune 1970; Rueschemeyer 2003; Tilly 1997). Two designs that have been intensively discussed for decades now are Mill's *method of agreement* (MoA) and *method of difference* (MoD) (1874, 278–81).[12]

The MoA is characterized by two cases that meet three criteria: they take similar scores on the outcome, they display the same score on one cause, and they are dissimilar with respect to all other causes.[13] The intuitively appealing idea behind the MoA is that differences cannot explain similarities. This statement is true but pertains only to correlational case studies, which can be exemplified with an empirical example contrasting the value of the MoA for correlational and set-relational comparisons.[14]

Hendriks and Michels' comparison of democratic reforms in the United Kingdom and the Netherlands from 1990 to 2010 is a design built on the idea of the MoA. The United Kingdom and the Netherlands can be described as archetypical cases of a majoritarian democracy and consensus democracy, respectively (Hendriks and Michels 2011, 307). The notion of majoritarian and consensus democracies goes back to Lijphart (1999). He distinguishes the two types on the basis of two dimensions, each of which consists of five variables. The first dimension is called executive–legislative dimension and includes variables such as the form of cabinet (single-party v. coalition) and the electoral system (majoritarian v. proportional). The second, the federal–unitary dimension, includes, among others, the variable territorial distribution of power (unitarism v. federalism). Hendriks and Michels' aim is to discern whether the international debate about more direct democracy leads to the introduction of elements of direct democracy in a majoritarian and consensus democracy, two types that differ on a range of institutional variables (2011, 307). Thus, the outcome to be explained is the move to a political system that offers a greater opportunity for direct political participation. Indeed, Hendriks and Michels' empirical analysis shows that the United Kingdom and the Netherlands invented elements of direct democracy (314–15).

The MoA that is reflected in this comparison is presented in Table 4.3. For ease of illustration, I present only three of the institutional variables on which the United Kingdom and the Netherlands differ. In a covariational view, Table 4.3 shows that the institutional differences between the United Kingdom and the Netherlands can be discarded as causes because the outcome is invariant. On the other hand, the debate about direct democracy qualifies as a *potential* cause. It can neither be eliminated as an independent variable nor can it be credibly claimed to be one because the debate about

Table 4.3 Method of agreement

Case	More elements of direct democracy (Y)	International debate about more direct democracy	Election system	Government	Organization of state
United Kingdom	Yes	Yes	Majoritarian	Single party	Unitary
the Netherlands	Yes	Yes	Proportional	Coalition	Federal

direct democracy and the adoption of such do not covary.[15] The eliminatory role of the MoA in correlational research explains why it is often considered to be a weaker design than the MoD (Gerring 2001, 212–13).

It is important to note that these inferences rest on the implicit assumption that the institutional variables do have a causal effect of their own (Lieberson 1991, 313–14). Once this assumption is relaxed and one allows for the presence of interaction effects, the picture gets more complex. It could be that an institutional variable correlates with the outcome in the absence of a debate about direct democracy or when one or more of the other institutional variables is invariant across both cases. For example, it could be hypothesized that one would observe a correlation between the organization of the state and the outcome when no international debate about direct democracy is occurring. The citizens of a federal country are closer to the political process because the subnational level has legislative autonomy in at least some of the policy fields. In unitary states, on the other hand, the citizens' distance from the political process could lead to demands for more direct democracy that are not evident in federal countries. This correlation is not apparent in Table 4.3 because both countries are influenced by the international debate about direct democracy. Since the effect of the international debate supersedes the effect of the type of state organization, the latter becomes visible only when comparing two countries for which one can deny an international influence.

Similarly, there may be an interaction effect between two institutional variables that goes unnoticed in Table 4.3. Let's assume that one can credibly hypothesize a correlation between the electoral system and the outcome in unitary countries. For unitary countries, elements of direct democracy should be invented in majoritarian electoral systems, while no such effect is to be expected in unitary countries with proportional systems. The rationale might be that unitary countries offer citizens fewer opportunities than do federal states and that majoritarian electoral systems capture the diversity of a society less well than does a proportional system. Both variables together – unitarism and majoritarian electoral system – trigger a demand for direct democracy that does not arise in a federalist state or in countries that have a proportional system. When such an interaction effect can be theorized, the MoA is indeterminate because multiple inferences are compatible with the cross-case data at hand.[16]

Having discussed causal inference in the correlational MoA, a final note on the of role case selection is in order. In principle, covariational comparisons for tests of cross-case hypotheses can be constructed via case selection on the dependent or the independent variables. Because of the importance of achieving a specific pattern of scores on the independent variables, however, there are clear advantages to case selection on the independent variables. Whether nature provides one with the ideal cases is an empirical question, but the purposeful choice of cases on the causes increases the likelihood of constructing the ideal MoA (or any other desired comparison). The alternative strategy involves case selection on the outcome with a great deal of hope that the cases take the required scores on all the independent variables of interest. Building on what was elaborated in Chapter 3, a mixed strategy is the approach that maximizes the chances of creating the ideal comparison. It calls for case selection on the outcome and all independent variables that represent rival explanations. The only variable that would not be considered in the case selection part is the independent variable of main interest. In respect to the example in Table 4.3, this would imply the choice of two countries that introduced elements of direct democracy and that differ on all institutional variables. The hypothesis test could then focus on finding out whether both countries were under the influence of an international debate about direct democracy. Although types of case studies, case selection, and the actual comparison are elements that should be kept separate analytically, this short digression shows that it is important to consider them in conjunction when crafting a case study.

In comparison with a correlational analysis, the MoA is somewhat more valuable for *set-relational* case studies. In inquiries into necessary conditions, the MoA in Table 4.3 allows one to infer that the international debate about democracy is the sole necessary condition because it is always observable when the outcome is present. The ideal MoA therefore makes it possible to draw inferences on necessity. To be more precise, the classic MoA is not required if one has strong theoretical expectations as regards potential necessary conditions. Assume that it is feasible to hypothesize that potential, individually necessary conditions for more direct democracy are a debate about direct democracy, a proportional electoral system, a coalition government, and federalism. Given these expectations, the empirical analysis of the Netherlands is superfluous because one can test all these expectations with the United Kingdom alone. If one wants to perform a comparison, one should search for a second country that is exactly the same as the United Kingdom. The inference that the international debate is necessary and the other conditions not gains credibility if it can be substantiated with two cases. On the other hand, the MoA is indispensable if one has no theoretical expectations as to whether the presence or absence of a condition is necessary (Mahoney 2000, 392). All four conditions in Table 4.3 automatically qualify as necessary if one were to examine only the United Kingdom or only the Netherlands. Yet both cases together demonstrate that the electoral system, the type of coalition, and the organization of the state cannot be necessary because then the outcome should be absent in one of the two cases.[17]

The strength of theory is equally crucial in analyses of sufficiency. When one has reasonably strong expectations about what a sufficient condition is, a test for sufficiency can be based on a single case having the condition in question in place. Still, the MoA has some merit in case studies on sufficiency because it allows one to diminish the potential indeterminacy of single-case studies a little bit. Imagine first that one aims to test for the sufficiency of an international debate about direct democracy. This requires the choice and comparison of two cases that have the debate in common and that check for the invention of elements of direct democracy. Since this applies to the two cases in Table 4.3, the condition 'international debate' passes the test for individual sufficiency. Regarding the other three conditions, presume that they represent rival explanations and that a majoritarian electoral system, a single-party government, and unitarism are hypothesized to be individually sufficient. What then does the MoA in Table 4.3 tell us about their individual sufficiency? The answer is 'nothing'. Each condition passes the test for individual sufficiency because the United Kingdom has the outcome in place. On the other hand, the Netherlands is wholly irrelevant because of the absence of all three conditions. The MoA hence is weak in terms of the testing of rival explanations for sufficiency.

When we extend the perspective beyond the sufficiency of individual conditions, it further becomes apparent that the pattern for the United Kingdom is compatible with a large number of more complex set-relational inferences that include equifinality as well as INUS conditions. For example, it could be that the three-way interaction of the institutional variables, any two-way interaction, and any institutional condition alone is sufficient for the invention of elements of direct democracy. Altogether, the MoA in Table 4.3 illustrates that this form of comparison has a high propensity to produce indeterminacy in analyses of sufficiency.[18]

The indeterminacy of the ideal MoA tends to get larger the more conditions one has because the size of the property space increases with each condition that is added to the analysis. Moreover, it holds that the weaker the theory is, the larger the extent of indeterminacy. In the previous example, we had specific expectations about what institution is sufficient for the invention of elements of direct democracy. Moreover, it was hypothesized that the presence of an international debate should lead to more direct democracy. The range of viable inferences is even larger if one hypothesizes that a factor is a condition but fails to specify what manifestation of the factor is sufficient for the outcome. Without attempting to provide a full list of the possible inferences here, it could be, for example, that an ongoing international debate about direct democracy forms an equifinal solution with the factor 'state organization', which takes the form of unitarism and federalism. Whether it is federalism or unitarism would be impossible to determine on the basis of the selected cases. It could also be that the combination of unitarism with a coalition government is sufficient or a majoritarian electoral system in a federal state. As these examples indicate and building on what was said before about correlational designs, weak theory contributes to the indeterminacy of cross-case comparisons.

Besides these problems and similarly to analyses of necessity, the discussion indicates that the ideal MoA is not necessarily the best cross-case comparison one can construct in an inquiry into sufficiency.[19] The goal of the MoA in Table 4.3 is to show that an international debate is sufficient. As explained in the previous paragraphs, adherence to the ideal MoA comes at the expense of eliminating rival conditions. The comparison in Table 4.4 illustrates an alternative to the MoA in Table 4.3. The Netherlands allow one to infer that an international debate about direct democracy is sufficient for the invention of more elements of direct democracy. At the same time, the examination of New Zealand, which is now being compared with the Netherlands, permits the inference that unitarism cannot be individually sufficient because we do not observe a trend to direct democracy in New Zealand.[20]

This comparison does not resemble the MoA anymore (it looks more like a suboptimal MoD, see below), but one should not to put the cart before the horse. Instead of considering what can be learned from Mill's methods, one should ask what the best comparison is for a test of a hypothesis (Savolainen 1994). The comparison in Table 4.4 might still leave a good deal of viable set-relational inferences untested. But such a comparison would achieve somewhat more than the MoA and would be particularly suitable if there are strong *ex ante* reasons to hypothesize that unitarism is individually sufficient.

Table 4.4 Test for sufficiency of an international debate about democracy and unitarism

Case	**More elements of direct democracy (Y)**	**International debate about more direct democracy**	**Election system**	**Government**	**Organization of state**
the Netherlands	Yes	Yes	Proportional	Coalition	Federal
New Zealand	No	No	Proportional	Coalition	Unitary

Having dealt with the MoA, the discussion now turns to the ideal MoD.[21] The MoD is the mirror image of the MoA and combines three features: two cases differ on the outcome and one cause and display invariant scores on all other causes. The simple logic behind the MoD is that differences must be explained with differences, which is again a compelling logic but pertains only to correlational case studies. This can be shown with the example of Zangl's study on the effect of the judicialization of international dispute settlement procedures on the compliance of a state with the outcome of these procedures (the rulings, so to say). Zangl (2008) performs four pairwise comparisons, each of which resembles the MoD. The comparisons are specifically concerned with the question of whether the United States complies with the dispute settlement procedures. Table 4.5 shows that one comparison covers disputes about hormone-treated beef. The independent variable of interest that varies across both cases is the degree of judicialization. It is high for procedures held under the auspices of the World Trade Organization (WTO) and low for procedures that took place within the context of the General Agreement on Tariffs and Trade (GATT). Three other variables that are included in the comparison are invariant: the dispute involved two powerful countries, namely the United States and the European Union, and the substance matter was identical – both conflicts centered on hormone-treated beef.[22]

The covariational logic behind the MoD is that judicialization is an independent variable because it covaries with compliance. The independent variables that are held constant can be discarded as causes in this analysis because invariant causes cannot account for variance of the outcome (de Vaus 2008, 254). A stronger inference is not possible unless one can credibly argue that the outcome is the result of monocausality, that is, that only one variable can produce change in the outcome. Since such a claim is hard to maintain in the social sciences (Bennett and Elman 2006; Franzese 2008), one has to allow for the possibility that the outcome is correlated with one of the other independent variables in another MoD. In other words, the MoD in Table 4.5 keeps the three control variables constant as a means to *control* them (de Vaus 2008, 254), which is not the same as elimination.

Table 4.5 Method of difference

Case	**Compliance of first country (Y)**	**Judicialization**	**First country powerful**	**Second country powerful**	**Substance matter**
WTO	Yes	Yes	Yes (USA)	Yes (EU)	Hormones-treated beef
GATT	No	No	Yes (USA)	Yes (EU)	Hormones-treated beef

One important caveat has to be attached to this conventional interpretation of the MoD because it is feasible only under the assumption of additive causal effects (Lieberson 1991, 312–13). If interaction effects cannot be ruled out, the cross-case pattern has to be interpreted more cautiously. It could be, for example, that the correlation between judicialization and compliance depends on the power of the first country or the power of the second country or the substance matter or any combination of these factors. In fact, it is likely that a powerful country is more likely to comply when it confronts another powerful country in the dispute settlement procedure (Zangl 2008, 840), the implication being that the ideal MoD is susceptible to indeterminacy even if only one independent variable covaries with the outcome.

This admonition is important to consider in light of the long-standing and repeatedly made claim that the MoD mimics the experimental ideal (Gerring and McDermott 2007; Lijphart 1971). If Table 4.5 is seen through the lenses of an experiment, case 1 is the treatment case and case 2 the control case, as judicialization – the treatment – is given in the WTO and not in the GATT.[23] Given that the two cases are similar in all other respects, the experimental template suggests that the difference in compliance is due to the difference in the extent of judicialization. The idea to resemble an experiment with the MoD is as appealing as it is misleading because qualitative case studies are, above all, observational designs. The key element of an experiment, the random assignment of the treatment to the cases, is not met. This is important to note because the intentional matching of cases on variables is not the same as the random assignment of cases to the treatment (Lieberson 1991; Zelditch 1971).[24] As a consequence of this, one confronts all the inferential problems that are attached to all kinds of observational research (Przeworski 2007). If one is lucky enough, the treatment is unrelated to other variables that have an effect on the outcome of interest, and one can refer to the design at hand as a natural experiment (Dunning 2008). However, the claim that one is dealing with a natural experiment should be based on careful reasoning and convincing empirical evidence. Whenever such a claim is not warranted, which is likely to be the rule and not the exception (Gerring 2007a, 172–3), one major problem in observational case studies is the possibility of interaction effects.

The benefits of the MoD for *set-relational* case studies are smaller compared with the MoA's. With respect to Zangl's example and Table 4.5, a test for the *necessity* of judicialization checks for the presence of the condition in the WTO case and its absence in the GATT case. Since the pattern of cross-case scores is in accord with a statement of necessity, one could infer that judicialization is necessary for compliance. However, if one theorizes that the presence of the condition 'powerful first country' and 'powerful second country' is expected to be individually necessary as well, the MoD is indeterminate because both countries are powerful in the WTO case, where we observe compliance (the

GATT case is irrelevant in this context). Uncertainty as regards the conditions 'powerful first country' and 'powerful second country' does not invalidate the inference that judicialization is necessary. However, there is no justification to argue that judicialization is *the* necessary condition.

The reason for the indeterminacy of the MoD lies in the fact that its construction is detached from theory. If theory predicts that the presence of a strong country is necessary for compliance, one should select two cases of compliance and test for the presence of the condition 'powerful first country'. An alternative pattern of scores that mirrors the MoD that would allow one to refute the necessity of 'powerful first country' (and 'powerful second country') is presented in Table 4.6. Since one can select cases only with respect to their scores on the outcome in research on necessity, all that one can do is to hope for cross-case scores on the conditions that permit the generation of inferences that are as unambiguous as possible.

Turning now to *sufficiency*, it can be easily seen that the MoD in Table 4.5 is not the best possible test of the hypothesis that judicialization is sufficient for compliance. One would not generate such an MoD because only the WTO case allows one to test this hypothesis. Moreover, the cross-case pattern of the WTO case is indeterminate because compliance is observable only when multiple purportedly sufficient conditions are present (assuming that the presence of a powerful first and second country are theorized to be individually sufficient). Although the GATT case does not allow one to test for the sufficiency of judicialization, this case has merit for the elimination of rival hypotheses (Mahoney 1999, 1158; Most and Starr 1989, 54–5). The GATT case allows one to reject the claim that having a powerful first and second country is sufficient for compliance.

Extending the perspective to conjunctural causation, one can infer that the two conditions are not jointly sufficient. Still, one cannot refute any influence of these conditions because they may produce the outcome in interaction with judicialization. In light of the WTO case, it could be that judicialization is individually sufficient, but it is equally justified to argue that a conjunction of 'judicialization' and 'powerful first country' is sufficient; or 'judicialization' and a 'powerful second country' or all three

Table 4.6 Alternative method of difference for set-relational case study

Case	Compliance of first country (Y)	Judicialization	First country powerful	Second country powerful	Substance matter
WTO	Yes	Yes	No	No	Hormones-treated beef
GATT	No	No	No	No	Hormones-treated beef

conditions combined; or that a powerful first and second country are INUS conditions because compliance could result when judicialization comes together with a powerful first country or a powerful second country.

One way to test the sufficiency of judicialization with a more suitable comparative design would require selecting cases on the basis of the conditions depicted in Table 4.6. The benefit of this comparison is that the WTO case now fully supports the inference that judicialization is individually sufficient. The conditions 'powerful first country' and 'powerful second country' can be ignored (not eliminated) as sufficient conditions because they are all absent. However, the GATT case is still a problem through the lenses of comparative hypothesis-testing because one cannot infer anything from it (except that one does not seem to have omitted a sufficient condition from the analysis). This example shows that, in constructing an MoD, one has to make a decision between two scenarios. First, as captured by Table 4.5, it cannot be shown that judicialization is individually sufficient, but the individual and joint sufficiency of the conditions 'powerful first country' and 'powerful second country' can be rejected. Second, it is possible to claim that judicialization is individually sufficient without being able to make any inferences about the rival conditions (see Table 4.6).

Mill's methods are the worst cross-case designs except…

The limitations of Mill's two well-known methods prompted many scholars to take a very pessimistic perspective on cross-case comparisons and inferences (Goldstone 1997; Goldthorpe 1997a; Lieberson 1994).[25] The abandonment of the two types of cross-case comparisons therefore seems to suggest itself. Before doing so, however, one has to ask what the alternatives are. The general logic behind the MoA – differences cannot explain similarities – and MoD – differences must be explained with differences – is quite intuitive (with the limitation that it is particularly accustomed to correlational designs). The simple aim is to avoid the problems of indeterminacy that derive from comparisons that deviate from the ideal comparison (Lijphart 1971).

The previous discussion of Trampusch's cross-case results was, among other things, implicitly about two suboptimal MoDs (and MoAs) because two potential causes covaried with the outcome.[26] If one has the choice between this pattern of cross-case scores and a perfect MoD where all but one variable are constant, one would always opt for the ideal comparison. The ideal MoD may still suffer from indeterminacy, but the range of viable causal inferences is always smaller than in an imperfect comparison. The imperfect MoD is always indeterminate, while the optimal MoD can be indeterminate, but to a lesser degree than an imperfect design. Similar arguments pertain to set-relational case studies, the MoA, and, in fact, all cross-comparisons that rely on many causes and few cases.

In view of these problems, the argument that the MoA and MoD should not be used for cross-case inferences entails the implicit claim not to perform cross-case comparisons at all (at least not those involving a small number of cases). The abandonment of cross-case comparisons cannot be the solution, however. Case studies that rely on extensive within-case analyses also aim to make inferences about whether a certain cause is relevant or not, which is the same as making a very basic cross-case inference (see Chapter 7 and Hall 2008). Even if we assume that no cross-case inference is intended, Chapter 3 shows that the cross-case level is vital to informed case selection for process tracing and that it presupposes knowledge about whether a cause is or is not relevant and, on a higher level, the nature of its causal effect. Unless one abandons cross-case analyses entirely for large-n methods such as regression analysis and QCA, there is no way to avoid cross-case inferences in case studies. The proper response is not to skip cross-case inferences but to admit the inferential uncertainty and report all theoretically viable inferences that are compatible with a specific pattern of cross-case scores.

4.3 Comparisons for the modification of hypotheses

Comparisons aiming at the modification of a hypothesis are built on the knowledge that a hypothesis test has produced a surprising result (Alvesson and Karreman 2007; Grofman 2001). A puzzling finding can have four sources: concept misspecification, measurement error, an omitted cause, and probabilism.[27] Of these four factors, only omitted causes are directly related to the formulation of a hypothesis and therefore are discussed in this section. Concept misspecification and measurement error are discussed in the section on general threats to cross-case inference below. A consideration of probabilism is postponed until Chapters 8 and 9 on generalization.

A hypothesis-modifying comparison starts out in full knowledge of the cross-case scores taken by the outcome and the causes. This knowledge is indispensable because the cross-case pattern forms the basis for case selection in process tracing, which in turn (probably) results in the modification of the hypothesis. In principle, this goal leaves it open whether one compares two anomalous cases or contrasts an anomalous case with a case that displays the expected result. In the following, ‘anomalous case’ is the general heading for deviant cases, failed-most-likely cases, and passed least-likely cases. The term ‘expected case’ subsumes typical cases, passed most-likely cases, and failed least-likely cases because they exhibit the expected cross-case patterns. Regardless of what types of cases are examined, the comparison of an expected and an anomalous case is preferable to a comparison of two anomalous cases because the goal is to learn something about the reasons for the anomaly. This can be best achieved by looking at a case that conforms to the expectation and one that does not. The precise implementation of a comparison between an expected and an anomalous case hinges on whether one is interested in a correlation, necessity, or sufficiency.

Correlational comparisons

The puzzle of a correlational case study lies in the absence of an expected covariation. Such a puzzle can take three forms.[28] First, the independent variable varies, but the dependent variable is invariant. In the context of hypothesis modification, I label this a *no-variance-on-Y comparison* so as to denote what the surprising cross-case feature is. Second, the outcome varies even though all independent variables are invariant, which is called the *variance-on-Y comparison*. Third, the observed correlation is the opposite of the expected correlation, which is dubbed the *inverse-correlation comparison*.

Lange's study on the development of former British colonies (2009) is motivated, among other things, by a no-variance-on-Y puzzle. The analysis considers whether the countries with direct rule – Mauritius and Guyana – experienced better economic development than the two former colonies with indirect rule – Sierra Leone and Botswana. Table 4.7 shows that this is the case when one compares Mauritius and Sierra Leone but not when one compares Guyana and Botswana, for example.[29]

One explanation for this puzzle could be that the correlation in the first pair of cases is coincidental. In fact, spuriousness is one potential reason for most of the puzzles that are addressed in the remainder of this section. For the analysis in Table 4.7, process tracing then should aim to identify an omitted and, of course, theoretically intelligible variable that is correlated with the outcome.

However, provided that the form of rule can be claimed to be causal for economic development, the goal must be to identify an omitted variable that interacts with the form of British rule. Such a variable must meet two requirements: it must be invariant for each pair of cases and must vary across the two pairs. The last column includes the variable ethnic homogeneity, which in this example is assumed to be high for Mauritius and Sierra Leone and low for Guyana and Botswana.[30] Assuming that one can rule out an independent effect of ethnic homogeneity, it then holds that the form of rule only matters when ethnic homogeneity is high.

The characteristic of a *variance-on-Y* comparison is the observation of variance on the dependent variable, even though all independent variables are

Table 4.7 No-variance-on-Y comparison

Type of pair	Country	Economic development (Y)	Form of British rule	*Ethnic homogeneity*
Expected	Mauritius	Good	Direct	*High*
	Sierra Leone	Bad	Indirect	*High*
Anomalous	Guyana	Bad	Direct	*Low*
	Botswana	Bad	Indirect	*Low*

Table 4.8 Variance-on-Y comparison

Type of pair	Case	Successful deterrence (Y)	Power asymmetry	*Perception of determination*
Expected	1	No	No	*No*
	2	Yes	Yes	*Yes*
Anomalous	Fashoda I	No	Yes	*No*
	Fashoda II	Yes	Yes	*Yes*

constant. This comparison mirrors what Mill (1874, 284–5) introduces as the *method of residues* (MoR). The logic of the MoR is that there must be an omitted variable that correlates with the outcome if all observed variables fail to account for the variance. An empirical example that resembles this type of comparison is Schultz's analysis of the Fashoda crisis (Table 4.8). Schultz (2001, chap. 6) is concerned with the involvement of democracies in international crises. He argues that the trajectory of an international crisis depends on whether a democracy can credibly deter its opponent and that the credibility of the signal depends on whether the domestic opposition supports the government. Schultz selects the case of the Fashoda crisis in order to test the argument that a supportive opposition accounts for credible deterrence. The Fashoda crisis involved Great Britain and France; the two countries were close to war, but the crisis terminated peacefully because France eventually backed down.

Schultz argues that the Fashoda case represents a puzzle for explanations focusing on the power balance between the countries involved in a conflict. Symmetric power relations are expected to be conducive to an escalation, while power asymmetries should lead to peaceful conflict resolution because the weaker state knows that an armed conflict would be devastating. The first pair of hypothetical cases in Table 4.8 confirms this line of reasoning. The anomalous pair of cases that constitute the Fashoda crisis contradict this line of reasoning (Schultz 2001, 177–8). In Table 4.8, Fashoda I stands for the first part of the crisis, during which neither France nor Great Britain backed down. Fashoda II denotes the second part of the crisis, at the end of which Great Britain gained the upper hand. This development cannot be explained with a change in the distribution of power because Great Britain was more powerful throughout the entire crisis. From the perspective of an explanation based on power asymmetries, the Fashoda crisis thus represents a puzzle because one observes an invariant cause and a varying outcome across the two cases.

Schultz proposes an alternative explanation and argues that France's perception of Great Britain's determination changed in the course of the crisis. France initially believed that Great Britain was bluffing and became later convinced that its opponent was determined, which eventually prompted France to back down (Schultz 2001, chap. 6). Adding the variable 'perception of determination', measuring whether a country perceives the opponent to be domestically united, resolves the puzzle because this variable covaries with the outcome. In effect, the inclusion of this variable turns a variance-on-Y comparison into the method of difference.

In light of the previous discussion about indeterminacy and the method of difference, one should carefully evaluate the causal inferences that the new pattern of cross-case scores allows one to make. Two reasons speak for this. First, the new variable may interact with the invariant one. In fact, Schultz (2001, 195) presents a quote of Joseph Chamberlain, a then member of the British government, that speaks against an individual effect of Britain's determination. Chamberlain remarks that Britain's success is attributable 'as much to the spectacle ... of an absolutely united people as it was to those military and naval armaments' of Britain (177–8). This quote suggests that unequal and high levels of armament played a role as well. Since this rival explanation cannot be examined with the Fashoda crisis, one would need to perform a separate test with the variable 'power asymmetry' taking different scores.

Second, the inclusion of the new variable requires one to reconsider the causal inference on the original variables in the expected pair of cases. For the empirical example, this means that the new variable 'perception of determination' might shed a different light on the inference that a power asymmetry is the cause of successful deterrence in the first pair of cases. With respect to the pattern in Table 4.8, the inclusion of the new variable leads to indeterminacy because two variables covary with the outcome now. A similar problem results when the new variable is invariant for the first pair of cases. This pattern of cross-case scores would raise the question of whether the correlation between power asymmetry and the outcome depends on the variable 'perception of determination' taking specific invariant scores across the two cases. Again, these competing causal inferences can be empirically evaluated only in a follow-up case study that is specifically designed to test them.

A third form of covariational comparison is motivated by the observation of an unexpected correlation. This probably not so common form of comparison can be coined the *inverse-correlation comparison*. Studlar's analysis of tobacco policy in the United States and Canada is an example of such a comparison (2010). Studlar aims to explain why the central government has been dominant in tobacco regulation in Canada, whereas the states were the main actors in the United States. This is striking because Canadian federalism can be described as decentralized and US federalism as centralized (391). If one takes the general level of decentrality as a cause of the jurisdictional level in charge of tobacco policy, one observes an unexpected correlation because there is no apparent reason why the relationship should be reversed in the case of tobacco policy. In Table 4.9, the expectation is captured by the hypothetical countries 1 and 2 and the inverse correlation

Table 4.9 Inverse-correlation comparison

Type of pair	**Country**	**Jurisdiction in tobacco policy (Y)**	**Federal system**	**Anti-tobacco groups**	***Cigarette industry***	***Government ideology***
Expected	1	Centralized	Centralized	Strong	Weak	*Left*
	2	Decentralized	Decentralized	Strong	Weak	*Right*
Anomalous	Canada	Centralized	Decentralized	Strong	Strong	*Left*
	USA	Decentralized	Centralized	Strong	Strong	*Right*

by Canada and the United States. The inverse relationship between the centralization of the federal system and the jurisdiction in charge of tobacco policy cannot be explained with reference to rival variables such as strength of anti-tobacco groups because they are identical for Canada and the United States (Studlar 2010, 389–90).

There are two ways in which the puzzle of an inverse correlation can be removed. In contrast to all other puzzles discussed in this section, the first variant is not related to the comparison but to the theory on which the comparison is based. There might be nothing wrong with an inverse correlation from an empirical point of view, meaning that it exists empirically. The problem that one has to address instead is rooted in the theory that fails to make sense of the inverse correlation. Consequently, the goal must be to find and assess empirically a compelling explanation for why the correlation between cause and outcome is the opposite of the expected association. The second possibility is that the degree of centralization in a federal system is unrelated to the level of centralization in tobacco policy. In this instance, one should search for an omitted variable, such as government ideology, that displays the same correlation in the expected and anomalous pair of cases.[31] Without going into the details here, it again holds that the inclusion of a new variable should prompt one to reconsider the original pattern of cross-case scores.

The arguments on correlational comparisons seeking to solve puzzles are summarized in Table 4.10. Each of the three types of comparisons is described with respect to its defining feature, potential theory-related sources for the puzzle, and the corresponding remedy.

Set-relational comparisons

Hypothesis-modifying comparisons that are embedded in a set-relational framework aim to shed light on why an expected set relation does not hold in a specific case. Owing to the nature of set relations, one can distinguish three variants of comparisons that address a set-relational puzzle. In case studies on necessity, a *presence-of-Y comparison* seeks to understand why the outcome is present in the absence of the necessary condition. Two

Table 4.10 Correlational comparisons for modification of cross-case hypothesis

Comparison	**Feature**	**Source of puzzle**	**Remedy for puzzle**
No-variance-on-Y	Varying X and invariant Y	Spurious correlation	Add covarying variable
		Interaction effect	Add invariant variable
Variance-on-Y	Varying Y and invariant X	Omitted variable	Add covarying variable
Inverse correlation	Unexpected covariance of X and Y	Weak theory	Explain unexpected correlation
		Spurious correlation	Add covarying variable

Table 4.11 Presence-of-Y comparison for necessity

Type of case	**State**	**Low level of inequality (Y)**	**Early childhood education available**	***Left government***	***Entry test***
Expected	Berlin	Yes	Yes	*Yes*	*No*
Anomalous	Hesse	Yes	No	*Yes*	*Yes*

comparisons are available in case studies on sufficiency: an *absence-of-Y comparison* asks why the outcome is absent when the sufficient condition is given, while a *presence-of-Y comparison* asks why the outcome is observable when all sufficient conditions are absent.

A *presence-of-Y comparison* should be the goal for case studies interested in necessity because the only puzzle that can motivate a comparison is the presence of the outcome in the absence of a necessary condition. Freitag and Schlicht's analysis of educational federalism in Germany (2009) can serve as an example for this type of comparison. Freitag and Schlicht find that a high availability of early childhood education is necessary for a low social inequality of education. Yet the pattern of cases is not fully consistent with one of necessity because some German states deviate from it. For instance, Hesse has an unexpectedly low level of social inequality in relation to the availability of early childhood education. Hesse thus is an anomalous case that could be contrasted with a state such as Berlin, which conforms to the pattern of necessity (Table 4.11).

The solution to this puzzle can be twofold. First, the proposition may be wrong and the availability of early childhood education not necessary. It may be that the empirical phenomenon of interest does not have any necessary conditions, but it could also be that an omitted condition is necessary. The omitted condition should be invariant across both cases because otherwise it does not qualify as necessary. In Table 4.11, this constellation is captured by the condition 'left government'.[32] Second, early childhood education and an omitted necessary condition could be functional equivalents (Ragin 2000). A functional equivalent for this condition could be entry tests that are performed when children enter the school and determine the assignment of children to catch-up courses. Let's call such a condition 'entry test' and suppose that Hesse performs such tests while Berlin does not. The inclusion of the condition 'early test' does not introduce indeterminacy *ex post* because early childhood education and early tests are not simultaneously observable. If one can make a credible point that these two conditions are functionally equivalent and the observed pattern is as in Table 4.11, the inclusion of the condition 'early test' solves the puzzle.

However, the search for an omitted functional equivalent does not necessarily have this implication because two conditions can be functional equivalents and simultaneously present in the expected case. In Table 4.11, this would mean that there would be no empirical basis for discriminating between the inference that early childhood education is spurious and that it is a functional equivalent of entry tests. If one aims to discriminate between these two inferences, one should perform a hypothesis-testing case study with one case (or more) for which the outcome is present and the newly discovered condition is absent ('early test' in Table 4.11). If early childhood education is present, one could infer that the two conditions are functional equivalents.[33]

In case studies interested in sufficiency, the *absence-of-Y comparison* is driven by the observation that the outcome is not observable, even though the sufficient condition is present. Table 4.12 draws on a study by Drezner (2000) as an illustration for the absence-of-Y comparison. Drezner notes that multilateral economic sanctions are sometimes effective and sometimes not. Since the absence of the outcome is puzzling when the condition is present, there must be something missing from an explanation of effective multilateral sanctions.

Puzzles like the ones identified by Drezner can be solved in two ways. First, the supposedly sufficient condition is spurious and wholly irrelevant

Table 4.12 Absence-of-Y comparison for sufficiency

Type of case	**Case**	**Economic sanctions effective (Y)**	**Multilateral cooperation**	***International organization involved***
Expected	1	Yes	Yes	*Yes*
Anomalous	2	No	Yes	*No*

Table 4.13 Presence-of-Y comparison for sufficiency

Type of case	Country	Low inequality (Y)	Type of union	*Organization of union confederations*
Expected	Sweden	Yes	Strong	*Fragmented*
Anomalous	Italy	Yes	Weak	*Inclusive*

for the outcome. In this instance, process tracing should aim at a condition that is present in case 1 and absent in case 2 so as to be able to explain the difference in the outcome. Second, the alternative is to search for an interaction effect between the original variable and an omitted one. Indeed, Drezner claims and finds evidence showing that multilateral cooperation is effective only if the sanctions are enforced by an international organization. In order to solve the puzzle, the minimum requirement is that the omitted condition 'international organization involved' must be observable in case 1 and absent in case 2 (last column in Table 4.12).

As holds true for the previous designs, the inclusion of a condition may introduce indeterminacy *ex post*. This is not an issue for the empirical example because the omitted condition cannot be sufficient on its own; if there is no multilateral cooperation that could be enforced, there is also no role that an international organization could play. In another analysis, however, it could be that the new condition is capable of producing the outcome on its own. Again, the only empirical way to discriminate between the two scenarios would be to select a case with the new condition present and the original condition absent. If the outcome is observable, the new condition would seem individually sufficient, whereas conjunctural causation would be in evidence if the outcome is absent.

Another puzzle that one may encounter in the analysis of sufficiency is the presence of the outcome in the absence of the purported sufficient conditions. I refer to this as a presence-of-Y comparison for sufficiency in the following. This comparison may seem surprising because it is concerned with the presence of the outcome and not with the presence of a condition. However, the presence of the outcome can be taken as a puzzle if all alleged sufficient conditions are absent and the case displays the outcome nevertheless. An example for a presence-of-Y comparison is Oliver's analysis of earnings inequality in Sweden and Italy. Sweden and Italy had about the same relatively low level of earnings inequality in the late 1990s. This is a puzzling finding because, then as now, labor unions were strong in Sweden and weak in Italy.[34] If 'strong unions' is taken as a sufficient condition for low inequality (and assuming that other possible sufficient conditions are absent for Italy), one can run a presence-of-Y comparison to address this puzzle (Table 4.13).

Process tracing should search for an omitted condition that is present in the anomalous case because otherwise the puzzle would persist. One such condition could be the organization of the union confederations because different forms of organization imply different capacities to negotiate wages and address wage inequalities (Oliver 2011, 558–9). In Sweden, one finds fragmented confederations because blue-collar and white-collar workers are organized in different unions. Italian unions are inclusive and include blue-collar and white-collar workers alike, meaning that the inclusion of the condition 'inclusive unions' is able to solve this puzzle.

Again, it is important to consider the cross-case score that the new condition assumes for the expected case. If the pattern is as in Table 4.13, one can conclude that strong unions and inclusive unions are equifinal (interaction effects are left aside here). This would be different if unions had turned out to be inclusive in Sweden and Italy. Such a pattern would make it impossible to distinguish between equifinality and the inference that strong unions are irrelevant and inclusive unions alone sufficient. This can be decided only in a follow-up study that checks whether the outcome is also observable in a country with strong and fragmented unions.

The discussion of set-relational comparisons aiming to modify a hypothesis is summarized in Table 4.14.

Table 4.14 Set-relational comparisons for modification of cross-case hypothesis

Type of set relation	Comparison	Feature	Source of puzzle	Remedy for puzzle
Necessity	Presence-of-Y	Y is present in the absence of X	Spurious necessity	Add invariant condition
			Omitted functional equivalent	Add condition that is present in anomalous case
Sufficiency	Absence-of-Y	Y is absent in the presence of X	Spurious sufficiency	Add invariant condition
			Omitted condition interacting with included condition	Add condition that is present in expected case and absent in anomalous case
	Presence-of-Y	Y is present in the absence of all plausible X	Omitted sufficient condition	Add condition that is present in anomalous case

In total, the discussion of set-relational and correlational comparisons shows that they are more intricate than they might seem at first sight. Depending on the design at hand, there might be more than one way to resolve the puzzle that motivates the comparison. Moreover, the inclusion of a new cause potentially introduces indeterminacy *ex post*. Because of these problems, it is important to carefully consider the cross-case pattern after a hypothesis has been modified and to report all the causal inferences one can read into it. This is a valuable service to researchers who aim to test propositions by drawing on the insight of a hypothesis-modifying case study.

4.4 General threats to cross-case inference

So far, the discussion of cross-case comparisons has focused on the inferential problems of various ideal–typical and suboptimal designs. The main message is that the generation of cross-case inferences is protracted even before additional complicating issues are introduced. Two such issues that deserve attention and have frequently been addressed in the literature are *omitted causes* and *measurement error* (Goldthorpe 1997a, 7; King et al. 1994, 208–13; Lieberson 1991). The previous section showed that omitted causes are one reason for puzzles that drive hypothesis-modifying comparisons. The same effect can result from measurement error, which can take the form of a misspecified concept, indicators with low reliability and/or validity, and the inappropriate use of sources (Adcock and Collier 2001; Munck and Verkuilen 2002). However, the following discussion shows that the consequences of omitted causes and measurement error are not limited to case studies seeking to modify a hypothesis. Equally important to note is that they do not necessarily have detrimental consequences on cross-case comparisons.

The implications of measurement error for cross-case comparisons are obvious. Let's assume that one has constructed an ideal–typical MoD but with a wrongly measured outcome. More precisely, the outcome turns out to be invariant when the measurement error is removed. What seems to be an ideal–typical MoD is therefore a suboptimal MoA because the outcome and multiple causes share the same scores across the two cases. More generally seen, mismeasurement of a single cause or the outcome suffices to turn an optimal comparison into an imperfect one. At the same time, however, this implies that a suboptimal design involving measurement error could be an ideal–typical one once it is implemented free of measurement error.

The second threat to cross-case inference is an omitted cause. The omission of a cause can lead one to investigate a puzzle, but may also undermine causal inference in case studies that build or test a hypothesis. Let's assume that one has constructed an ideal–typical MoD for hypothesis testing and that someone can make the credible claim that a cause has been omitted. In contrast to measurement error that automatically undermines an optimal comparison, neglecting a cause is not necessarily a problem. Taking the example of the MoD, the omitted cause may vary across the two cases and turn the design from an optimal comparison into a suboptimal one if it were to be included. However, the omitted cause may also be invariant and may not have an adverse effect on the generation of cross-case inferences (except that it may increase the range of viable interaction effects). Evidently, which of the two scenarios holds can be known only after the omitted cause is included in the comparison.

The need to protect the analysis against the claim of omitted causes may lead one to include a whole range of causes that are potentially relevant according to the substantive literature underlying the case study. However, the probability of constructing a suboptimal comparison naturally increases with an increasing number of causes. There is therefore a trade-off between the problem of omitted causes, on the one hand, and the problem of suboptimal comparisons, on the other hand (Lijphart 1971). Thus, there is no easy way to salvage one's comparison against the claim of omitted causes because this is likely to come at the expense of creating an imperfect comparison.

4.5 Conclusion

Cross-case comparisons have been subject to debate for decades now. Although not all arguments in this chapter are new, I attempted to give a comprehensive and systematic elaboration by explicit reference to the covariational and set-relational view on causation. The chapter confirms the conclusion made by Lieberson (1994, 1991) and others (Zelditch 1971) that causal inference in small-n comparisons rests on a set of demanding assumptions. These problems plague correlational and set-relational comparisons regardless of whether they rely on Mill's methods or any other pattern of cross-case scores. On the cross-case level, it is accurate to refer to covariational case studies as the 'statistical method writ small' (Hall 2008, 308) and to set-relational designs as 'Qualitative Comparative Analysis writ small' with all the associated problems that have been discussed in this chapter.

The problems of arriving at internally valid cross-case inferences pertain to all three types of research purposes. While this insight is not new for hypothesis-testing case studies (Lieberson 1991), I have shown that it is equally important to take into account in exploratory case studies because hypothesis-building and hypothesis-modifying analyses are supposed to lay the best possible basis for subsequent tests of hypotheses. At the end of the case study, it is therefore mandatory that one makes explicit all propositions that are compatible with the collected evidence. The hypotheses should be phrased as unambiguously as possible so as to give the strongest guidance to researchers who aim to perform a hypothesis test.

5 Enhancing Causal Inference in Comparisons

The previous chapter showed that cross-case comparisons tend to be plagued by manifest problems connected with the generation of unambiguous causal inferences. Being cognizant of these problems, the small-n literature made several recommendations that aim at improving cross-case comparisons (King et al. 1994, 208–13; Lijphart 1971, 686–90). The present chapter serves to discuss the potential of these and additional instruments for the enhancement of cross-case inferences. It starts with the problem of suboptimal comparisons and relates them to the multidimensional nature of cases. It is shown how one can invoke these dimensions in the construction of comparisons in order to approach as closely as possible the envisaged ideal design.

The next sections specifically focus on the size of property space and present four instruments for the improvement of suboptimal and ideal comparisons. If the problem is one of a large property space and few cases, two solutions that suggest themselves are a smaller property space and a larger number of cases. The degree to which more than two cases can diminish the inferential intricacies is elaborated in Section 5.2. The three remaining tools for enhanced cross-case comparisons all aim at a reduction of the property space.[1] In Section 5.3, I discuss the role of the level of measurement aggregation. The level of measurement aggregation pertains to the number of categories on which causes and/or outcomes are measured, the basic distinction being between bi- and multicategorical measurement of causes and/or the outcome. An additional tool for reducing indeterminacy is strong theory. In Section 5.4, I elaborate what strong theory is in qualitative case studies and how it can contribute to causal inference. The fifth and final instrument, discussed in Section 5.5, has been largely neglected so far and pertains to the transformation of causes into scope conditions.

5.1 Units of analysis and time: Comparability v. generalizability

Every case can be located on a temporal and substantive dimension as well as on at least one additional dimension (see Section 2.1). In this section, I highlight the role of these dimensions for comparisons and argue that each of these dimensions is characterized by a *trade-off* between the comparability of cases and the generalizability of cross-case inferences. In order to simplify the following discussion at its onset, I start with a distinction between the temporal dimension and spatial dimension and introduce other dimensions only later in the discussion. As the following discussion shows, the temporal dimension enjoys a special status in comparisons, while all other, nontemporal dimensions have equivalent implications. The different implications of the spatial and nonspatial dimensions therefore justify treating the former on its own terms and taking the spatial dimension as illustrative for other dimensions. Before the functions of the dimensions are addressed, it is first necessary to introduce the concept of a unit of analysis and relate it to the dimensions on which a case is described.

Units of analysis and cases in case studies

The unit of analysis is often described as the object of interest in an empirical analysis. If one hypothesizes that welfare state spending of OECD countries increases with increasing economic openness, the units of analysis are OECD countries because the hypothesis makes a prediction about their spending behavior (Johnson et al. 2007, 77–8). Though intuitively plausible, the description of units as the objects of analysis is misleading because the empirically relevant object is the case. A case is fully described by boundaries on at least three dimensions (the temporal, the substantive, and a third one), therefore offering full information about what is examined in the empirical analysis. In contrast, knowing the unit of analysis is not very illuminating because a huge number of cases and research questions are compatible with the same unit. In the given hypothesis, we assumed that the units of analysis are OECD countries. But the information that a case study is about OECD countries provides little insight because the question as to the interesting feature of the OECD countries is left answered. It could be welfare state spending, the share of women in the national parliament, the stability of governments, the illiteracy rate, the share of children in public childcare, and so on.

Moreover, the multidimensional character of cases suggests that there is not something like *the* unit of analysis, but that every case can be assigned to multiple units. Consider the hypothesis ‘welfare state spending of OECD countries increases with increasing economic openness’. Arguably, most researchers would take OECD countries as the unit of analysis. Now consider the hypothesis ‘welfare state spending increases with increasing economic openness’, which is the same as the first except for the reference to OECD countries. However, it is evident that both hypotheses can be tested with exactly the same design, cases, and data. Since the second hypothesis does not designate the territorial unit of interest, the unit of analysis cannot be defined in spatial terms. Instead, the second hypothesis implies that the welfare state is the unit of analysis because that is the entity emphasized by the hypothesis. The lack of reference to the territorial entity OECD countries is therefore not accidental but serves to highlight that the interest lies on the welfare state. Neither of the two definitions of the unit of analysis would be right or wrong; they simply denote a different perspective on a similar issue: the effects of globalization on spending.

On the basis of these arguments, it is now possible to clarify the relationship between units of analyses, on the one hand, and the multidimensionality of cases, on the other. The dimensions on which a case is described give one a broad sketch about the nature of that case. For example, when a case is described by a temporal, substantive, and spatial dimension, we know that the analysis concerns a territorial entity rather than an institution or organization (see Chapter 2). However, without the imposition of bounds on each underlying dimension, the nature of the case of interest remains opaque.

As I understand them in the following, units of analysis are closely related to the boundaries of a case and its nature as a multidimensional phenomenon. Imagine that we take radical welfare state retrenchment in New Zealand in 1991 as the case in an analysis of the determinants of radical retrenchment. In terms of dimensions, the description of the case implies that it is located on a territorial, temporal, and substantive (the institutional dimension is left aside). The unit of analysis is now understood as the place a case takes on a *single* dimension owing to the specification of boundaries. This means that there are always multiple units of analysis involved in the analysis of a single case. With respect to the territorial dimension, New Zealand is the spatially defined unit of analysis in the given example. This is relevant information because it is more specific than a reference to the spatial dimension alone. However, it is still incomplete because it is not clear what matters about New Zealand. The additional information is provided by the specification of the unit of analysis on the other two dimensions. On the temporal dimension, the unit of analysis is simply the year 1991, whereas radical welfare state retrenchment is the substantively defined unit of analysis. A case therefore is the intersection of all units of analysis into which it can be decomposed by taking a multidimensional view on cases.

Having clarified what the unit of analysis is, I now focus on the role of the unit of analysis for comparability and the generalizability of cross-case inferences. One can distinguish two nonexclusive strategies of realizing cross-case comparisons (Gerring 2004, 343).[2] First, one can compare cases within and across units (Smelser 1973, 63) and, second, at the same point in time – cross-section comparisons – and over time – longitudinal comparisons (Bartolini 1993). The two strategies are now discussed in turn.

Cross-unit and within-unit comparisons

The first choice one has to make when constructing a comparison is between cases that belong to the same unit and those that belong to different units of analysis. The rationale for a within-unit comparison is that cases belonging to the same unit are likely to be more comparable than cases from different units (George and Bennett 2005, 166). Jakobsen's (2010) analysis of the timing and content of liberalization in utilities sectors exemplifies the benefits of a within-unit comparison. Jakobsen's goal is to test hypotheses that predict an effect of Europeanization, globalization, and national politics on the trajectory of liberalization in utilities sectors. He takes Denmark as the spatial unit of analysis and compares liberalization in the telecommunication and electricity sector. On the spatial dimensions, this is advantageous in respect to the control of relevant factors such as the type of political regime and growth of the GDP because they are similar. However, the enhanced comparability comes at the expense of a somewhat increased uncertainty of generalization across units. Jakobsen's insights hold for two utility sectors in Denmark, but it is not known if confirming evidence is available for countries other than Denmark.

A cross-unit comparison does not suffer from this problem in the same way if one can find two countries that lend themselves to a comparative analysis of liberalization in the utilities sector. In contrast to a within-unit design, a cross-unit analysis slightly increases confidence in the veracity of the generated inference across units. The advantage of a cross-unit comparison hinges on the condition that the cases from different units are comparable to the degree of cases belonging to the same unit. Whether one can find such cases is an empirical question, but it is likely that a cross-unit analysis results in a suboptimal comparison. Consequently, the choice between a within-unit and cross-unit comparison tends to be characterized by a trade-off between unit-related comparability and generalizability.

In order to demonstrate the consequences of the multidimensional nature of cases for comparisons, additional insight can be gained by taking a second dimension into the picture. In Jakobsen's study, the telecommunication and electricity sectors represent two different substantive units of analysis.[3] This means that Jakobsen's design entails a spatial within-unit analysis and a substantive cross-unit analysis. With respect to the trade-off between comparability and generalizability, the analysis therefore leans toward comparability on the spatial dimension and generalizability on the substantive dimension.

This example shows that a researcher faces the trade-off on each dimension on which a given case is located. Moreover, it indicates that the decisions to be made for each dimension are independent of one another. When the highest value is attached to comparability, one has to perform a within-unit comparison on each dimension. On the other hand, generalizability is maximized when one realizes cross-unit comparisons on all dimensions. In between these extremes, one can construct a mix between comparability and generalizability by combining a within-unit comparison on one or more dimensions with a cross-unit perspective on one or more other dimensions.

Cross-section and longitudinal comparisons

The second fundamental dimension that underlies comparisons concerns time. The two strategies that are on offer are a cross-section comparison, which contrasts cases at similar points in time, and a longitudinal comparison, which contrasts cases over time. In a cross-section design, comparability is more difficult to achieve than one may think at first sight because one can distinguish two subvariants. In the first variant, one focuses on what can be called *chronological time*. This design entails that cases are compared at exactly the same year, month, or whatever time unit matters for the research question at hand. The second variant emphasizes what can be referred to as *theoretical time*. This form of cross-section comparison relates the temporal dimension of a comparison to the theory and hypotheses that are under scrutiny, which does not necessarily mean that a comparison at exactly the same point in time is the best option.

As an example for a comparison centered on chronological time, consider Eckert's analysis of delegation in the postal sector in France, Germany, and the United Kingdom (2010). The comparative case study takes a cross-section perspective because in all three countries the delegation processes took place in the 1990s and early 2000s. Thus, the three cases cover the same period of time and are influenced by the same events, such as an EU directive in the postal sector.

Eckert's study arguably relies on the classic version of cross-section designs. Depending on the research question, though, cross-section comparisons at the same point in time may in fact undermine the comparability of cases. In this instance, a focus on theoretical time can be the better variant of a cross-section comparison. Lange's analysis of former British colonies can be taken as an example for such a comparison. His comparative case study is concerned with the development of former British colonies and includes Botswana, Guyana, Mauritius, and Sierra Leone (2009). When one takes the year in which these colonies became independent as a decisive event, a comparison at exactly the same point in time is not always feasible because the four colonies gained independence over the course of several years. For instance, Sierra Leone became independent in 1961 and Mauritius in 1968, and so it is impossible to compare the two states as independent countries at any point in time before 1968. A comparison at a later point in time, say 1970, is also not without its problems because Sierra Leone had experienced nine years of independence by this time and Mauritius only two. The seven years that set the two countries apart may make a difference for the case study as Sierra Leone is likely to have been a more consolidated country at this stage and already undergone some economic development.

This problem can be taken into account by crafting a comparison accordingly. For instance, one could compare Sierra Leone and Mauritius ten years after each country became independent, that is, one would contrast Sierra Leone in 1971 with Mauritius in 1978. Since both countries had ten years to develop and consolidate, they are similar in this respect and the comparison would allow one to analyze differences in the degree of development. However, the downside of this cross-section comparison is apparent because a series of economic and political events, for example, the first oil crisis in 1974, happened between 1971 and 1978. Such an event may have affected the development of Mauritius but is necessarily irrelevant for Sierra Leone because it happened after 1971. In an attempt to handle this problem, one could include a control cause 'oil crisis' and score it as absent for Sierra Leone and present for Mauritius. But this is the very problem of a cross-section comparison centered on theoretical time because every additional cause tends to make the generation of cross-case inferences more protracted (see Chapter 4).

This example illustrates that cross-section designs are characterized by a trade-off between comparisons at exactly the same point in time and comparisons at points in time that are comparable in light of the theory at hand. In an ideal constellation, it is possible to find cases that are comparable in both respects. For the empirical example, this would call for the analysis of two colonies that became independent in the same year and that are compared at the same point in time. When such a comparison is not feasible because nature does not provide one with the ideal cases, one should carefully consider which subvariant of cross-section design is more appropriate for the research question at hand.

In contrast to both types of cross-section comparisons, a genuine longitudinal comparison compares two cases over time. This form of comparison is also known as a *before–after design* and *interrupted time-series design* (Collier 1993; George and Bennett 2005, 166). In accordance with the experimental ideal (Gerring and McDermott 2007), a longitudinal comparison conceives of a change of a potential cause as equivalent to the treatment in an experiment. The change distinguishes two cases, one of which falls into the prechange (pretreatment) period and one into the post-change (posttreatment) period.[4] The opportunity to trace a cause over time and observe change is what sets the temporal dimension of a case apart from all other dimension on which it is located. From the viewpoint of causal inference, the appealing feature of longitudinal comparisons is that they are necessarily within-unit comparisons that exhibit the advantages and disadvantages elaborated above. The two (or more) cases are likely to be comparable, as they belong to the same unit, but at the expense of increased uncertainty about the generalizability of causal inferences.

The enhanced comparability of cases in longitudinal comparisons explains why before–after designs are often deemed to be a very attractive form of comparison, if not the most attractive one (Caramani 2010; Collier 1993, 113; George and Bennett 2005, 166–7; Gerring and McDermott 2007; Lijphart 1971, 689).[5] While within-unit comparisons are appealing, three qualifications are in order when it comes to the generation of causal inferences. First, the influence of a cause might be discernible only in the long run. A long-run effect might be in place because the cause has a cumulative effect, or because the outcome is characterized by a cumulative process, or for both reasons simultaneously (Grzymala-Busse, 2011; Pierson 2004, chap. 3). The compensation hypothesis can be taken to illustrate the idea of a cumulative cause. Imagine that economic liberalization is hypothesized to increase welfare state spending. In the empirical analysis, one observes that a country reduces its tariffs and liberalizes the capital market, that is, it moves from a closed to an open economy. Economic liberalization does not immediately produce an increase in welfare state spending, though; it happens only in the long term because the competitive pressure on domestic producers and the employees' demand for compensation accumulates over time. At some point in time that could be distant from the actual move to an open economy, the government then decides to respond and starts to compensate employees.

This example points to important ramifications of cumulative causes and outcomes for comparison because it is necessary to choose an appropriately long period of analysis.[6] When one takes a short-term perspective in an analysis of higher welfare state spending, one may attribute the increase in spending to proximate causes such as massive lobbying of labor unions and domestic producers. In contrast, a long-term perspective would show that massive lobbying is attributable to increased competitive pressure, which in turn derived from economic liberalization a long time ago. The point is that the longer the period of analysis is due to cumulative causes and/or outcomes, the more likely it is that multiple potential causes change over time and create an indeterminate pattern of cross-case scores. The degree to which this is an issue in a case study depends on the subject matter but tends to be particularly pertinent for comparative historical analysis because of its interest in long-run processes, cumulative causes and outcomes, and periods of analysis covering dozens or even hundreds of years (Mahoney and Rueschemeyer 2003a).

A second, related aspect that complicates longitudinal comparisons hinges at least partially on the research question. It concerns the proper choice of a negative case for comparison (Mahoney and Goertz 2004), which may be a case that is temporally distant from the positive case. Suppose one wants to explain why countries adhere to the most favored nation (MFN) principle in international trade cooperation. Put simply, the MFN principle mandates nondiscrimination between treaty partners, for example, granting all countries entitled to MFN status the same tariff. The case selected for empirical analysis is France in 1860, which is the year when it granted Great Britain MFN treatment after decades of discriminatory trade policy (Pahre 1998). This positive case is contrasted with a negative case – any trade agreement negotiated by France and Great Britain or a comparable country that does not include an MFN provision. In the case of France, the problem is that no such agreement was negotiated in the 1850s. Consequently, there is a temporal gap of more than ten years between the positive case and a viable negative case that thereby increases the likelihood that the cases differ on more than a single cause that could explain the adherence or nonadherence to the MFN principle.

The extent to which this problem occurs in a case study is an empirical one. If France had negotiated a comparable treaty without MFN treatment in 1858, the period of analysis would be rather short, and the two cases would be likely to be similar in many respects. However, there are some research questions that are more likely to call for extended periods of analysis and have a higher propensity to include more than one cause that changes over time. Examples for studies that are likely to suffer from such problems include institutions because institutions are rarely subject to abrupt large-scale change and incremental change stretches over an extended period of time (Mahoney and Thelen 2009).[7]

A final issue that limits the value of before–after comparisons is their emphasis on observing variance over time. While it is appealing to observe a change on the outcome in the aftermath of a change on a (purported) cause, the focus on longitudinal variance excludes the implementation of some valuable cross-case designs, such as the method of agreement (Caramani 2010). In total, longitudinal comparisons are valuable because they contrast cases belonging to the same unit. However, there are also several provisos attached that one should take into account in the generation of causal inferences.

Unit-related and time-related comparability and generalizability

The discussion of different forms of cross-case comparisons shows that there is no single best form if the analysis is limited to two cases. For each of the dimensions on which a case is located, there is a trade-off between the comparability of cases and generalizability. Table 5.1 captures these trade-offs by classifying two-case designs with respect to the comparability of cases and generalizability of causal inferences. For ease of presentation, I present only the temporal dimension and one other unit of analysis, which could be spatial, substantive, and so on.[8] A cross-section comparison within the same unit performs best with respect to comparability and worse as regards generalizability across units and time (upper-left cell). The before–after comparison (lower-left cell) and cross-unit comparison (upper-right cell) are equally suitable when the

Table 5.1 Unit-related and time-related comparability and generalizability of two-case comparisons

		Unit	
		Comparability	*Generalizability*
Time	*Comparability*	Within-unit cross-section comparison	Cross-unit comparison
	Generalizability	Longitudinal comparison	–

goal is to achieve comparability across both time and units. The diminished comparability on one dimension is matched by the advantage of somewhat increased generalizability on the other. The lower-right cell is empty because a longitudinal cross-unit comparison is not possible, since a comparison over time requires a comparison of cases within the same unit. In a pairwise comparison, one has to make a choice between a longitudinal comparison within a unit and a cross-section comparison either within or across units.

The trade-offs and the fact that the lower-right cell remains empty for two-case comparisons points to the benefits of a four-case comparison. The four-case design combines two longitudinal designs covering two pairs of cases that belong to different units, meaning that the comparison takes place over time as well as within and across units.[9] This design can be exemplified with a modification of Jakobsen's analysis of liberalization in utilities sectors (2010). The design would involve a four-case comparison if one were to examine the Danish telecommunication and electricity sectors before and after liberalization. This comparison would combine a within-unit perspective with a cross-unit perspective by tracing each sector over time and by comparing the two sectors. Still, the case study would be of the within-unit type in spatial terms because it covers only Denmark. If one were to take a cross-unit view on the spatial dimension as well, one would have to add four more cases to the analysis. More generally, this means that solving the trade-off on each dimension requires a doubling of the number of cases. The dimension-specific trade-off between comparability and generalizability therefore can be solved only by performing multicase comparisons, the pros and cons of which are subject to the next section.

5.2 Comparisons with more than two cases

The reason for indeterminacy on the cross-case level is the result of a mismatch between a large property space and a small number of examined cases. In this light, it is self-evident that an increase in the number of cases is, in principle, suitable for the diminishment of indeterminacy (King et al. 1994, 213–17). The plea for an increased number of cases is often countered with the argument that it is futile (George and Bennett 2005, chap. 1). Qualitative case studies have to emphasize the depth of the within-case analysis and therefore have to focus on a small number of cases (Gerring 2004, 347–8). This is a valid point because one should not select more cases at any price – only to the extent that the depth of process tracing for the expanded set of cases allows one to deliver convincing within-case analyses (Hall 2008).

There are no hard and fast rules for finding the optimal balance between breadth and depth because the depth of process tracing hinges on a multitude of factors, the availability of research resources, for example. It is therefore up to the small-n researcher to make this decision anew for each case study. Within these limits, however, it is desirable to think about the inclusion of more cases.[10] The degree to which more than two cases can diminish inferential problems and the pros and cons of this instrument are addressed in the following. Since a comprehensive discussion of various suboptimal and ideal comparisons is beyond the scope of this section, I rely on illustrative treatments of some comparisons and specific cause–effect relationships.

The comparison in Table 5.2, which is a modified version of Trampusch's comparison (2010; discussed in Chapter 4) exemplifies that more cases may reduce the indeterminacy problem of imperfect designs (Frendreis 1983, 266–7). For purpose of illustration, assume one hypothesizes that the nature of the powerful companies – small/medium v. large – is correlated with the type of institutional change. The comparison of Austria and Germany does not allow one to make this inference because the type of coalition covaries with the type of institutional change as well. Taking the hypothetical case 3 into the picture (highlighted in italics in the following), we see that this case and Austria constitute an indeterminate MoD as well, as we now observe a correlation between the outcome, the type of the powerful company, and the strength of unions. At the same time, however, Germany and case 3 form an ideal–typical MoA. The MoA yields the leverage for the elimination of independent variables because the type of institutional change is invariant, whereas the type of coalition and the strength of unions vary. These inferences can be taken and transferred to the MoD formed by Austria and Germany in order to justify the inference that only the type of powerful company correlates with the type of institutional change. This inference becomes possible in a three-way comparison because we can eliminate an effect of the type of coalition on the basis of the MoA between Germany and case 3. This example exemplifies the benefits of a seemingly paradoxical strategy, which is the strengthening of causal inference by combining two suboptimal MoDs. A look at Table 5.2 shows that this strategy works out

Table 5.2 Imperfect method of difference and multicase comparisons

Case	Institutional change (Y)	Powerful company	Type of coalition	Unions
Austria	Self-preserving	Small/medium	Protectionist	Strong
Germany	Transformative	Large	Liberal	Strong
Case 3	*Transformative*	*Large*	*Protectionist*	*Weak*

Table 5.3 Indirect method of difference

	Case	Compliance of first country (Y)	Judicialization	First country powerful	Second country powerful
MoA 1	WTO 1	Yes	Yes	Yes	Yes
	WTO 2	Yes	Yes	No	No
MoA 2	GATT 1	No	No	Yes	Yes
	GATT 2	No	No	No	No

here because case 3 and Germany constitute an ideal MoA that, in combination with the original comparison, promotes causal inference.

Chapter 4 showed that cross-case inferences are not free of problems even if they meet the standards of an ideal comparison, which entails that there is also benefit to an increase in the number of cases in this instance. One ideal multicase comparison one can strive for is Mill's *indirect method of difference* (IMoD) as a formalized design comprising four cases (Mill 1874, 283–4; Ragin 1987, 38–44). Table 5.3 is built on a modified version of the example by Zangl (2008) presented in Chapter 4. It exemplifies that the IMoD is a combination of two ideal MoAs that can be decomposed into their constituent cases and rearranged so as to obtain two ideal MoDs.

The first two WTO cases constitute an MoA with the outcome present, whereas the second pair of GATT cases produces an MoA for which compliance is absent. At the same time, the cases WTO 1 and GATT 1 and, respectively, WTO 2 and GATT 2 form two MoDs. The IMoD is superior to a stand-alone MoA and MoD for correlational and set-relational case studies. Each MoA allows one to refute a powerful first and second country as variables having an independent causal effect but renders it impossible to demonstrate that judicialization is a cause of compliance. Both MoAs together, however, render it possible to make a positive causal inference as regards the effect of judicialization. Moreover, one can rebut some interaction effects between judicialization and the other two variables. A comparison of the cases WTO 1 and GATT 1 and WTO 2 and GATT 2 reveals that judicialization correlates with compliance regardless of whether both involved countries are powerful or not. Still, one could argue that there would not be any correlation if one country were strong and the other weak. Moreover, one cannot evaluate the argument that the power of the first country correlates with compliance when the power of the second country is invariant across both cases (and vice versa). The IMoD thus permits the empirical examination of more cross-case inferences than the MoD and MoA, but some degree of indeterminacy might remain.

Similar arguments extend to set-relational case studies, whereas there is no need for the IMoD at all if the case study is about necessity. The first MoA suffices to reject the power of the first and second country as individually necessary for compliance. This holds true even if theory is weak and does not allow it to hypothesize whether the absence or presence of a strong country is necessary. In contrast, an IMoD is valuable in inquiries into sufficient conditions. The GATT 2 case indicates that one has not omitted a condition because the outcome is absent in the absence of all supposedly sufficient conditions. Moreover, the WTO 2 case highlights that a powerful first and second country are neither sufficient individually nor when combined in a two-way interaction. If one takes all four cases together, judicialization therefore appears to be necessary and sufficient for compliance. However, this inference again hinges on the availability of good theory because it is justified only when one can argue that the presence of each condition is sufficient (either individually or in conjunction). If this claim is not tenable, the IMoD is indeterminate because there is a multitude of unobserved configurations that include the absence of one or two conditions – for example, no judicialization in combination with a strong and a weak country.

In total, the illustrative discussion of multicase comparisons shows that more cases can diminish the inferential problems of cross-case designs. Whether more cases eliminate indeterminacy is an empirical matter depending on the analysis at hand. When one recalls that the number of cases is still limited in expanded design and assuming that none of the other instruments discussed in this chapter are applied, more cases are likely to change something about the feasibility of cross-case inferences in degree but not in kind.

5.3 Comparisons with multicategorical measures

The methods literature on cross-case comparisons predominantly focuses on dichotomous causes and outcomes (DeFelice 1980; Sartori 1991; Tarrow 2010), presumably because of the preoccupation with the MoD and MoA that are built on binary causes and outcomes. Dichotomous measurement of course is mandatory when a hypothesis stipulates a causal relationship between binary phenomena. When this is not the case, one can choose between binary and more fine-grained measurement, the latter subsuming multicategorical and continuous measures. With the MoA and MoD arguably being the most widely known of Mill's methods, Mill (1874, 285–9) offers the *method of concomitant variation* (MoCV) for the analysis of nonbinary causes and outcomes (Mahoney 2000, 399–406). The idea behind the method of concomitant variation is to discern a potentially existing correlation between multicategorical causes and outcomes, which implies that the classic MoCV is limited to covariational case studies. As this chapter shows, it is possible to rely on more fine-grained measurement in set-relational case studies as well (Cronqvist and Berg-Schlosser 2008) but with different implications than in correlational designs. The most apparent difference is that set-relational case studies can use only multicategorical conditions and outcomes, whereas one can, in principle, rely on multicategorical and continuous measures in covariational small-n analysis (but see below).[11]

In the following, the focus is on what I call the *level of measurement aggregation* because this is the aspect that matters in generating cross-case inferences.[12] One can distinguish binary, multicategorical, and continuous measurement as the three basic levels of measurement aggregation.[13] Nominal and ordinal measures can be bi- and multicategorical but not continuous, whereas continuous variables can be measured continuously and broken down to multiple and two categories. The highest level of measurement aggregation that one can achieve is binary measurement. Correspondingly, a low level of aggregation denotes multicategorical measurement (for reasons detailed below, continuous measures are left aside in the following). It further follows that one can compare the level of measurement aggregation of causes and outcomes. One cause or outcome has a higher or lower level of measurement aggregation when based on fewer or more categories than another cause or outcome. For example, measuring economic growth in terms of high and low growth implies a higher level of aggregation than the distinction between high, moderate, and low growth.[14]

5.3.1 Correlational comparisons

In principle, one can use continuous measures in qualitative case studies. However, little is gained by using continuous measures because their main rationale is to estimate a causal effect or at least to calculate a correlation coefficient as a measure for the strength of an association. This is a pointless endeavor with two cases because the outcome is necessarily fully predicted by the cause owing to the lack of degrees of freedom. Continuous measurement makes sense only in a multicase comparison, and all the more so the larger the number of cases. However, there is a tight limit on the number of cases one can handle in in-depth qualitative case studies (Gerring 2004), the implication being that continuous measurement is not meaningful in small-n research.

Given that continuous measurement does not offer inferential benefit, the only viable variant of nonbinary measurement includes multicategorical causes and outcomes. An example for a multicategorical cross-case study is Eckert's analysis of delegation in the European postal sector of the United Kingdom, Germany, and France (2010). The three selected countries are instances of different varieties of capitalism; the United Kingdom is a liberal market economy (LME), Germany a coordinated market economy (CME), and France a state-led market economy (SME).[15] Among other things, Eckert relates the type of capitalism to the de facto independence of the regulatory agency in the respective national postal sector. More specifically, Eckert distinguishes between a strong, medium, and weak degree of de facto independence. The expected link between the type of capitalism and the level of de facto independence is depicted in Table 5.4. Eckert's empirical analysis shows that her hypothesized correlation between the type of capitalist system and the level of independence has empirical resonance. For purposes of illustration, Table 5.4 additionally includes the expected competitiveness of the national postal company in a liberalized European postal market as a rival explanation. In this example, the competitiveness is assumed to be high for all three countries, meaning that it can be discarded as a standalone independent variable in this study.

This example denotes two important aspects of multicategorical measurement. First, a low level of measurement aggregation renders it necessary to compare more than two cases in order to be able to test the hypothesis under analysis. A comparison of the United Kingdom and Germany would point to a correlation between the type of capitalism and independence. Yet one cannot fully confirm the hypothesis because the analysis lacks a capitalist system of the SME type. Compared with the use of dichotomous measures, for example, simply distinguishing between LME and non-LME countries, three-categorical measurement increases the property space with direct implications for the number of cases that the small-n analysis should include. If one does not want to perform a counterfactual analysis of the degree of independence in a state-led economy, a country such as France must be added to the comparison if one is to distinguish between three types of capitalist systems. On a more general level, multicategorical measurement calls for correlational multicase comparisons because one needs to have as many cases as one distinguishes categories on the independent and dependent variables.

Table 5.4 Comparison with multicategorical causes

Country	***De facto* independence (Y)**	**Capitalist system**	**Competitiveness of postal company**
United Kingdom	Strong	LME	High
Germany	Medium	CME	High
France	Low	SME	High

A second reason complicating multicase comparisons is the increase in the number of possible interaction effects that derives from multicategorical measurement. The comparison in Table 5.4 exemplifies the point. The inference that the capitalist system correlates with the level of de facto independence implicitly assumes additive causation. Yet the cross-case pattern is also compatible with the inference that this correlation is in place only when the national postal company is highly competitive. Assuming that competitiveness is measured binarily in terms of high and low, a test of this counterargument requires the selection of three additional cases. One would be able to demonstrate that no interaction effect is in place when the type of capitalist system and independence also correlate with each other when the postal company assumes low levels of competitiveness. When competitiveness is measured multicategorically in terms of high, moderate, and low levels, the number of additional cases required for an empirical test increases to six. One would have to choose two pairs of three countries that are instances of LMEs, CMEs, and SMEs, respectively. Each pair then needs to be combined with the variable 'competitiveness', taking moderate and low values and yielding six additional cases altogether. In addition to the fact that it is getting more difficult to create an ideal–typical cross-case comparison with one or more multicategorical independent variable, the example further indicates that the number of possible interaction effects quickly extends beyond the number of cases that one can reasonably examine in qualitative case studies.

So far, the discussion presumed that the number of categories is similar for the independent and dependent variable. This may be the case but not necessarily because of the possibility to assign more categories to the cause than the outcome and vice versa. In practice, a design measuring the independent variable on a lower level of aggregation suffers from an inferential pitfall, whereas it has merit to distinguish more categories on the dependent than the independent variable. Table 5.5 illustrates the benefits of the latter type of cross-case comparison with a modification of Eckert's case study. The outcome, de facto independence of the national regulatory agency, is still measured in terms of the three categories strong, medium, and high. The two independent variables under scrutiny are measured binarily and

Table 5.5 Comparison for assessment of additive causation

Country	***De facto* independence (Y)**	**State intervention in market**	**Competitiveness of postal company**
United Kingdom	Strong	Low	High
Spain	Medium	Low	Low
Portugal	Low	High	Low

concern the intervention of the state in the postal market and the competitiveness of the national postal company. The correlation between state intervention and de facto independence is expected to be negative, while the correlation between competiveness and independence is hypothesized to be positive. Table 5.5 presents the hypothetical empirical results of the case study and shows that the two propositions are empirically accurate.

With respect to Mill's methods, such a comparison can be conceived of as a mix of the MoD and the MoCV. The outcome brings in an element of the MoCV because of multicategorical measurement, whereas the MoD is represented by binary independent variables. As can be seen in Table 5.5, this measurement strategy makes it possible to examine *additive causation* in the context of a case study. A comparison of Spain and Portugal shows that a decrease in state intervention goes along with an increase in de facto independence from low to medium. This change cannot be attributed to the competiveness of the postal company because it is the same for both countries. A comparison of Spain and the United Kingdom further indicates that an increase in the company's competitiveness leads to a shift from medium to high levels of independence. This change is not attributable to the intervention of the state as it is invariant across the United Kingdom and Spain. Binary measurement of the independent variables and multicategorical measurement of the outcome therefore renders it not only possible to establish an association between state intervention, competitiveness, and de facto independence, but also to offer evidence that the individual causal effects are additive. For the reasons detailed above, though, the ability to discern additive causation in multicase comparisons comes at the price of a potential increase in the number of possible interaction effects. While acknowledging this complication, however, it can have inferential merit to distinguish more categories on the dependent variable than the independent ones.[16]

It should be understood that distinguishing more categories on the dependent variable is not always a beneficial strategy. In particular, one runs into problems when the dependent variable is linked to a single independent variable. The problem can be exemplified with Lange's analysis of the effect of the form of British colonial rule on the economic development of the former colonies (2009). Lange distinguishes between indirect and direct colonial rule and relates it to low, medium, and high economic development. The expectation is that the development is better under direct rule than indirect rule. A comparison of Sierra Leone, Botswana, Mauritius, and Guyana yields the picture summarized in Table 5.6 (Lange 2009, 14).

One can argue that the expectation of a correlation between the type of British rule and level of development is corroborated. However, Botswana and Guyana seem to constitute a puzzle because, although the type of rule is different, they have the same degree of development (Lange 2009, 14). This puzzle should not come as a surprise because the independent variable has a higher level of measurement aggregation than the outcome. This makes

Table 5.6 Comparison with more categories on Y than X

Case	Economic development (Y)	Colonial rule
Sierra Leone	Low	Indirect
Botswana	Medium	Indirect
Guyana	Medium	Direct
Mauritius	High	Direct

it possible for two cases with different scores on the independent variable to fall into the same category on the dependent variable, as is the case with Botswana and Guyana in Table 5.6. Because of this, a higher level of measurement aggregation on the dependent variable has little use in correlational case studies if one is examining a single independent variable with fewer categories. Without going into the details here, it holds that the distinction of more categories on the independent variable runs into similar problems.

Set-relational comparisons

Multicategorical measurement has slightly different implications for set-relational designs owing to the different nature of correlational and set-relational cause–effect relationships. First of all, it is to note that the outcome's level of measurement aggregation does not matter in analyses of necessity. Whether the outcome is binary or multicategorical is not relevant because one is interested only in the necessary conditions for a specific outcome set. In order to clarify this point, consider a hypothetical modification of Eckert's comparison of delegation in the postal sectors of France, Germany, and the United Kingdom (2010). Let's imagine the outcome of interest is the strong independence of a regulatory agency as opposed to a moderate or low level of independence. The level of measurement aggregation is irrelevant because we are interested only in the necessary conditions of strong de facto independence. Consequently, we look only at countries that have installed regulatory agencies that are strongly independent and ignore the other cases.

The situation is different for case studies that are interested in sufficient conditions. Continuing with another modification of Eckert's research, suppose you hypothesize that a high level of competitiveness is sufficient for a strong degree of de facto independence (Table 5.7). Furthermore, suppose that the initial case study includes only the United Kingdom and Belgium.

If one were to perform an ordinary MoA, the pairwise comparison could not show that the effect of a high level of competitiveness is independent of the type of capitalist system because one does not observe a highly competitive company that is operating in an SME. Consequently, a complete empirical assessment of the individual sufficiency of high competitiveness requires the comparison of three cases that are instances of LME, CME, and SME,

Table 5.7 Set-relational comparison with one multicategorical condition

Country	*De facto* independence (Y)	Capitalist system	Competitiveness
United Kingdom	Strong	LME	High
Belgium	Strong	CME	High
Italy	Strong	SME	High

respectively. On a more general level, this example again points to the problems of expanded property spaces owing to multicategorical measures. The more multicategorical conditions that one has and the more categories that one distinguishes, the greater the number of viable causal inferences tends to be.

At the end of this section, it should be emphasized that theoretical and conceptual reasons should drive the choice of the measurement approach. If one believes that binary measurement is too broad and conceals interesting patterns that can be discerned only via more fine-grained measurement, a multicategorical measurement approach should be followed. But in doing so, one should take note of the implications that an increased property space has for the indeterminacy problem and causal inferences. In all other instances wherein theory speaks neither for binary nor multicategorical measurement, a case study researcher should carefully weigh the benefits and costs of this decision.

5.4 Better theory

In parts of the previous sections and Chapter 4, the assumption was that theory is weak. 'Weak theory' here means that it does not allow one to reject some empirically possible cross-case inferences as too implausible on theoretical ground.[17] If one relies on theory after the empirical analysis, this tool is essentially similar to counterfactual reasoning, which is reserved for discussion in Chapter 7.[18] In hypothesis-testing case studies, the careful consideration of theory prior to the empirical analysis offers additional benefits in the generation of causal inferences (Dür 2007a).[19] For example, Bueno de Mesquita (2003, 57) points out that realism as laid out by Morgenthau and Thompson (1985) witnesses a logical inconsistency.[20] According to realism, states are above all concerned about their survival and act accordingly in international relations. On the one hand, it is then argued that all states seek power in order to survive. On the other hand, it is claimed that there are states that aim to maximize power and those that do not. In the face of such an inconsistency, it is not clear what observable implications

Table 5.8 Three steps toward stronger cross-case inferences

Step	Question to ask
Specifying causes	What causes are plausibly related to the outcome?
Specifying the nature of the causal effect	Is the causal effect correlational or set-relational? If it is correlational: is the correlation positive or negative? If it is set relational: is the condition necessary and/or sufficient?
Specifying interaction effects	What interaction effects are and are not plausible in light of theory?

one can legitimately derive from the theory. It is arguably possible to get rid of logical inconsistencies by recasting the theory. However, it is not up to a researcher to improve rival explanations if they suffer from contradictory claims. Instead, one can confine oneself to the assertion that the theory, in its present state at least, cannot be tested empirically by formulating unambiguous hypotheses.

A look at the inferential problems that were discussed in the previous chapter hints at three issues that make a theory good in relation to cross-case comparisons (Table 5.8). First, one needs a recipe of causes to be included in the analysis, one that is to be derived from current theories. The more causes one can discard as irrelevant at this stage on the basis of theory (Dür 2007a; Nassmacher 2010), the smaller the potential problems of indeterminacy will be when interpreting a cross-case pattern at a later stage of the analysis. This criterion is limited to hypothesis-testing case studies because the other two research goals call for exploratory small-n research that cannot start with a predefined set of potential causes.

Second, one should hypothesize for each cause whether it is expected to correlate with the outcome or follow a set-relational pattern. If one theorizes a correlation, it should be specified whether it is positive or negative. When the causal effect is expected to be set-relational, one should clarify whether the presence or absence of a condition is necessary or sufficient for the outcome. Third, regardless of the type of causal effect, one should additionally consider what interaction effects may be in place.[21] The more logically possible interaction effects can be claimed to be theoretically implausible, the smaller indeterminacy becomes in this respect.

If these three steps are followed, cross-case inferences can be generated in a theoretically disciplined and systematic way. This is not to say that this will be an easy endeavor because many fields of the social sciences are currently characterized by weak theory and it is unlikely that the range of theoretically and empirically convincing inferences can be boiled down to one. However, when complaining about the state of social science theory, it should not be forgotten that theory is only as good as social scientists make it and that there are manifest benefits to the strengthening of theory.

5.5 Turning causes into scope conditions

The previous discussions of the MoA and MoD demonstrates that their classic (correlational) interpretation rests on implicit assumptions about the role of those causes that are not of prime interest. In the MoD, these are the invariant causes, in the MoA, the varying causes.[22] However, unless one is able to make strong claims about the absence of interaction effects and equifinality, the inference that invariant causes do not play a role in the MoD or that the varying causes can be ignored in the MoA is not justified.[23] In principle, though, it is a valuable idea to reduce indeterminacy by removing some causes from the picture.

A strategy that achieves exactly that is the transformation of potential causes into scope conditions. The rationale is that scope conditions constitute the bounds *within* which a specific causal relationship is expected to be hold. A boundary condition is taken as a given in the analysis, and there is no obligation to consider whether its absence would have an impact on the causal relationship of prime interest.[24] Therefore, the replacement of a cause by a proper scope condition achieves for the generation of causal inferences what the purposeful matching of cases can achieve only under certain and usually demanding assumptions. However, the following discussion shows that this tool does not offer an easy solution to all inferential problems because its value hinges on the nature of the comparison and whether it is ideal or imperfect.

For presentational purposes, this section has to be limited to the suboptimal and ideal MoA and MoD, what are arguably the most widely applied types of comparisons. This section is then organized as follows: the section first turns to the ideal and suboptimal MoD and then considers the ideal and imperfect MoA. Within the discussion of each variant of comparison, correlational causal relationships are discussed first, followed by an elaboration of set relations.

The discussion of an *ideal covariational MoD* is based on Zangl's (2008) comparison of compliance in the WTO and GATT (see Chapter 4). As Zangl acknowledges, the inference that judicialization has an independent effect is seconded by the claim that there may be an interaction effect between judicialization and the power of the first and/or second country (Table 5.9). For instance, one can argue that a powerful country may comply with the dispute settlement mechanism only if the opponent is powerful as well (Zangl 2008, 832).

As a means to reduce indeterminacy, suppose that one takes a powerful second country as a scope condition (denoted by the parentheses in Table

Table 5.9 Scope conditions and an ideal method of difference

Case	Compliance of first country (Y)	Judicialization	First country powerful	(Second country powerful)	Substance matter
WTO	Yes	Yes	Yes	(Yes)	Hormones-treated beef
GATT	No	No	Yes	(Yes)	Hormones-treated beef

5.9 and all following tables). In this example, the substitution of a cause with a corresponding scope condition has two interrelated consequences. First, the property space declines. The scope condition 'powerful second country' implies that one looks only at cases where the second country is powerful, so that one is relieved from controlling for the power of the second state in the actual comparison (Walker and Cohen 1985, 297). Second and as a consequence of this, the second country's level of power ceases to be an independent variable with which any other variable could form an interaction effect that one needs to consider in this case study. Indeterminacy thus decreases because all inferences that include the power of the second country in the original comparison drop from the list of possible inferences in a comparison for which a powerful second country is a scope condition.

As regards *set-relational causation*, necessary-condition case studies need not be discussed here because the MoD is not suitable in analyses on necessity (see Section 4.2). Table 5.9 shows that this differs from research on sufficiency, in which the transformation of causes into scope conditions is viable. As explained in Chapter 4, the GATT case allows one to refute the individual and joined sufficiency of all conditions that are present in this case. However, one cannot infer whether the cause of interest – judicialization – is individually or jointly sufficient in the WTO case. Taking a powerful second country as a scope condition reduces indeterminacy because the range of possible inferences on sufficient conjunctions decreases.

While the invocation of additional scope conditions can promote causal inferences, it is no panacea because it cannot turn a *suboptimal* MoD into an ideal one. This argument is first illustrated with a modified version of the Zangl comparison and the discussion of a correlational causal effect of judicialization (Table 5.10). The modified variant additionally includes the concerned trade volume, representing the argument that countries may comply when there is not much at stake in the dispute but fail to comply if the concerned trade volume is high. The comparison in Table 5.10 is imperfect because the variables 'judicialization' and 'trade volume' covary with the outcome.

In the face of this problem, it might be tempting to get rid of this problem by taking a low trade volume as a scope condition.[25] This would remove

Table 5.10 Scope conditions and a suboptimal method of difference

Case	Compliance of first country (Y)	Judicialization	First country powerful	Second country powerful	(Trade volume)
WTO	Yes	Yes	Yes	Yes	(Low)
GATT 1	No	No	Yes	Yes	(High)
GATT 2	*No*	*No*	*Yes*	*No*	*(Low)*

Table 5.11 Scope conditions and an ideal method of agreement

Case	More elements of direct democracy (Y)	International debate about more direct democracy	Electoral system	Government	(Organization of state)
The Netherlands	Yes	Yes	Proportional	Coalition	(Federal)
United Kingdom	Yes	Yes	Majoritarian	Single party	(Unitary)
United States	*Yes*	*Yes*	*Majoritarian*	*Single party*	*(Federal)*

the GATT 1 case from the population, which implies that the WTO case needs to be compared with another case such as GATT 2 (cases included in the comparison in a second step are highlighted in gray in Table 5.10 and all following tables). The GATT 2 case is necessarily comparable to the WTO case as regards the trade volume because otherwise it would have been eliminated from the population. However, the two cases must constitute an imperfect comparison as well because otherwise one would have compared them at the outset. In Table 5.10, the WTO case and the GATT 2 case represent an imperfect comparison because they differ with respect to the level of judicialization and the power of the second country. The specification of the scope condition 'high trade volume' therefore does not change anything about the imperfection because now one faces a similar problem with respect to two other variables.

The inferential merit of additional scope conditions is slightly different for ideal and imperfect MoAs. The first leading example of this discussion is presented in Table 5.11 and concerns an extended variant of the study by Hendriks and Michels (2011) introduced in Section 4.2. In a *correlational view*, an ideal MoA allows one to reject all variables that take different scores as individual causes of the outcome. Still, there is a rationale for turning a variable into a scope condition, as it may be possible that an interaction between two or more variables is at play. For instance, assume one hypothesizes that the organization of the state correlates with the invention of elements of direct democracy if there is no debate about direct democracy. If we cannot examine a country that is not influenced by the international debate, there is no way to assess this hypothesis empirically.[26]

One can salvage the case study against this rival argument by turning federalism into a scope condition, which means that the variable 'state organization' is not part of the empirical analysis anymore. One consequence is that the United Kingdom ceases to be part of the population and that one has to find another federal country for comparison with the Netherlands. In this example, one selects the United States in order to construct another ideal MoA. The comparison might still involve some degree of indeterminacy because the three remaining variables could be involved in interaction effects, as well. Nevertheless, treating federalism as a scope condition reduces indeterminacy because all inferences that include the organization of the state are taken out of the empirical analysis.

In inquiries into *necessary conditions* that rely on the ideal MoA, the transfer of causes into scope conditions is *not* warranted. For Table 5.11, a comparison of the United Kingdom and the Netherlands permits the inference that an international debate is necessary.[27] The Netherlands additionally show that the three institutional variables are not individually necessary for more direct democracy. If theory allows one to theorize the individual necessity of specific conditions (such as unitarism as opposed to a general reference to the organization of the state), nothing is gained by treating any condition as a scope condition. In fact, it is inferentially harmful to pursue this strategy. Consider the implication of taking unitarism as a boundary statement. When unitarism is a scope condition, one finds it impossible to empirically show that this condition is not necessary because the Netherlands is excluded from the analysis. On the other hand, if one treats federalism as a boundary condition, the analysis of the UK demonstrates that unitarism is not necessary for more direct democracy. Since the same can be achieved without invoking federalism as a scope condition, there is no added value to this strategy in this setting. Similar arguments apply when theory is weak in the sense that one does not know whether the presence or absence of a condition is necessary. As explained before, one must then compare two cases in which the condition under scrutiny is present and absent, respectively, and it is impossible to do so when the presence or absence of one condition is taken as a scope condition.

The situation is slightly different for case studies that aim at *sufficient conditions* and rely on an ideal MoA. In Table 5.11, the analysis of the Netherlands permits the inference that an international debate is individually sufficient.[28] The indeterminacy of the ideal MoA derives from the United Kingdom because it is compatible with multiple causal inferences involving claims of equifinality (see Section 4.2). Consequently, causal inference can be promoted by turning potentially sufficient conditions into

Table 5.12 Scope conditions and a suboptimal method of agreement

Case	More elements of direct democracy (Y)	International debate about more direct democracy	Electoral system	Government	(Organization of state)
Canada	Yes	Yes	Majoritarian	Single party	(Federal)
the Netherlands	Yes	Yes	Proportional	Coalition	(Federal)

scope conditions. If one takes federalism as a scope condition, for instance, all set-relational inferences that include unitarism as a condition cease to be relevant, and the indeterminacy of the comparison declines.

In an *imperfect* MoA, adding scope conditions is a good strategy regardless of the nature of the cause–effect relationship, which distinguishes this comparison from the MoA. This can be exemplified with the suboptimal MoA in Table 5.12, which is another modification of the study by Hendriks and Michels.

In a *correlational* perspective, the invariance of the variables state organization and international debate is problematic because both qualify as potential independent variables. A similar issue pertains to case studies aiming for necessity and sufficiency because both conditions can be treated as necessary conditions and sufficient conditions, respectively. As Table 5.12 shows, the indeterminacy that is inherent in a suboptimal MoA vanishes when federalism, for instance, is treated as a scope condition. The cross-case scores then resemble an ideal MoA, and indeterminacy is automatically reduced regardless of whether the causal relationship is about correlations, necessity, or sufficiency.

Altogether, this section shows that the transformation of causes into scope conditions can diminish indeterminacy and improve cross-case inferences. However, there is no unambiguous link between the replacement of causes with scope conditions. Whether more scope conditions have positive or negative ramifications for cross-case inferences depends on a variety of aspects such as the type of hypothesized causal relationship and the cross-case pattern at hand. It is thus necessary to carefully consider the consequences of this strategy before putting it into practice.

Another issue to be discussed at the end of this section concerns the repeated use of this instrument. In the previous examples, only one potential cause was transformed into a scope condition. Owing to the nature of scope conditions, it is evident that the pursuit of this strategy decreases the size of the population. How large the decline is cannot be determined in the abstract because it depends on how many cases meet the new scope condition. However, it is clear that the more causes are substituted with scope conditions, the smaller the population becomes. In the extreme scenario, one would take all but the cause of prime interest as scope conditions. While viable in principle, such an extreme strategy would eliminate all rival explanations from the empirical analysis. Hence, the comparative assessment of hypotheses becomes more restricted, the larger the of causes that are transformed into scope conditions. Consequently, the transfer of causes into scope conditions entails a trade-off between the comparability of cases on the one hand, and the feasibility of comparative hypothesis-testing. A second trade-off to be noted in this context concerns the reduction of indeterminacy and the breadth of generalization. As mentioned above, more scope conditions imply smaller populations, meaning that more confidence in one's causal inferences comes at the expense of a diminished number of cases to which the inferences are generalized. From the viewpoint of the case study method, there is no reason why larger populations are preferable to smaller ones, but one should keep this trade-off in mind when evaluating the transfer of causes into scope conditions.

5.6 Conclusion

The problems that plague the generation of cross-case inferences are manifold and severe. This chapter served to discuss five instruments that can be used to promote causal inference in comparative case studies. As these instruments are not mutually exclusive, good cross-case analysis combines multiple tools, such as a longitudinal comparison with elaborated theory and the substitution of some causes with scope conditions. However, it should be openly acknowledged that, with the exception of improved theory, neither of the tools is without its costs; a longitudinal comparison comes at the expense of confidence in the generalizability of causal inferences; binary measurement yields a more coarse-grained picture of the social world than multicategorical measurement, and so on. If the diminishment of indeterminacy is not the only goal for a case study researcher, one should be cognizant of the compromises attached to each instrument and make a conscious choice for or against each specific tool.

6

Process Tracing: Theory, Temporality, and Method

Process tracing represents the empirical core of many, if not most, case studies because inferences on causal mechanism and processes often are at the heart of small-n research (Mahoney 2007a, 2007b). The literature on process tracing and mechanisms burgeoned over the last years and contributed to the improvement of causal inference in within-case analyses. As I explain in Chapter 8, however, many of the inferential issues dealt with in the context of process tracing fully extend to cross-case analysis. A discussion of tools, such as Bayesian inference, with an exclusive focus on process tracing therefore misrepresents the potential of these instruments. For this reason, I reserve a consideration of many aspects that one knows from the process tracing literature for Chapter 8 as this chapter takes a view on causal inference that covers cross-case and within-case analysis alike.

The present chapter is interested only in issues that are specific to process tracing, which is examined from two different, though complementary perspectives. In the first half, I point out potential pitfalls in achieving a fit between the hypothesis of interest and the way process tracing is implemented.[1] The pitfalls are related to *realized* and *anticipated* processes as two different ideal types of processes that one can theorize in case studies. In Section 6.1, I first introduce various empirical examples of process tracing to demonstrate that there are different ways to theorize and examine processes. Section 6.2 takes a more general perspective and distinguishes process tracing on realized and anticipated processes along four dimensions. One central insight is that the same empirical phenomenon can be examined via the analysis of realized and anticipated processes. This makes it particularly necessary to understand their respective advantages and disadvantages, which are discussed in Section 6.3.

In the second half of this chapter, the focus rests on the role of time and temporality in theory and the ramifications for process tracing. The notion of temporality is particularly salient for Comparative Historical Analysis, but it can play an important role for case studies more generally as it offers the opportunity to derive additional, time-related observable implications from a theory. Section 6.4 discusses time-related concepts such as sequencing and path dependence with an eye on their implications for the case study method and process tracing.

6.1 Processes and process tracing: introductory examples

In contrast to the ongoing debate about the nature of causal mechanisms (see Section 2.3), there seems to be more of a consensus about the definition of process tracing. George and Bennett understand process tracing as 'attempts to identify the intervening causal process – the causal chain and causal mechanism – between an independent variable (cause) and the outcome of the dependent variable' (George and Bennett 2005, 206). Similar understandings are, for example, shared by Hall (2008) and Checkel (2008, 115), who defines process tracing as a method that 'identifies a causal chain that links independent and dependent variables'. Even scholars who take a more critical perspective on process tracing and within-case analyses subscribe to this definition (Gerring 2007a, chap. 7).

These definitions imply that the cause of interest is the inception of a process that was realized empirically and that resulted in the outcome (George and Bennett 2005, 177). This point and the definition of process tracing more generally is illustrated by the example of international tax competition in Figure 6.1 (for illustrative purposes, I focus on events here that are the product of entities and activities). Imagine that the outcome that one wants to explain is the reduction of company tax rates by a certain country in a specific year (referred to as the domestic country in Figure 6.1.

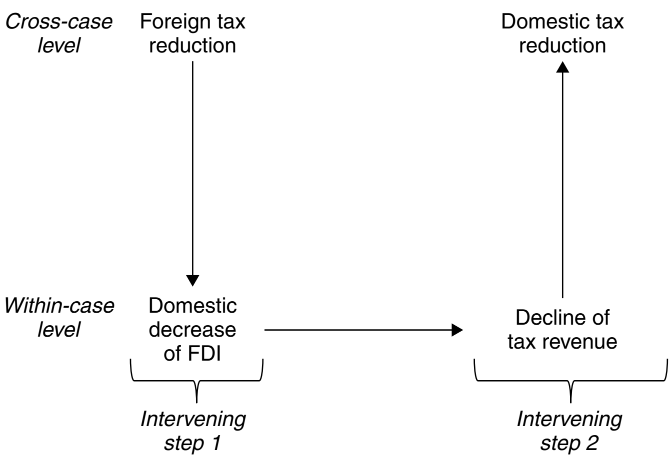

Figure 6.1 Process tracing by the example of tax competition

The hypothesis under analysis stipulates that a country reduces its company tax rates in response to the company tax reduction of another country.

Since the rationale of process tracing can be exemplified with a simple process, the intervening steps are limited to two. In the first step, the company tax reduction of a foreign state leads to a decline in foreign direct investment (FDI) of the other country. The decline in FDI diminishes economic growth and competitiveness of the domestic companies, which, in the second step, leads to decline of tax revenue. The drop in tax revenue creates pressure for a tax reform, which eventually comes about in the form of a reduced company tax rate in the domestic country.

This simple example points to three elements of process tracing as expressed in the definitions presented above. First, the hypothesized cause is, at the same time, the triggering event that starts an empirical process. Second, the arrows in Figure 6.1 represent temporal order and causal influence. A decline in FDI is caused by the foreign tax reduction, which necessarily implies that the tax reduction precedes the decline in FDI in time. Similarly, decreasing FDI causes declining tax revenue, which in turn is a cause of lower domestic taxes. In set-relational terms, this means that each intervening step can be taken as a sufficient condition for the subsequent step (Goertz 2003).

Third, it is possible to theorize a fixed sequence of intervening steps. The sequence of steps depicted in Figure 6.1 can be theorized to occur in exactly this order and then submitted to empirical scrutiny in process tracing (George and Bennett 2005, 30). This form of within-case analysis sometimes goes under the rubric of pattern matching because one matches the hypothesized to the observed sequences. The opportunity to engage in *pattern matching* has implications for causal inference because one should observe an uninterrupted causal process by finding causal process observations (CPOs) for every intervening step of the theorized path (218–22). The plea for this seems straightforward because if one can theorize a full sequence of steps, one should also test for the presence of each step because otherwise parts of the argument remain untested. In set-relational terms, this means that each step is a necessary component and they are jointly sufficient for inferring that the purported cause indeed is a cause.[2]

The literature on process tracing sometimes takes a broader understanding of pattern matching. In a relaxed view, it simply denotes the comparison of multiple observable implications with the CPOs that one collects (Campbell 1975). This is good practice in empirical research regardless of the employed method (Lave and March 1975), but it is not entirely clear what the patterns are that are matched. Nevertheless, I follow the distinction between a narrow and a broad, or relaxed, conception in the following sections.[3]

This perspective on process tracing with its three characteristic elements does apply to analyses concerned with hypotheses on tax competition and many propositions dealing with other causal relationships. However, there are also hypotheses that fail to meet these criteria. A prominent example of such a hypothesis is the democratic peace hypothesis, which states that a democratic dyad accounts for peace between countries. The cause 'democratic dyad' denotes that a specific pair of countries represents two democracies, which is not an empirical event that can initiate a process at the end of which one observes peace. The tax and peace examples can be phrased in terms of a cross-case hypothesis, therefore highlighting that the differences between them do not rest in different theoretical ends.

This claim can be substantiated by formulating within-case hypotheses supplementing the cross-case relationships. For the tax competition example, one can hypothesize that a country reduces its company tax rates in response to the company tax reduction of another country because it leads to a decline of FDI and revenue in the domestic country. A within-case hypothesis predicated on the democratic peace phenomenon is that a democratic dyad accounts for peace between countries because members of the political elite are committed to democratic norms (Rosato 2003, 586). In the case of an emerging conflict, a democracy is aware that another democracy adheres to these principles as well. The other state can be respected and trusted, which altogether accounts for peaceful conflict resolution among democracies (Rosato 2003, 584). If one takes 'commitment to democratic norms' as the causal mechanism, it again holds true that the mechanism does not trigger a process that results in peace. The fact that 'democratic dyad' and 'commitment to democratic norms' are not triggering events becomes apparent in case studies of the democratic peace phenomenon that start with an event that creates a crisis between two democracies and nearly brings them to war (for example, Owen 1994; Schultz 2001).[4]

One could continue with additional examples of hypotheses that are similar to the tax competition hypotheses and the democratic peace hypothesis. However, the goal of this section is not to explore which of the classes of hypotheses is more prevalent but to illustrate that there apparently are different ways to theorize processes. In the following section, I systematize these differences, develop criteria distinguishing both classes of processes from each other, and detail the implications for process tracing.

6.2 Theorizing realized and anticipated processes

The literature on process tracing sometimes hints at the reason underlying the claim that one can theorize two different types of processes with implications for how process tracing is realized (for example Hall 2008). However, to date, this has not been thoroughly addressed. The differences between the hypotheses and processes illustrated in the previous section can be grasped by distinguishing between hypotheses that are concerned with *realized* processes and those concerned with *anticipated* processes. As this section shows, this distinction is not meant to develop a new logic of process tracing. It simply aims to bring to the forefront the two ideal–typical variants of hypotheses and process tracing that one finds in the empirical literature that have not been sufficiently appreciated so far.

The tax competition example is an instance of a hypothesis on a realized process as it was described above. The proposition has the cause as the starting point and the outcome as the end point of a process that was realized empirically, that is, can be traced and observed in a within-case analysis. On the other hand, the democratic peace hypothesis is implicitly concerned with what I call *anticipated processes*. A hypothesis explaining an outcome with anticipated processes focuses on the considerations that actors make before coming to a decision and/or committing a specific action. The consequences that actors expect will unfold if they take a specific action then account for their performing the action that results in the outcome.

An empirical example can best illustrate what a hypothesized anticipated process is. Suppose that two democracies are caught in a trade war that is causing domestic turmoil in both states because of devastating economic consequences. The economic conflict's potential for escalating into war is thwarted because of the influence of democratic norms. In order to demonstrate the involvement of anticipated processes, the democratic norms argument needs to be elaborated further. Before the trade conflict de-escalates, the government in each country weighs the two alternatives, war and peace (assuming the government decides the course of action). If a government decides in favor of war, it will be perceived as violating democratic norms. This decision in turn will lead to a decline of trust by other states; foreign relations will worsen; other countries will take a more hostile attitude toward that government or show a decreased willingness to sign trade or environmental agreements that would benefit the belligerent country. When a government opts for peace, on the other hand, it expects to be still perceived to be trustworthy and its foreign relations do not worsen. The two outcomes that a government anticipates to ensue from war and peace are depicted in Figure 6.2 (again, one can think of more complex processes and alternative processes but a simple scenario suffices for presentational purposes). In a case study on anticipated processes, one hypothesizes that the government expects these processes to unfold and opts for one of the two according to which it prefers most.

The example of democratic norms may appear to imply that hypotheses on anticipated processes presume a rational choice framework. Indeed, there is an affinity between rational choice and the democratic norms example as it was presented here; governments eschew war because of their belief that choosing conflict will be to the detriment of their foreign relations. In other words, the *expected* utility of peace is larger than the *expected* utility of war, a conclusion that points to the suitability of case studies for rational choice research (Bates 2007; Pahre 2005). The close tie between process tracing

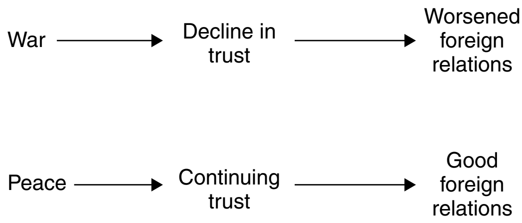

Figure 6.2 Democratic norms and anticipated processes

and rational choice becomes particularly apparent in game theoretic models that focus on sequential games. In sequential games, actors make a decision at present depending on the expected utility of various courses of actions taken at different points in time by this actor and other actors, mirroring an anticipated process as it was defined above.

However, there is nothing inherent in the inquiry of anticipated processes that demands the adoption of a rational choice perspective. The democratic norms example works equally when one assumes that actors follow the logic of appropriateness instead of the logic of consequentiality (March and Olsen 2006). In a broad sketch, an actor following the logic of appropriateness acts according to principles common in the environment in which it is embedded. With regard to the democratic peace example, a democracy knows that it is appropriate, that is, it is expected that the democracy will resolve a conflict with another democracy peacefully because that is the code of conduct in the international democratic community. The logic of appropriateness places emphasis on norms, but this does not imply that actors ignore the implications of their behavior. Without going into detail here, one can generally argue that an actor follows specific norms of behavior because it is expected that the violation of these norms would be disapproved of by other actors. The analysis of anticipated process can therefore be open to a variety of reasons why actors decide and act in specific ways, including narrow material interests and ideational factors.[5]

Independent of this point, Figure 6.2 demonstrates that hypotheses on anticipated processes can be visualized in ways similar to realized processes (Figure 6.2 ignores the distinction between the cross-case and the within-case level because it is not significant here). However, there are salient differences between anticipated and realized processes of which one needs to be aware in order to do process tracing properly. First, it is obvious that the anticipated processes cannot be the empirically relevant process for process tracing because the outcome is the starting point of the anticipated process. Neither of the two anticipated processes can offer empirical evidence that is relevant to the explanation of the outcome. In order to illustrate this point, assume that a government opts for peace. If the governments expectations prove correct (and nothing else changes in the country's environment), trust will remain at high levels, and foreign relations will continue to be solid. However, this is irrelevant to the evaluation of the hypothesis that worries about the consequences of violating democratic norms accounts for peace. This decision shows only that the executive could correctly anticipate the future, but the accuracy of predictions made by actors concerns a research question different from the one dealing with the actual causes of peace.[6] For similar reasons, one should not draw any inferences from the observation that an anticipated process does not occur. All that matters is that the government decided for peace because of certain consequences that it anticipated would unfold, which is independent of what happens after the decision has been made.

The second difference between hypotheses on realized and anticipated processes is closely related to the first. In the analysis of realized processes, as was explained above, the hypothesized process is also the process that is subject to process tracing. As this does not hold true for anticipated processes, the question arises of what the process is that one should focus on in the empirical analysis. The democratic peace example showed that actors evaluate different potential processes before a decision is made because the expected consequences inform an actor's decision. It follows that the empirically relevant process is the *decision-making process* that brings about the outcome. As regards the empirical example, the decision-making process of the government finally led to opting for peace. The goal of process tracing is to collect multiple pieces of evidence demonstrating that the government weighed the consequences of the two available courses of action and decided for peace because of the concerns that war would have for the country's trustworthiness and foreign relations.[7] (A detailed discussion of causal inference in process tracing can be found in Chapter 7).

Third, the decision-making process that one reconstructs via process tracing can be visualized as a realized process. Figure 6.3 contains a simple version of a decision-making process that one can trace empirically. The trade war, which triggers the crisis that could lead to a military conflict, first prompts the government to meet and to discuss the proper course of action. Empirically, this means that the government's decision-making process is triggered by an *exogenous event* – a trade war – and not by the hypothesized cause (that is, concerns about the implications of war and peace). In a second step, the leaders of both countries come together for diplomatic consultations, each leader hoping to determine if the other country appears to be committed to democratic norms. With the insights gained from this summit, the governments meet again and decide for de-escalation and peace.[8]

Although this figure looks similar to Figure 6.1 on tax competition, there are salient differences between the two illustrations. In the empirical analysis of anticipated processes and, thus, decision-making processes,

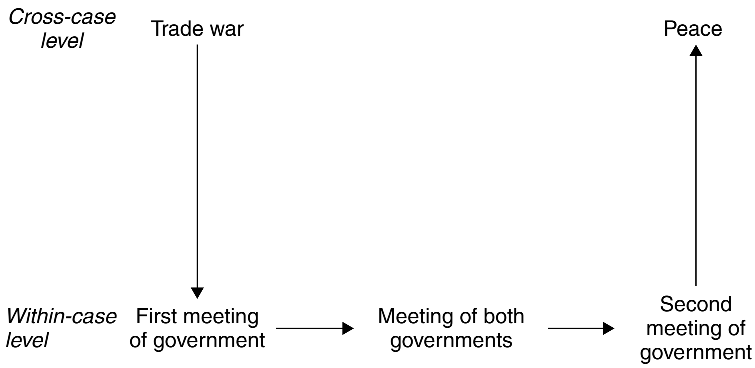

Figure 6.3 Visualization of the decision-making process

Table 6.1 Characteristics of hypotheses on realized and anticipated processes

Characteristic	**Realized processes**	**Anticipated processes**
Starting point of process	Hypothesized cause	Exogenous event
Relevant process for process tracing	Hypothesized process	Decision-making process
Meaning of arrows in visualizations of observable process	Causal influence	Temporal order
Fixed sequence of intervening steps	Can be hypothesized and is relevant	Not relevant

the arrows leading from one phenomenon to the next represent temporal order but not causal influence. Empirically, it may be that the governments decided at the first meeting that the countries' leaders should confer and that these decisions are a cause of the meeting. However, the theoretical inquiry of anticipated processes does not demand it to theorize a sequence because it is empirically irrelevant whether a summit took place after the first meeting or the second meeting or never. Instead, at every single step of the decision-making process, one should search for evidence that the political actors reflected upon the hypothesized anticipated processes. For example, this research calls for interviews with government officials and the analysis of protocols of the meetings in order to see whether the consequences of war and peace were weighed by the governments. Fourth and as a corollary of the third point, the theoretical analysis of anticipated processes does not demand it to specify a specific sequence of intervening steps. The differences between realized and anticipated processes are summarized in Table 6.1.

On the basis of this comparison, four additional remarks on the distinction between realized and anticipated processes are in order. First, one should have a conceptualization of process tracing that applies regardless of the type of process that one theorizes. A definition meeting this criterion is that *process tracing is a method for the collection of causal process observations in order to be able to reconstruct the process that leads to the outcome of interest.* The first attribute of the definition – a method for the collection of CPOs – underscores that process tracing aims at a specific type of data that is not amenable to a cross-case analysis (see Chapter 2 and Collier 2004b). It is worthwhile to emphasize that process tracing is a tool for the collection of CPOs used for the generation of causal inferences.[9] The second attribute of the definition – reconstructs the process leading to an outcome – is deliberately held in broad terms because it does not impose any constraints about the type of process that one is interested in.

Second, the two variants of processes that one can theorize are ideal–typical and one can think of hypotheses and process tracing taking intermediate positions. This point can be exemplified with a modification of the democratic norms example. Suppose a hypothesis states that democracy accounts for peace between two countries because of democratic norms. This hypothesis reflects neither a realized process nor anticipated processes because no argument about actor expectations is involved. Still, process tracing would focus on the decision-making process in order to gather causal process observations indicating that norms mattered for conflict resolution. For example, a CPO could be an interview statement of a government member: 'We did not attack the other country because of our commitment to democratic norms.' Although an argument about anticipated processes is implicit in the proposition that democratic norms matter, the hypothesis formulated above does not meet the criteria of one of the two ideals presented in Table 6.1.

Third, as complex as the debate about causal mechanisms is, the comparison of realized and anticipated processes indicates that the differences between them are not related to mechanismic theorizing and different conceptions of mechanisms. A substantiation of this assertion requires following a specific definition of 'causal mechanism', a difficult task, as so many definitions are on offer (see Chapter 2 and Hedström and Ylikoski 2010). In my reading, however, no conceptualization of mechanism is directly tied to what I call realized and anticipated processes. Since there are many different conceptualizations of mechanisms, I leave it with a discussion of two conceptualizations here.

In one perspective, a mechanism is invariant and instantiated in a given empirical situation (for example Waldner 2012). For instance, one could say that rationality is a mechanism (Mahoney 2003b), meaning that actors are rational and that rationality plays out when actors operate in a specific constellation.[10] This understanding of mechanism is compatible with both examples presented above. One can argue that it is rational for a government to respond to a tax cut of a foreign country, much as it is rational for it to be committed to democratic norms while involved in a conflict with another democracy.

The conceptualization of mechanisms in terms of entities and activities (Machamer et al. 2000) is equally applicable to the tax and democratic norms examples. In both processes, actors (the entities) make certain decisions and take actions (the activities) that result in the outcome. In light of the plethora of definitions of mechanisms, it is not possible to provide an exhaustive discussion of how mechanisms are related to the distinction between two types of processes. However, the consideration of two prominent definitions suggests that the differences between realized and anticipated processes are independent of the way in which mechanisms are defined.

Fourth, there is no inherent link between any empirical phenomenon and one of the two types of processes. The same subject matter can be theorized in terms of realized and anticipated processes alike. Theory therefore determines whether one examines an empirical phenomenon by following one perspective on processes or the other. Let me illustrate this with the tax competition example. (One could also think of a case study approaching the democratic norms argument from the perspective of realized processes.) In a hypothesis on realized processes, the cause of a tax cut in a given country, also called the domestic country, is a foreign tax cut. A case study that focuses on anticipated processes could have as the cause a country's goal to generate a sufficient level of tax revenue in the future. This goal can be threatened by a foreign country realizing a tax reduction *or* announcing to commit to one soon. An actual or potential foreign tax cut is an exogenous event that prompts the domestic government to think about the consequences of a stable and reduced domestic tax rate for tax revenue. The government thus weighs two anticipated processes and makes a decision according to how they are expected to affect tax revenue.

Continuing with tax competition, the case study now deals with anticipated processes with the consequences for theory and process tracing that are summarized in Table 6.1. There thus are two equally viable ways to theorize processes that create different demands for process tracing. Since they emphasize different aspects of the same subject matter, they also come to different conclusions; for this reason, a small-n researcher should be aware of the type of process that a hypothesis represents.

Realized and anticipated processes have been discussed in the context of hypothesis testing so far. However, it is equally important to be cognizant of this distinction in case studies that build and modify hypotheses because the exploratorily generated insights can refer to realized and anticipated processes alike. In formulating inductively generated causal inferences, one should clearly communicate the nature of the process in order to provide the best possible basis for a subsequent test of a corresponding hypothesis. For the latter purpose, one should also know the specific advantages and disadvantages that are linked to the two types of processes, the subject of the next section.

6.3 Challenges in research on realized and anticipated processes

The choice between the analysis of realized and anticipated processes should depend on theory because every phenomenon can be approached from two perspectives. This section shows that different problems and caveats are attached to inquiries into the two types of processes, problems that should be known in order to be clear about both the potential and limits of one's own analysis. Five issues are addressed in the following, the first three pertaining to challenges in case studies on realized processes. First, a hypothesis on realized processes is likely to be indeterminate as regards the reasons why an actor makes a specific decision (the mechanism, if you like). Second, hypotheses on realized processes that are built on the narrow idea of pattern matching are much more demanding than is currently acknowledged because one must explain why one and only one specific process should unfold. Third, hypotheses on specific sequences of intervening steps lack parsimony in the sense of requiring a lot of theoretical input for explaining a single causal chain. The fourth issue pertains to the depth of causal process observations that one needs for credible causal inferences, which is likely to be deeper in research on anticipated processes than on realized processes. Finally, it is important to emphasize that neither perspective on processes is superior to the other; they are simply two different possible views on the same empirical phenomenon.

First, by definition, hypotheses on realized processes remain silent on the actual reasons that a cause leads to a specific outcome. This is a problem because there often are multiple reasons why actors try to achieve a given outcome in face of the same exogenous event. In the tax competition example, a tax cut occurs because a foreign cut causes a decline of FDI and decreases tax revenue. The decision of the domestic government to lower taxes could reflects its goal of attracting more FDI and restoring revenue to higher levels in order to maintain its spending capacity. However, it is also possible that the decrease in FDI and revenue is taken as an indication of decreasing competitiveness of the domestic economy, which may also witness adverse consequences on the labor market. In this instance, lower tax rates are attributable to government concerns about competitiveness and the unemployment rate. One can think of more explanations – for example, that lower revenue threatens the country's security – that are all compatible with the realized process in Figure 6.1. The compatibility of the realized process with multiple motivations for the ultimate decision means that the hypothesis on the realized process is *indeterminate* as regards the reasons why the government responds to the foreign tax cut with a domestic tax reduction.

One may counter that the realized process in Figure 6.1 is coarse grained and that it is possible to theorize a much more specific process. Of course, the opportunity to discriminate between different actor motivations depends on the number and specificity of the intervening steps constituting the realized process (George and Bennett 2005, chap. 10). However, it should be difficult to single out all actor motivations but one by looking at realized processes only.[11]

Another objection might be that realized processes are simply not interested in the reasons why actors make a certain decision or take a certain action because this is the subject of process tracing on anticipated processes. This assertion is true (see above) but needs to be put into perspective by the argument that process tracing is often presented as a means to diminish indeterminacy (Bennett 2008; Campbell 1975; Hall 2008). Indeterminacy can be reduced as regards the process by which the cause leads to the outcome but it is unlikely to decrease with respect to the underlying actor motivations (see Section 7.2).

A second problem with the analysis of realized processes concerns the idea of pattern matching. In a narrow view one theorizes that multiple intervening steps occur in a specific order. In principle, it is possible to engage in this version of pattern matching, but it is likely to be more protracted than has been envisaged so far, the reason being that there are usually a large number of nontrivial sequences in which a cause can produce the outcome. Singling out one of these sequences requires the specification of a set of auxiliary conditions explaining why *one and only one* chain of steps should unfold. In set-relational terminology, the auxiliary conditions must be individually necessary and jointly sufficient for a particular sequence to occur.[12]

The tax competition example illustrates this argument. According to the realized process in Figure 6.1, one first observes that FDI declines. In the next step, tax revenue drops and in turn prompts the domestic government to cut taxes. Even when one is dealing with such a coarse-grained process, one can easily think of alternative, equally plausible ways for the outcome to occur. For instance, it may be that the domestic government expects that the decline in FDI will lead to a loss of tax revenue in the near to medium-term future. This expectation suggests itself because FDI is important to economic growth and tax revenue. If the government anticipates that declining FDI has this effect, why should it then wait until revenue drops? It may be that the government is uncertain about the effects of lower FDI and wants to wait before it alters its tax rate. But there are also good reasons to rush ahead and reduce the tax rate in order to stimulate FDI before revenue is reduced. Even if revenue declines before the domestic tax cut can take effect, a proactive tax cut should restore revenue sooner than a tax reduction that waits until income declines. Theoretically, it is therefore not clear why a tax reform should occur only after revenue has gone down. Consequently, a hypothesis that predicts this specific sequence – first lower revenue, then lower taxes – must explain under what conditions this happens as opposed to a tax reduction that precedes lower income.

Similar arguments can be made for the intermediate step 'lower FDI' and the cause as such. Foreign tax cuts are rarely prepared in secrecy, meaning that the domestic government should know that a foreign government is planning a tax reduction. Again, the question is why the former should wait until the adverse consequences of the foreign tax cut become empirical reality? If the government wants to prevent this outcome in the first place, it may implement a tax cut before the other country can reduce its tax rate. The hypothesis on the realized process in Figure 6.1 would be disconfirmed by this finding because the hypothesized outcome would occur *before* the cause. In contrast, a hypothesis predicting anticipated processes would find empirical confirmation because the preemptive domestic tax cut is made to avoid a decrease in tax revenue.

Jakobsen's study on the Danish liberalization of the electricity and telecommunications markets underscores that these arguments are not hypothetical but have relevance for process tracing (2010). Among other things, Jakobsen hypothesizes that Europeanization explains the liberalization processes in the two Danish sectors. One observable implication is that national liberalization begins after a corresponding European initiative was started. Jakobsen finds supportive evidence but also observes in one instance that Danish ministerial actors anticipated that liberalization in the telecommunications sector was shaped by what the European Union was expected to prescribe in the near future for telecommunications liberalization (2010, 901). We thus see an influence of the EU before a formal EU policy measure came into existence. Although this shows that Europeanization matters, it is disconfirming evidence for the hypothesis that national liberalization processes follow actual EU action.

The tax example and the Danish case study demonstrate that hypotheses on the order in which events occur must include auxiliary conditions as to why it is this order and not another one. Moreover and building on what was explained above, the Jakobsen example implies that hypotheses on anticipated processes do not care about the order in which events materialize. All one needs to show is that actor behavior was shaped by what was expected to follow from different courses of action. In Jakobsen's study, it seems that Danish ministerial actors weighed liberalization proposals that disregarded future EU action and that considered it, respectively. The former course of action was deemed too risky because if the EU prescribed a specific form and extent of liberalization, Denmark may have needed to adapt and commit to two reforms in a short period of time.

Third, these insights hint at the additional point that hypotheses on realized processes lack parsimony in the sense that they require much theoretical input to explain relatively little. Two reasons account for this. First, the specification of a cause such as a foreign tax cut limits the applicability of the hypothesis to cases wherein a foreign tax reduction occurred.[13] Second, complexity of the explanation is increased by the auxiliary conditions that explain why one particular process has to unfold. In total, one explains little – a single sequence of steps – with much theoretical input. Case studies on anticipated processes perform better on both ends. The tax example underscores that the empirical starting point is an event that is exogenous to the actual hypothesis. The government might weigh a tax cut against the status quo because a foreign country reduced its tax rate and FDI dropped, or because of a tax cut that is expected to influence FDI in the future, or because it is known that the foreign country plans to lower its taxes. The process that is theorized in a case study on realized processes therefore is only one process out of many that is compatible with a hypothesis on anticipated processes. Moreover, research on anticipated processes does not need to explicate a set of auxiliary conditions because the sequence in which intervening steps occur is irrelevant. For these reasons, hypotheses on anticipated processes tend to explain more with less theoretical input.

So far, the discussion has centered on the narrow understanding of pattern matching that is interested in the precise order of intervening steps. Research on realized processes following the more relaxed perspective on pattern matching is relieved of some of these problems. To recall, a broad understanding of pattern matching means that one derives multiple observable implications from a hypothesis, gathers empirical evidence, and evaluates the pattern in terms of the overall fit of expectations and evidence (see Section 7.2). The order of intervening steps is irrelevant in the broad variant, meaning that the hypothesis is compatible with a broader array of realized processes and that there is no need to stipulate auxiliary conditions. Still, a hypothesis on realized processes focuses on a specific triggering cause and presumes that the cause occurs before the outcome. The tax example and Jakobsen's case study show that this can prove incorrect if actors operate proactively. Consequently, even the broad view on pattern matching covers only a subset of possible processes that are compatible with an analysis of anticipated processes.

The fourth point on which the analysis of realized and anticipated processes differs is the requirement as regards the depth of the empirical evidence. The evidence that one needs to substantiate a realized process depends on the precision of the predicted intervening steps. However, one can reasonably argue that it is more difficult to gather evidence that actors made a decision because of what they expected to flow from it and from alternative courses of action that are not taken. As explained above, one should refrain from inferring anything from what happened after actors decided or operated in a specific way because doing so leads to the problem of assuming revealed preferences (Pierson 2000). Although problematic, empirical research such as Jakobsen's case study shows that it is possible to gather causal process observations that yield insight relevant to hypotheses on anticipated processes (Pahre 2005). Nevertheless, the requirements as regards the collection of unambiguous empirical evidence should generally be higher for case studies on realized processes.

Fifth, concluding this discussion, it should be emphasized that I do not make a plea for the exclusive analysis of either realized or anticipated processes. A case study researcher should decide about what he or she finds more interesting and should then handle the issues that are attached to case studies on one of the two variants of processes. One way to almost get the best of both worlds is a combined inquiry into realized and anticipated processes. Such a case study would theorize a realized process as well as the reasons why actors respond to the cause, thereby producing the outcome. For the tax example, a hypothesis could read 'A country reduces its tax rate in response to a foreign tax cut because of the belief that this will ensure tax revenue and the opportunity to engage in policy making at pre–foreign tax cut levels.' This hypothesis could be expanded by theorizing intervening steps that would occur before a government takes a certain action. Beneficial as an integrated analysis is, it has two limitations: one needs specific CPOs for the assessment of the predicted anticipated processes, and it is limited to instances in which an actual foreign tax cut triggered a domestic response. As detailed above, the latter restriction does not extend to hypotheses on anticipated processes; thus, an integrated inquiry into realized and anticipated processes has to make some compromises as well and invites the challenges attached to the analysis of the two types of processes.

6.4 Time and temporality

The previous discussion of realized and anticipated processes points, at least implicitly, to the importance of time and temporality for process tracing. Naturally, a temporal element is inherent in process tracing because a process unfolds over time (George and Bennett 2005, chap. 10). On the most basic level, the opportunity to collect multiple causal process observations over the course of a process allows one to increase the number of observations that are in line with one inference or the other.

Beyond a focus on the temporal unfolding of processes and the analysis of causal mechanisms over time, the idea of temporality can be invoked in a variety of ways in order to increase the theoretical leverage of a case study. Since different manifestations of temporality have been lucidly discussed before (Grzymala-Busse, 2011; Pierson 2004; Slater and Simmons 2010), I stay with a short, illustrative discussion of temporality here. Comparative Historical Analysis (CHA) particularly offers a rich set of time-related concepts such as sequencing (Büthe 2002), cumulative effects and outcomes (Pierson 2004), critical antecedents (Slater and Simmons 2010), critical junctures (Capoccia and Kelemen 2007, Hogan and Doyle 2009; Soifer 2010), and path dependence (Immergut and Anderson 2008; Thelen 1999). Although these concepts are closely linked to CHA and historical institutionalism, the basic ideas behind these notions of temporality might travel to case studies that are rooted in a different theoretical framework.

As an illustration of a specific concept of temporality and the challenges it creates for process tracing, I focus attention on the concept of critical antecedents in the following (Slater and Simmons 2010). This concept must be seen in combination with the more established concept of critical junctures (Collier and Collier 1991). Critical junctures describe (usually short) periods of time during which several favorable conditions come together to give actors the opportunity to engage in significant policy or institutional change. For instance, the fall of the Berlin Wall in 1989 is a critical juncture to the extent that it opened the opportunity for the reunification of West and East Germany. In 1989 and 1990, Helmut Kohl, then German chancellor, seized the opportunity before the chance for reunification vanished again.

For some time, CHA was concerned with the question of what a critical juncture is and under what conditions it goes along with change, on the one hand, and inertia, on the other (Hogan and Doyle 2009). The notion of critical antecedents goes one step back in time and asks what conditions might increase the likelihood that the process of interest will develop in one direction or the other (Slater and Simmons 2010). The global financial crisis of 2007 and 2008 can be used to exemplify the meaning of critical junctures and critical antecedents. After a long period of liberalization in the finance sector, the recent financial crisis created the opportunity for the reregulation of this sector so as to prevent future crises of similar scale. Although regulations were tightened to some degree, massive state intervention in the sector, as demanded by the public, did not come into existence in 2008, 2009, or 2010, which are the years during which reregulation seemed possible. As it seems, a major reason why regulations were not tightened more – the outcome to be explained – had to do with the lobbying activities of the financial sector. In this reading, lobbying of antiregulation groups explains why the critical juncture was not followed by major policy change.

A case study on critical antecedents tries to explain why massive lobbying could play such a decisive role during the critical juncture and inhibited tighter regulations. One could argue that one reason why the financial sector played a vital role during the critical juncture was the economic importance of banks, hedge funds, and so on, in some countries, which is in turn attributable to financial deregulation and growth of this sector since the 1980s. Financial liberalization thus contributed to the crisis and opened the critical juncture (which is not necessarily a feature of critical antecedents) but at the same time created private actors so strong that they could work against reregulation during the critical juncture. It seems justified to argue that lobbying did not determine what happed during the critical juncture because major change seemed possible immediately after the crisis. However, the economic importance of banks and lobbying power increased the chances that major policy change could be prevented, which meets the understanding of critical antecedents by increasing the likelihood that the process during a critical juncture is channeled into a specific direction.

Having illustrated the concept of critical antecedents, the question is what the implications for process tracing are. First of all, it is evident that temporality is involved here: critical antecedents are temporally prior to the causes that are operative during a critical juncture, and these causes, in turn, explain why a process took the direction that it did (no tight regulation in the case of the financial crisis). In order to substantiate the argument, process tracing would need to deliver multiple pieces of evidence showing the following: first, that deregulation in the 1980s successively contributed to the growth of the financial sector, increasing its economic importance in some internationally influential countries such as the United States and the United Kingdom; second, that growing economic importance generally increased the political influence of financial actors and lobby groups; third, that massive lobbying took place during the financial crisis; fourth; that these lobbying activities eventually accounted for the opposition of some countries toward tighter regulation; fifth, that less deregulation (or none) in the 1980s would have led to lower growth of the financial sector, less political influence, and less powerful lobbying activities during the critical juncture, ultimately raising the chances for policy change.[14]

This list of observable implications, which could be expanded, illustrates that temporality and specific manifestations, such as critical antecedents, imply challenges that are common to case study researchers. First, the theoretical arguments that are attached to concepts such as critical antecedents and path dependence must be translated into specific, observable implications that one can examine via process tracing. In the first step of the empirical example, this involves tracing the link between a liberalized financial sector and its growing economic importance and, as a consequence of that, increased political influence. Similar and common process tracing tasks apply to the other three steps that establish a temporal sequence of events.

Second, the arguably more challenging task is to present evidence that the outcome would have been different if the critical antecedent would have been different. Since the empirical example is not a comparative case study but only about the critical juncture following the financial crisis, one essentially needs evidence that substantiates a counterfactual (see Section 7.3). This second task is again common to case studies and does not present a new challenge per se. Causal arguments involving notions of temporality obviously have different implications than simple nontemporal claims and therefore might raise the bar when it comes to causal inference. For instance, the argument that the chances for policy continuity increase with increasing lobbying activities is relatively easier to examine than a proposition that additionally invokes critical antecedents attributing policy inertia in the years prior to 2010 to deregulation in the 1980s. This is certainly not a plea against critical antecedents, but one should be alert to the implications that elaborate causal arguments about time and temporality have for process tracing causal inference.

6.5 Conclusion

Process tracing lies at the heart of many case studies and is an indispensable tool for causal inference. This chapter aimed to highlight two issues that are specific to process tracing. First, in order to achieve a fit between theory and method (Hall 2003), one should be cognizant of whether one is interested in realized or anticipated processes. The type of process has theoretical and empirical ramifications that were outlined in the first two sections of this chapter. In addition, the chapter reflected on the role of time and temporality for process tracing. On the most basic level, the longitudinal perspective that one takes in process tracing allows one to collect numerous causal process observations and, if the case study is about realized processes, a sequence of events. In addition, the subject matter might permit it to introduce notions of temporality in the form of concepts that are particularly common to CHA. With respect to causal inference, temporality-related concepts increase the theoretical leverage of a case study beyond inquiries into causal effects and causal mechanisms. In order to achieve more leverage, however, it is essential that one carefully considers the observable implications that are attached to a specific concept. The analysis of temporality-bound causal arguments tends to be more demanding than nontemporal ones precisely because one makes claims about how multiple events unfold and are related to other events over time. The invocation of time and temporality thus is as promising as it is challenging because of increased requirements as regards the alignment of theoretical claims about temporality and process tracing (Hall 2003).

7

From Evidence to Inference: Use of Sources and Counterfactuals

The challenge of generating internally valid explanations is to separate the causes that bring about the outcome from those factors that are not of general relevance (Geddes 2003, chap. 1, King et al. 1994, chap. 3). This chapter and the following one provide a comprehensive discussion of causal inference and internal validity, partially drawing on the previous three chapters on cross-case and within-case analysis. The discussion is split into two chapters to highlight the different stages and elements that play a role in the generation of causal inferences. Chapter 8 presents an in-depth discussion of frequentist and Bayesian causal inference and the similarities and differences between them. The present chapter describes the methods-related steps and problems that one confronts in the collection of evidence and in inferring that a cause or causal mechanism is or is not operative. These issues are common to frequentist and Bayesian causal inference and thus are dealt with in a separate chapter. Moreover, Chapter 8 will show that it is analytically useful to distinguish the inference that a cause or causal mechanism is causally relevant from the implication that such an inference has for the hypotheses under scrutiny. Bayesian causal inference, in particular, does not stop with a positive or negative inference but uses it to reflect on one's confidence in the hypothesis to which the inference is related.

Section 7.1 lays the foundation for the discussion of frequentism and Bayesianism and starts with what has become known as the *source coverage problem* and the *source coverage bias* (Lustick 1996). The general issue that is elaborated upon is that each source provides only a probably small and biased part of the empirical picture that one needs to gather for valid causal inference. In Section 7.2, I reconsider the problem of indeterminacy with the specific focus on causal inference via process tracing. Process tracing is often presented as one means of diminishing, if not eliminating indeterminacy at the cross-case level because of the opportunity to collect causal process observations (CPOs) that far outnumber the number of data set observations (DSOs) at hand. However, it is shown that this line of reasoning conflates the number of collected CPOs with the number of inferences that one makes in process tracing. Once one focuses on the number of inferences, the potential of process tracing to eliminate or at least diminish the extent of indeterminacy hinges on the question of whether the outcome is attributable to one or multiple causes, which is simply an empirical question.

Presuming that more than one causal inference remains viable after process tracing has been done, one can rely on *counterfactual reasoning* for the strengthening of causal inference. Counterfactuals are needed if the evidence at hand does not allow one to assess all potential causal inferences empirically because of a lack of appropriate cases. Although the value of counterfactuals is not undisputed, counterfactual analysis can be a valuable tool if it follows certain guidelines that are detailed in Section 7.3 of this chapter.

7.1 Source coverage problem and bias

Case study researchers derive their data in the form of CPOs drawn from a variety of sources. A common classification of sources distinguishes between primary sources, secondary sources, interviews, and newspapers (George and Bennett 2005, chap. 5). Regardless of whether one would like to distinguish more or fewer types of sources, it is important to be aware of what has become known as the source coverage problem and the source coverage bias. The goal of empirical research must be to acquire an empirical picture of the process and phenomenon of theoretical interest that is as complete as possible. The collection and evaluation of sources is the means of putting the picture together. The fundamental problem that one confronts in achieving this goal is that every source covers only a certain fraction of the relevant empirical evidence. Thus, reliance on a single source leaves one with an incomplete picture of what occurred in the case of interest.

Let us suppose that you are interested in the explanation of trade liberalization and hypothesize that the lobbying of exporters plays a crucial role (see Dür 2010). One such case is the US Reciprocal Trade Agreements Act (RTAA) of 1934, which marked the beginning of the United States' turn to liberal trade (Haggard 1988). The political process that resulted in the RTAA continued for over a year and offered exporters many opportunities to lobby for liberalization vis-à-vis the government. Now imagine that you gain access to numerous primary documents from the second half of the process and that you do not find any evidence of exporter lobbying. Would you reject the hypothesis that the lobbying of exporters is not relevant for making sense of the RTAA? This may be the correct conclusion, but it may also be that exporters lobbied successfully very early in the political process. Once they put US trade policy on the track of liberalization, there was no further need to lobby, which accounts for the absence of evidence in the part of the process that is covered by your sources. This example points to the central question that one needs to carefully consider in the evaluation of sources: when can I take the absence of evidence as the evidence of absence of a cause (Sober 2009)?

In the given example, one may sense that there is a problem with the available primary sources because it is easy to determine that they do not cover the first half of the process at all. However, a more balanced coverage of the process does not guarantee that you get a better understanding of what happened as long as parts of it are not covered by your sources. Let's assume that you have access to the protocols of every second meeting at which members of the US administration discussed the RTAA, which would be a quite voluminous body of documents. In this example, the overall coverage of the process would be good. But how do you know that the other half of the protocols includes information that would lead you to making the same inferences as for the half that you can access? This example shows that a sampling of sources is fraught with severe inferential problems regardless of what sampling rule you use.

Besides the fact that every source yields only a small share of all relevant information, it is likely that the information that it offers is *biased*, thereby introducing a source coverage bias into the analysis (Lustick 1996, 606). Continuing with the example of the RTAA, suppose that you rely on a historical book concerned with the link between a country's security policy and its trade policy. Since the secondary source is about the impact of security policy, it should come as no surprise that it deals barely, if at all, with organized interest groups as another potentially relevant factor that drives trade policy making. Even if the secondary source is a piece of high-quality empirical research, it is a biased source *with respect to the hypothesis of interest* because its own research focus is at least partially different. This means that the source coverage problem is not one of low-quality sources (though this may be the case). The bias is simply due to the fact that both studies are motivated by different research interests.

This insight can be generalized to other types of sources by stating that each source must be put into context. A primary source such as a protocol of a meeting is not a neutral documentation of the meeting but may deliberately include some information while leaving out other, sensitive information. Similarly, interview partners may have be prepared to share some pieces of information while not reporting others. Therefore, an assessment of the interests and involvement of the interview partner in the phenomenon of interest is important in order to evaluate the quality of the interviewee's statements.

The fact that the source coverage problem and bias pertain to all types of sources underscores that no one source is superior to another one. Because of this and the fact that the empirical picture becomes clearer the more observations one gathers from disparate sources, it is a common recommendation to *triangulate* sources (Lustick 1996; Thies 2002). Instead of relying on a small number of sources that probably are all of the same type, one triangulates the information that one derives from a diverse set of independent sources. For example, this means that one cross-validates the statements of an interviewee with information derived from primary sources, and vice versa. Similarly, one can use secondary sources and newspapers for the contextualization of the process of interest. In an analysis of trade policy making, these two types of sources may allow one to determine the specific period during which the government actually decided to liberalize trade. Such information then permits one to narrow down the time frame and the number of primary sources that must be read. Although triangulation is clearly superior to the reliance on a single type of source, one must clearly acknowledge that some uncertainty about the presence and extent of a source coverage problem and bias will always remain. This uncertainty should be taken into account in the context of deriving causal inferences because they cannot be more certain than the confidence in the accuracy of the underlying evidence.

7.2 Causal process observations and the indeterminacy problem

The collection of multiple causal process observations from a broad range of diverse sources plays a central role in the debate about indeterminacy in case studies (Bennett and Elman 2006, 459). When the within-case analysis, serves only as a means to measure the score of a cause on the cross-case level (see Chapter 1 and Mahoney 2010), the indeterminacy-related issues discussed in Chapter 4 and 5 apply because then the cross-case level is the theoretical end. The situation is different when process tracing (also) aims at the collection of causal process observations in order to distinguish between competing hypotheses on the within-case level. This difference is rooted in the distinction between CPOs and DSOs (see Chapter 2 and Collier et al. 2004b). The number of DSOs is small in qualitative case studies because they are tied to the cross-case level, but the number of CPOs can be very large because they are located on the within-case level (Campbell 1975).[1] Since the indeterminacy problem is in essence one having few DSOs and many potential causes, the collection of multiple CPOs seems to offer the leverage for diminishing if not eliminating this problem (Bennett 2005, 29).

In case study research, the importance of the distinction between different types of observations cannot be overestimated (Mahoney and Goertz 2006, 241–2). At the same time, however, one should be careful not to make too much of the fact that one can gather a large number of CPOs via process tracing. The elaboration of this admonition demands a reflection on how process tracing is performed in practice. Three steps can be distinguished, steps that reflect the nature of all theory-centered empirical research (Brady et al. 2004; George and Bennett 2005, chap. 10; Lave and March 1975). First, one formulates hypotheses and derives multiple observable implications from each hypotheses in order to increase the inferential leverage of the analysis (King et al. 1994, chap. 1). Second, one collects causal process observations via process tracing. Third, one subsumes the observations under the previously formulated observable implications and evaluates the fit between expectations and evidence (see Chapter 8 and Adcock and Collier 2001). Evidently, the first step does not apply to hypothesis-building and hypothesis-modifying case studies because they do not start with the formulation of a hypothesis, but the second and third steps apply to exploratory case studies as well. For purposes of illustration, however, the focus is on hypothesis-testing case studies in the following.

The second and third steps of the three-step procedure particularly show that observations do not speak for themselves but that they must be tied to observable implications and are instrumental for making inferences. More precisely, the purpose of a within-case analysis and the collection of CPOs is to infer that there *is or is not* a causal mechanism operative between a specific cause and the effect of interest. Process tracing therefore comes down to making a single inference in the end about the presence or absence of a mechanism (Gerring 2007a, 173; Munck 2005, 4).

The fact that CPOs serve to make a single inference on a causal mechanism can be exemplified with a modification of the tax competition example introduced in Chapter 6.[2] In addition to the tax reduction of a foreign country, one develops the hypothesis that the diffusion of liberal economic ideas causes the domestic tax reduction. The stylized process by which diffusion causes domestic tax reduction is the adoption of liberal

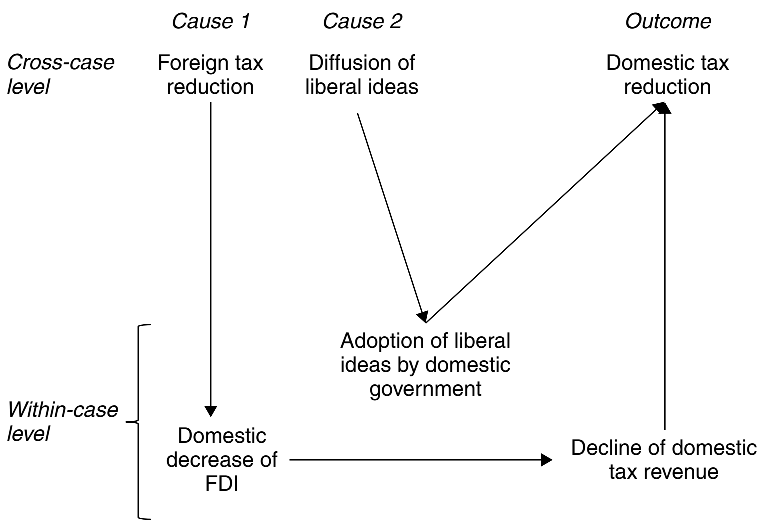

Figure 7.1 Illustration of process tracing and indeterminacy

economic thinking by the domestic government. Figure 7.1 visualizes both a hypothesis and a very simple causal process connecting the respective causes to the outcome (The respective intervening steps are all located on the within-case level. For simplicity of presentation, they are not horizontally aligned.)

Imagine now that one examines a single country on the cross-case level and that both a foreign tax cut and diffusion of liberal ideas are observable. Furthermore, process tracing then shows that the hypothesized process between a foreign tax reduction and a domestic tax reduction is in place. Ceteris paribus, the more supportive CPOs one can collect for each of the hypotheses, the more credible the inference becomes, but it nevertheless remains a single inference (see Chapter 8). What do we make of the insight that a causal mechanism (however defined) underlies each of the two cross-case relationships? This is a valuable finding, but that's exactly where we started; on the cross-case level, two causes qualified as causal before process tracing, and a foreign tax cut and liberal economic thinking still count as causal after the within-case analysis that discerned the respective causal mechanisms. Regardless of how convincing the within-case analysis is and how numerous the CPOs (see also Chapter 8), process tracing does not diminish indeterminacy in this instance. This would be different if one were unable to gather CPOs supporting the tax hypothesis or the diffusion hypothesis because then one could infer that only one of the two causes is connected to the outcome.

This example shows that it is an *empirical* question as to whether process tracing can diminish or even eliminate indeterminacy. When one collects evidence that multiple causes are connected to the outcome via a causal mechanism, that's simply the way it is. Indeterminacy continues to be a problem because the cases at hand do not allow one to discriminate between the mechanisms. On the other hand, if one is lucky enough to infer that there is no causal mechanism operative between one potential cause and an effect, the extent of indeterminacy declines because one hypothesis fails to receive support on the within-case level. Since the link between process tracing and the reduction of indeterminacy is an empirical one, there is not much logic in getting involved in principled arguments about the value of process tracing simply because the answer is 'It depends'.

So far, the discussion has focused on what was called effect-related indeterminacy in Chapter 4, meaning that the cross-case pattern is compatible with two different inferences on what the causes of an effect are and what the nature of their effect is. The previous arguments fully extend to *mechanism-related* indeterminacy. To review, mechanism-related indeterminacy denotes a situation where we make valid inferences about causal effects, but the cause in question can bring about the outcome via multiple causal mechanisms. A case in point for mechanism-related indeterminacy is the democratic peace phenomenon because peace could be attributable to the fact that the political elite is committed to democratic norms or that the political actors are constrained by a public opposing war, just to mention two viable explanations (Rosato 2003). In principle, process tracing can shed light on this because it may hint at a single element of democracy and corresponding mechanism that bring about peace. This is not necessarily the case, however. In a comparative test of different causal mechanisms related to democratic norms, institutional constraints, and so on, one may find confirming evidence only for one mechanism. However, the mechanisms are not mutually exclusive: democratic norms may prompt peace much as an institutional constraint on the executive might. Thus, process tracing could also deliver supportive evidence for two or more causal mechanisms. In total, these examples show that effect-related and mechanism-related indeterminacy might disappear in the light of causal process observations but not necessarily because this is an empirical question.

The claim that process tracing results in a single inference as to whether or not a mechanism is operative also has important implications for the debate about Mill's methods and cross-case comparisons more generally. In Chapter 4, it has been explained that case studies have been long criticized for suffering from indeterminacy in the generation of cross-case inferences. In addition to the argument that process tracing can diminish indeterminacy, one finds the claim that case study researchers do not generate cross-case inferences at all but are interested in and concerned with process tracing and inferences on mechanisms and processes (Goldstone 1997). The focus on the within-case level certainly is a feature of process tracing, but in producing an inference on a mechanism connecting cause to outcome, one cannot but infer that this cause has an effect on the outcome. This close link is due to the fact that a causal mechanism underpins a causal effect (see Section 2.2), which implies that the inference on a causal mechanism and a basic inference on an effect cannot be disentangled. The argument that causal inference via process tracing allows one to ignore the problems of cross-case inferences therefore is an illusion.

The tax example presented above illustrates this claim. Before process tracing has been done, one is uncertain about whether a foreign tax cut and the diffusion of liberal ideas have an effect on a domestic tax cut. The collection of causal process observations via process tracing then permits it to explain how each of the factors brings about the come (presuming we find supportive evidence). At the same time, the evidence is taken to infer that a mechanism is in place and, consequently, that each of the causes seems to have an effect on domestic tax policy. The inference on the causal mechanisms is therefore automatically linked to an at least very basic inference as to whether an effect is in place or not. As elaborated above, it *seems* as if they have an effect because the single-case study does not allow discrimination between the two determinants of a domestic tax cut. This in turn has implications for the within-case inferences because a mechanism is causal only when it underlies a causal effect (much as the other way round). The inability to make an unambiguous inference about causal effects thus feeds back on the within-case level and translates itself into indeterminacy as regards the underlying causal mechanisms.

In conclusion, I note that the discussion of causal process observations and indeterminacy is often tied to the idea of Bayesian causal inference. In this view, the potential of process tracing to overcome the indeterminacy problem specifically hinges on the opportunity to gather confirming causal process observations that were unlikely to be collected because the discovery of surprising supportive evidence is more valuable than confirming expected evidence. As explained previously, systematic reasons account for reserving a discussion of Bayesianism for the following chapter. Besides the fact that Bayesian causal inference is limited to hypothesis-testing case studies, one can and, in fact, should initially rely on counterfactual reasoning in order to diminish indeterminacy. Even if one is concerned with hypothesis-testing, it is recommended that one first engage in counterfactual analysis because frequentist and Bayesian causal inference becomes easier the more causal inferences can be rejected beforehand via counterfactual thinking (see Chapter 8).

7.3 Counterfactual reasoning

Counterfactual analysis addresses the indeterminacy problem, meaning a lack of appropriate evidence, by adding hypothetical cases to the analysis of observed cases, also referred to as factual cases in the following (see Northcott 2010). A counterfactual case is obtained by manipulating a factual case in a specific way (following the guidelines described below) in order to assess whether the manipulation would have made a difference. If one phrases it as a question, a counterfactual asks, what would the outcome have been if the factual case were different in a certain respect? Counterfactuals are common in everyday reasoning – if I would not have been in a hurry, I would not have had a car accident – and the social sciences alike. In the social sciences and history, as well (Bunzl 2004, 845), counterfactual reasoning is not without its critics because one relies on fictional cases instead of observed ones.

While one may be skeptical about incorporating unobserved cases into the analysis, the literature on counterfactuals has developed criteria that distinguish good counterfactuals from bad ones (Lebow 2010). Five criteria are distinguished in particular in the following paragraphs. The first one concerns the *transparency* of what the relevant counterfactual state of interest is and what the consequence is. The second and third principles deal with the quality of the manipulation of the cause and cover the *minimum rewriting rul*e and *empirical plausibility of the manipulation*. The remaining two guidelines pertain to the consequences that one derives from manipulation. They call for the *use of theoretical and empirical knowledge* for the substantiation of the consequences and for the *empirical plausibility of the consequences* that follow from the manipulation.

First of all, there is a twofold need for clarity in counterfactuals. It is evident that one should specify in what way the observed case is manipulated in order to obtain a counterfactual case. Imagine that you are interested in the decision-making capacity of federal and unitary countries. The hypothesis is that unitary countries have a greater capacity than do federal countries because decision making is hampered in federal states owing to the involvement of more actors. In the empirical analysis, you observe a federal country and its decision-making capacity (measured, for example, in terms of the time it takes to enact counterterrorist measures in response to a terrorist attack). In the absence of an observed unitary country, you must construct a counterfactual unitary state for an assessment of the hypothesis under scrutiny. This is achieved by manipulating the score on the variable 'territorial organization of the state' and changing it from 'federal' to 'unitary'.

On the basis of the manipulation, the second element of clarity pertains to the goal of the counterfactual analysis (Lebow 2000, 583). Two different goals are on offer in theory-centered case studies; one can try to discern if the manipulation of a cause leads to a different outcome or the same outcome (probably at a different point in time). In the former instance, one wants to show that the manipulated cause qualifies as a cause indeed, while, in the latter scenario, it is the goal to demonstrate that the outcome would have occurred anyway.[3]

The second criterion of good counterfactuals concerns the *minimum rewriting rule*, which states that one should change as little as possible about the factual case. In other words, the counterfactual case should be as close as possible to the observed case and differ in one respect only (Hawthorn 1991, chap. 1). The relevance of this criterion is evident in light of the discussion of causal inference in Chapter 4 and Section 7.2. The more we alter the factual case, the more difficult it becomes to assess what the outcome would be in the counterfactual case and the harder the causal inference is to defend. With an eye to what was discussed in Chapter 4, the minimum rewriting rule avoids the introduction of indeterminacy into a counterfactual analysis deriving from a suboptimal comparison. When one asks what the decision-making capacity would be in a unitary country, one effectively creates the method of difference with respect to the causes because the observed and counterfactual cases differ only in respect to the form of territorial organization. If we now additionally consider what the decision-making capacity would be if the federal country has a low GDP as opposed to a high one, we are comparing a rich federal state with a poor unitary country. This is a suboptimal comparison when the method of difference and minimal rewriting are the goal. Theoretically, it may be that one is interested in the joint effects of unitarism and low GDP on decision making and that the minimum number of causes to be manipulated is two. However, it should be clear that the counterfactual analysis of a poor unitary country implies two further counterfactuals: one involves a poor federal country and the other a rich unitary state. These two counterfactuals adhere to a strict minimum rewriting rule and should augment the counterfactual case of key interest, which differs on two variables from the observed case.

The third criterion asks for the feasibility of an isolated manipulation and concerns the plausibility of the manipulation in light of the broader empirical context in which the factual case is embedded. In the leading example, the isolated manipulation of the territorial organization of a country from federal to unitary may be hard to maintain. The country once installed federalism for a specific reason that may have driven the choice of other institutions as well. Indeed, federalism often coincides with second chambers, corporatism, and a powerful central bank (Lijphart 1999). These institutions constrain the government and force it to engage in consensual policy making, which could be the reason that constraining institutions often occur together. In this perspective, a counterfactual that seeks only the consequences of unitarism lacks plausibility because it is unlikely that this country could be unitary while still having a corporatist system of interest mediation, a second chamber, and so on. A counterfactual of the decision-making capacity in unitary countries would therefore require the manipulation of more institutional characteristics than would federalism alone (Lebow 2000, 584).

It is apparent that this principle is in conflict with the minimum rewriting rule (Lebow 2000, 582). The rule stipulates that there should be as little change as possible, whereas a plausible counterfactual may call for extensive differences between the observed and counterfactual case. If one takes a broader definition of what the 'minimum' is that one rewrites, the two rules become less contradictory because the plausibility of the manipulation determines what the minimum is that must be changed. In any instance, the tension between the minimum rewriting rule and the feasibility principle needs to be handled transparently by a case study researcher.

The final two criteria of good counterfactuals relate to the quality of the consequences that one derives from the manipulation of the observed case. The fourth guideline requires using theoretical and empirical knowledge so as to lend credence to the counterfactual inference (Lebow 2000, 583). Admittedly, there is little established knowledge in the social sciences. Nevertheless, this should not excuse one from augmenting one's conclusions, to the extent that it is possible, with supplementary evidence. For instance, one could use insights from research on group decision making in order to determine if the involvement of fewer actors in public policy making is likely to enhance or impede its progress.

The fifth and final criterion broadly refers to the empirical consistency of the consequences that one derives from the manipulation. Lebow (2000, 584) specifically draws attention to the co-tenability of the counterfactual inference, the interconnectedness of causes and outcomes, and second-order counterfactuals. Co-tenability demands that attention be paid to the possibility of equifinality in counterfactual analysis. As regards the empirical example, at first glance, it seems plausible that a country could produce policies more quickly when it is unitary. However, one could argue that those actors with stakes in public policy making – political parties, lobby groups, trade unions, and so on – would try to access the policy-making process at some other point. Unitarism would deprive these actors of one point of access, but they would adapt, and the country's policy-making capacity would remain unchanged.

Lebow (2000, 584) gives an illuminating example of what he calls the interconnectedness of causes and outcomes. The question is whether the Cuban missile crisis would have resulted in war between the United States and the Soviet Union if Richard Nixon had been president instead of John Kennedy. In this example, the cause is the incumbent president – Kennedy or Nixon – and the outcome is the consequence of the missile crisis – war or peace. One could argue that Nixon would have taken a more aggressive stance and war would have resulted. However, the Bay of Pigs invasion took place in 1961, that is, in the interim between Nixon (counterfactually) assuming the presidency and the missile crisis. The invasion failed under Kennedy, but Nixon may have taken a more determined stand in 1961 and overthrown Castro. The deployment of Soviet missiles from Cuba would have become impossible, thereby demonstrating that the manipulation of a cause may lead to consequences in the counterfactual causal chain that make it impossible for the actual outcome to occur.

At the same time, one should consider second-order counterfactuals. This term denotes that the manipulation of a cause may put the counterfactual on a different track at first but ultimately return it to the original path. In the Nixon example, a successful Bay of Pigs invasion would have made the Cuban missile crisis untenable. But if Khrushchev had been determined in his actions, he may have found another way to threaten the United States. At that point, Nixon would have taken a forceful stand in this conflict, and war would have resulted from an incident other than the Cuban missile crisis.

This discussion shows that counterfactual reasoning is not simply an act of imagination but a technique for disciplined thinking about alternative ways in which events in the past could have transpired. Under the premise that one follows the existing guidelines and makes the manner in which one arrived at counterfactual conclusions sufficiently transparent, a counterfactual is a valuable tool in the hands of case study researchers.

7.4 Conclusion

The road leading from the collection of evidence to the generation of causal inferences is a long and winding one. The basis for sound causal inference is always the amount and quality of the evidence collected via process tracing (see also Chapter 8). It is for a good reason that social scientists should follow the standards of historical research in the data-gathering part of the case study because high-quality process tracing is necessary though not sufficient for the generation of valid causal inferences.

If one is still confronted with indeterminacy following process tracing, counterfactual reasoning can prove valuable to diminish that indeterminacy if it is done in a disciplined way. Given the relative weakness of theory in the social sciences, counterfactual analysis is likely to be the rule in empirical small-n research. When the process tracing evidence is sufficiently strong and counterfactual analysis can reduce the number of viable causal inferences to a single one, a case study researcher is in the lucky position of avoiding indeterminacy. As has been shown in this chapter, it is an empirical question and a matter of whether an outcome can be ascribed to one or multiple causes.

Regardless of whether one or multiple causal inferences can be read into the evidence, counterfactual thinking is not the end point of producing causal inferences. The collection of evidence and counterfactual analysis allows one to infer that a causal mechanism is or is not operative in the cases under analysis and what the nature of the causal effect is. The implications of such an assertion for causal inference depends on whether one follows the frequentist or Bayesian modes of causal inference, which is the subject of the next chapter.

8

Frequentist and Bayesian Causal Inference in Tests of Hypotheses

The previous chapter dealt with challenges that one confronts in collecting and using evidence for the generation of causal inferences. The challenges and means to address them pertain to case studies that build hypothesis, test them, or seek to modify them in order to make sense of puzzling cases. A separate and important topic reserved for this chapter concerns frequentist and Bayesian causal inference as two ways of producing inferences in tests of cross-case and within-case hypotheses. Frequentist causal inference is based on the number of observations and the premise that the more supportive or disconfirming observations one collects, the stronger causal inferences are. Bayesianism emphasizes the theoretical impact and likelihood of collecting individual observations as opposed to their number. Among other things, a discussion of these two modes of causal inferences closes the circle with respect to Chapter 3. As is detailed below, distribution-based case selection is integral to frequentist causal inference, whereas Bayesianismrelies on the theory-based choice of cases.

The chapter starts with a brief review of two prominent arguments on causal inference and the use of process tracing for the evaluation of hypotheses. The review first focuses on the role of frequentism and Bayesianism. Bayesianism also plays an important role in the second argument that builds on van Evera's typology of four types of hypotheses tests (1997, 31), a typology recently rediscovered by the case study literature (Bennett 2008, 2010; Collier 2010, 2011). One dimension of this typology captures whether the certainty of a prediction is high or low, which is the Bayesian element, and whether a certain prediction is or is not unique to one hypothesis. Section 8.2 starts with a critical appraisal of the dimension of uniqueness.

In Sections 8.3 to 8.5, I reconsider the existing claims on Bayesianism v. frequentism and deliver an in-depth discussion of formalized Bayesian causal inference. The common basis for these sections is the reconstruction of the research process with a special emphasis on the formulation of hypotheses, their operationalization, the collection and use of observations, and the reconsideration of those hypotheses in light of evidence. In Section

8.3, I introduce this scheme and apply it to an empirical example to highlight the differences and similarities between frequentism and Bayesianism. The two subsequent sections, which are specifically concerned with the dimension of certainty and (moderately) formalized Bayesian causal inference, use Bayes' theorem and Bayes factor for causal inference.

8.1 Frequentism, Bayesianism, and hypothesis testing: a review

After having collected a broad range of evidence and, probably, invoked counterfactuals, the foundation is laid for the generation of frequentist and Bayesian causal inferences. For some time, the debate about causal inference has been primarily concerned with the cross-case level under the rubric of a many-causes-few-cases problem (see Chapter 4). On the other hand, the recent literature on causal inference in case studies is deeply concerned with mechanisms and process tracing.[1] Beyond the arguments made in Chapters 6 and 7, the debate focuses on two interrelated aspects. The first one builds on the distinction between a frequentist and a Bayesian logic of causal inference. The second line draws on the difference between the *uniqueness* of a prediction derived from a hypothesis and its *certainty*. In the present section, I review both arguments in the order in which I just presented them because the second point partially draws on the first.

Frequentism v. Bayesianism

In the debate about causal inference in small-n research, two counterarguments have been made against the charge of being liable to the generation of indeterminate conclusions. The first counterclaim, which has been critically discussed in detail in Section 7.2, is that process tracing and the collection of multiple causal process observations (CPOs) mitigate this problem. The second refutation, which has been reserved for this chapter, is that the critics of case studies take a frequentist perspective on causal inference (Bennett 2008, 708; George and Bennett 2005, chap. 1; McKeown 1999; Rogowski 1995). Frequentism means that the quality of causal inferences hinges on the *number* of observable implications and observations that receive empirical confirmation. A frequentist perspective, it is argued, is deficient because it ignores the conditional likelihood of gathering an observation and its bearing on our confidence in a hypothesis (see Chapter 3 and Bennett 2006, 341). A Bayesian perspective on causal inference is considered to be more appropriate because it emphasizes the probative value of a CPO. An isolated piece of within-case evidence that discriminates between competing explanations is said to boost confidence in one hypothesis and strongly discredit its rivals (Beach and Pedersen, 2012; Bennett 2008, 711; Van Evera 1997, 30–2). A single, unexpected CPO that can be linked to only one hypothesis can therefore be more valuable than a large number of CPOs that offered few surprises to collect in the first place.

Four types of tests

The emphasis of a Bayesian logic of causal inference is linked with another central argument that builds on a two-by-two typology of hypotheses tests (Bennett 2010; Collier 2010; Van Evera 1997, 30–2). The first dimension is called *certitude* or *certainty*, which is equivalent to the conditional likelihood of finding a proposition confirmed (see Chapter 3). The more likely it is to gather a specific observation in the selected case, the higher is the certainty of the prediction. The second dimension is labeled *uniqueness* and captures whether a prediction is made by one or multiple hypotheses.

To exemplify the meaning of the dimensions, consider two well-known hypotheses from the field of welfare state research. The compensation hypothesis predicts that spending increases with an increasing degree of economic openness. On the other hand, the efficiency hypothesis claims that spending decreases with increasing openness (Obinger et al. 2010, chap. 1). It follows that both hypotheses make unique predictions about the way in which globalization affects spending. The constellation is different when one tests the compensation hypothesis against the proposition that left-wing governments are more frequently associated with higher levels of spending than are right-wing governments. If one were to compare two cases, one with a closed economy and a right-wing cabinet and one with an open economy and a left-wing government, the compensation hypothesis and government ideology hypothesis lack uniqueness because both predict low spending in the first case and high spending in the second.

If one treats the criteria of certainty and uniqueness as dichotomous, one obtains the four types of tests that are presented in Table 8.1.[2] Each test has different implications for the hypothesis under scrutiny, depending on whether it passes or fails the test. In relation to this point, the typology was recently amended by specifying whether a criterion is necessary or sufficient for inferring causation after a test has been successful (Bennett 2010, 210–11). Certainty is related to necessary criteria for causal inference, while uniqueness is tied to sufficient criteria. High certainty is argued to be necessary, while low certainty is nonnecessary for inferring causation. Correspondingly, high uniqueness is sufficient and low uniqueness insufficient for causal inference. The four types of tests and the criteria of necessity and sufficiency are now clarified step by step.

The weakest of all tests is the *straw-in-the-wind test* because it is marked by low uniqueness and low certainty. A passed test is not sufficient for inferring causation because at least one other proposition is confirmed as well because of low uniqueness. Similarly, a successful test is not necessary for inferring that the hypothesis could be correct because certainty was low and failure to be expected. If a proposition fails a *hoop test*, the information gained is of particular value because it provides an unexpected outcome

Table 8.1 Types of hypothesis tests and causal inference

		Certainty (necessary for inferring causation)	
		High (Yes)	*Low (No)*
Uniqueness (sufficient for inferring causation)	*High (Yes)*	Doubly decisive test	Smoking gun test
	Low (No)	Hoop test	Straw-in-the-wind test

equivalent to a failed most-likely test. Since it is necessary for a hypothesis to master a test with high certainty, failure allows it to infer with a high degree of confidence that the hypothesis is incorrect. On the other hand, passing the test is not sufficient for inferring causation because, again, at least one other proposition receives empirical support as well.

Smoking gun tests have high inferential value if the hypothesis can be confirmed because this is equivalent to a passed least-likely test. Moreover, the discriminatory power of this test is high because of its high uniqueness, which is equivalent to a sufficient criterion for causal inference. On the other hand, a negative test constitutes a failed least-likely test and offers few interesting insights, which is the reason that passing a smoking gun test is not necessary for inferring causation. Tests of the *doubly decisive type* are deemed to be the most powerful ones (Bennett 2010, 210–11; Van Evera 1997, 31–2). They combine the beneficial features of a hoop test and a smoking gun test and have relatively strong theoretical implications in the face of a failed and passed most-likely test alike. For this reason, passing a doubly decisive test is necessary and sufficient for inferring that a hypothesis is correct.

The four types of tests are mostly discussed in the context of process tracing and hypotheses tests (Bennett 2008, 2010; Collier 2010). This is a legitimate level of analysis for the tests but unnecessarily limits the value of the typology. There is nothing that speaks against its extension to the cross-case level because one can equally well determine the certainty and uniqueness of cross-case and within-case hypotheses.[3] For this reason, the subsequent discussion of frequentist and Bayesian causal inference is not specifically limited to process tracing and the within-case level.

In the next sections, I take an integrated perspective on causal inference and reconsider the two major lines of reasoning reviewed in this section. First, I deal with the implications of the confirmation of a unique prediction and then turn to the criterion of certainty in the context of a more general discussion of Bayesian causal inference.

8.2 Uniqueness v. contradiction

Supposedly, a doubly decisive test has the highest inferential leverage of all four types. Failure disconfirms a hypothesis, whereas a successful test confirms this proposition and disconfirms all others (Bennett 2008, 711; Van Evera 1997, 30–2). In this section, it is maintained that this holds true only if two hypotheses make exactly the opposite prediction, which is fundamentally different from making two unique but complementary predictions. For instance, when two hypotheses center on the same outcome, contradictory predictions imply that the outcome of interest is attributable to *monocausation* because there is only one case related to the outcome (for a more detailed discussion, Rohlfing 2012). When two hypotheses make unique predictions on Y that are not contradictory, the could be characterized by *equifinality* in most fields of the social sciences, this is a much more defendable assumption than monocausation (Bennett and Elman 2006, 457; Mahoney 2007b, 135).

This line of reasoning can best be illustrated with an example. Imagine that you are interested in the determinants of high welfare state spending and test the compensation hypothesis against a left-wing government hypothesis.[4] The former predicts that high openness produces high spending because the government aims to compensate the losers from globalization, whereas the latter stipulates that left-wing governments use ideological reasons to account for high spending. You choose a country that is a most-likely case for both propositions, that is, one that has an open economy and a left-wing government and is characterized by other factors rendering the outcome most likely to occur.

For purposes of illustration and without loss of generality, I focus on only one observable within-case implication of each hypothesis in the following. The implication of the compensation hypothesis is that members of the cabinet justify the high level of spending with the need to do something for those constituents who suffer as a result of globalization. In respect to the left-wing government hypothesis, you expect statements from the members of the cabinet that emphasize the benefits of big government on the basis of ideological principles. Both predictions are unique in that the selected country constitutes a doubly decisive test for both hypotheses owing to high uniqueness and high certainty. In the empirical analysis, you begin with the collection of multiple, unambiguous CPOs from internal documents, secondary sources, interviews, and newspapers, all of which turn out to be in line with the compensation hypothesis.

What can we now infer from the collected CPOs? Evidently, it is justified to conclude that the compensation hypothesis is found confirmed. In light of the current interpretation of the doubly decisive test, it is possible to infer something about the left-wing government hypothesis as well. The doubly decisive test implies that one observation is doubly decisive because it confirms one hypothesis while, at the same time, it disconfirms another. As regards the example, the confirmation of the compensation hypothesis allows one to disconfirm the left-wing proposition *without* having analyzed it empirically. Certainly, there are good reasons that one might be reluctant to drop a hypothesis without an empirical analysis. However, it should be noted that this is a direct consequence of the claim that one piece of evidence simultaneously entails implications for two hypotheses.

Reluctance to dismiss a hypothesis on the basis of evidence tied to another hypothesis is well founded because the confirmation of a unique prediction is not automatically the same as disconfirming competing propositions. This argument can be illustrated with the empirical example at hand. The hypothesis that economic openness produces high spending as a means of compensating employees does not deny that other conditions, such as a left-wing government, bring about the same outcome for different reasons. The compensation hypothesis makes a specific prediction about why spending increases in response to globalization and, at the same time, allows for the presence of other causes, that is, equifinality is possible.

At this point, it is useful to return to the exposition of causation in Chapter 2. The compensation hypothesis assigns the government's aim to compensate employees the status of a sufficient condition. Sufficiency hypotheses are concerned only with the consequences of the hypothesized condition and do not rule out that other conditions or conjunctions can produce the same outcome. If the compensation hypothesis were to make an exclusive causal claim, it would have to read 'countries have high spending *if and only if* they aim to compensate employees for the pressure from globalization'. This proposition would claim that compensatory aims are a necessary and sufficient condition for high spending, which means assuming monocausation.[5] The compensation hypothesis and, in fact, most hypotheses do not make exclusive causal claims. One therefore does harm a hypothesis if one takes confirming evidence for this proposition as contradictory evidence for others (see Rohlfing 2012).

The fallacy of the doubly decisive test becomes further apparent when taking into account the left-wing government hypothesis. In the empirical example, I assumed that one first collected evidence for the compensation hypothesis, with the consequence that the left-wing government proposition was rejected without having been examined at all. Imagine now that you first test the government hypothesis and that all collected causal process observations clearly confirm it. In the conventional reading of the doubly decisive test, a successful test now prompts one to reject the compensation hypothesis. Evidently, the conclusions that one draws should not depend on the order in which one examines hypotheses, thereby demonstrating that the doubly decisive is not doubly decisive. In the first place, the confirmation of a unique prediction is neither more nor less than a confirmation of a unique prediction and thus does not entail consequences for other hypotheses.

While the doubly decisive test does not actually capture what it is purported to capture, the idea of disconfirming one hypothesis by confirming another is appealing and, in fact, can be incorporated in the typology presented above. In order to do so, it is necessary to add a third dimension called *'contradiction'* or 'mutual exclusiveness' to the typology. The dimension contradiction captures whether the implication of one hypothesis is directly contradicted by another proposition. A straightforward cross-case example for a contradiction can be found in the efficiency and compensation hypotheses because they predict exactly the opposite development of welfare state spending as a consequence of globalization. According to the efficiency hypothesis, spending should decrease, whereas the compensation hypothesis predicts that it increases.[6] In this instance, finding confirmatory evidence for one hypothesis automatically refutes the other because only one can be accurate. This differs from the previous example, in which the compensation hypothesis and left-wing hypothesis predicted the same outcome but for different, nonexclusive reasons.

If one expands the typology by adding the dimension contradiction, one obtains a typology with three dimensions, each of which is binary – certainty is high or low, uniqueness is high or low, and contradictions are either present or absent. There are various ways to rank the eight types of tests that result from the intersection of the three dimensions, depending on the weight that one attaches to each of the three criteria. I refrain from discussing possible rankings here and leave it to the reader to identify the type of test that fits the respective research interest most.

In concluding the discussion of uniqueness, one may wonder whether it is really necessary to make the neat two-by-two typology more complex by doubling the number of possible types. The previous discussion should have demonstrated that it is highly recommended. One could leave the typology as it stands and interpret the criterion of uniqueness properly, namely that evidence in line with a unique implication confirms one and only one hypothesis. Yet it was demonstrated that unique implications cover two subtypes: unique and contradicted implications and unique and noncontradicted ones. Since contradictions have important consequences for causal inference, they are better brought to the forefront by creating a dimension of their own and increasing the number of possible tests to eight.

8.3 Frequentism and Bayesianism: friends or foes?

The discussion of Bayesian causal inference requires an initial consideration of the relation between frequentism and Bayesianism.[7] The common ground for this is provided by Figure 8.1. It provides a stylized presentation of the research process differing from that in Chapter 1. Figure 8.1

emphasizes those aspects of a case study that are more closely related to causal inference and notes for each step of the process whether the relevant criterion is frequentist or Bayesian.

The process starts with the formulation of a working hypothesis, that is, a hypothesis one wants to examine empirically. A hypothesis (represented by H in the following) can cover predictions on correlational and set-relational cross-case patterns and/or on causal mechanisms and causal processes that one expects to observe in a single case study or in a comparative case study (see Western 2001). Altogether, this means that Bayesian causal inference does not face any constraint in qualitative case studies.

A Bayesian perspective on case studies additionally mandates it to determine the *prior probability* that the hypothesis is true (step 1.a in Figure 8.1). The probability is abbreviated $p(H)$ and is in the following also referred to as the *ex ante* likelihood or simply the *prior*. The likelihood must be derived from the existing state of theory and empirical research, which is not always an easy endeavor. As regards the democratic peace hypothesis, for example, the specification of the prior is not so difficult because one can be rather confident that two democracies do not fight each other. The corresponding *ex ante* likelihood thus can be fixed at, say, 0.95. The probability that the working hypothesis is wrong is captured by the null hypothesis. It is referred to as $\sim H$ (~ denoting 'not') and is linked to a probability of $p(\sim H)$, which simply is $1–p(H)$.

In the second step, one specifies as many observable implications as possible that should be found confirmed if the working hypothesis is true (George and Bennett 2005, chap. 10; King et al. 1994, chap. 1; Lave and March 1975). This criterion is good practice in empirical research and a frequentist one because the more observable implications one can specify, the higher the leverage of the hypothesis is and the easier it is to find disconfirming evidence. In general, an observable implication can refer to the predicted cross-case score of the outcome or cause of interest or to auxiliary implications that are located on either the cross-case or the within-case level (see below and Mahoney 2010).

The third step involves case selection, which is distribution based in frequentist case studies and theory based in Bayesian research. As regards the latter, I explained in Chapter 3 that the choice of a case influences the conditional likelihood of finding evidence (abbreviated E) that confirms an observable implication. This element was formally captured by $p(E|H\ \&\ case)$, where 'case' emphasizes the salience of the selected case for Bayesian causal inference and represents case-specific features that are relevant against the backdrop of the working hypothesis. The evidence E does not refer to any kind of empirical insight, but specifically means evidence that confirms the working hypothesis. As a consequence, it is also possible to gather evidence that is in discord with the working hypothesis, denoted by $\sim E$ in the following. Step four comprises the collection of as many CPOs as possible. The rationale for this frequentist criterion is that causal inferences are the more compelling the larger the number of CPOs on which they rest (Thies 2002).

From stage five onward, one is working backward from the gathered evidence to the hypothesis of interest. In step five, one subsumes the individual CPOs under the previously specified observable implications or, if this is not possible for all CPOs, relates them to constructs one has not considered prior to the empirical analysis (Adcock and Collier 2001). Again, the criterion for this step is frequentist, as it is more compelling to argue that an observable implication is found confirmed the more CPOs it subsumes. The sixth stage captures the generation of the inference that the hypothesis has empirical resonance or not.[8] It follows from the previous step that the guideline here is frequentist; the more observable implications that are found confirmed and the more CPOs previously subsumed under each implication, the more solid the inference.

Figure 8.1 shows that frequentist case studies are complete after the sixth step. On the basis of the evidence, one infers whether the hypothesis is true or not and concludes the case study. In contrast, Bayesian case studies call for two additional steps. In step seven, one calculates the *posterior probability* that the hypothesis is true, conditional on the collected empirical evidence that was previously subsumed under observable implications. This is also

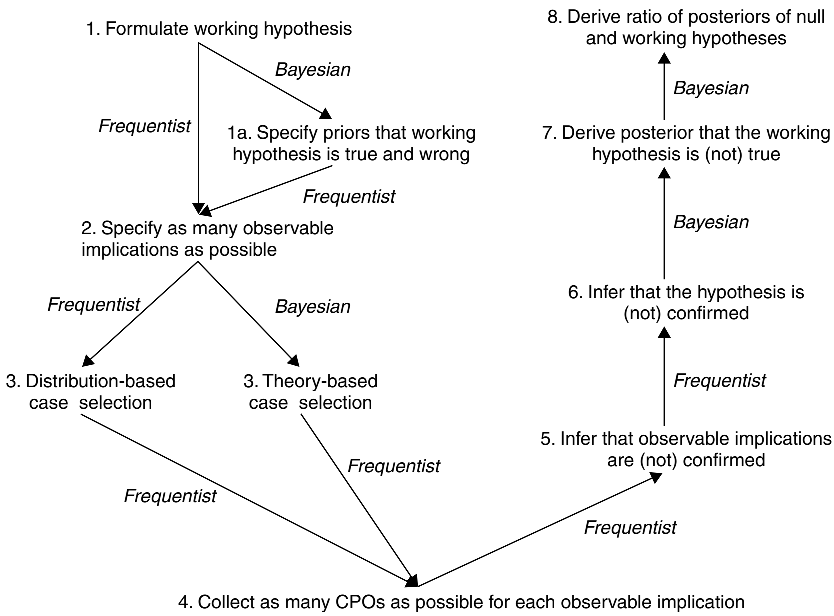

Figure 8.1 The research process and frequentist and Bayesian causal inference

called the *posterior* or *ex post* likelihood, formally denoted by $p(H|E)$ and calculated with Bayes' theorem in ways elaborated below in the context of an example. The formal representation of the posterior, which is standard Bayesian notation, directly relates the evidence to the Bayesian updating of confidence in a hypothesis. However, steps five and six show that it is not possible to move directly from the collection of observations to Bayesian causal inference. Observations do not speak for themselves (Bartelborth 2004) but make sense only when they are tied to observable implications, covered in step five in Figure 8.1. Observable implications are in turn closely related to a hypothesis that is evaluated in step six in light of what was produced in the fifth step. Only after this has been done is it possible to update confidence in the working hypothesis by determining its posterior likelihood. One can therefore focus on the posterior $p(H|E)$ only after establishing that the collected observations confirm or do not confirm an observable implication. In looking at a piece of evidence, the question one needs to answer is if this observation is, formally seen, of type E or $\sim E$. If this is not done prior to the actual generation of causal inferences, we simply would not know whether the relevant posterior is $p(H|E)$ or $p(H|\sim E)$.

Finally, step eight concerns the calculation of the posterior probability that the null hypothesis is true *relative* to the posterior probability that the working hypothesis is true. This final step is important because, as is illustrated below, the mere change of confidence as it is expressed in the difference between the prior and the posterior might give misleading impressions about how much trust we should, in fact, have in the hypothesis.

With respect to the topic of this section, the relation between frequentism and Bayesianism, the previous discussion yields two insights. First, it is possible to run a self-contained frequentist case study void of any elements of Bayesianism allowing one to communicate levels of confidence in a hypothesis. Second, Bayesian case studies *do* involve elements of frequentism. A good Bayesian case study takes full recognition of its frequentist elements in order to strengthen causal inference and to avoid falling prey to pitfalls that exist if one pits frequentism and Bayesianism against each other and treats them as mutually exclusive. In this sense, the Bayesian and frequentist logic of causal inference are more friends than foes. In order to highlight potentials pitfalls and flesh out the scheme in Figure 8.1, I will now use the compensation hypothesis of welfare state research as an example in the context of a general exposition of Bayesian causal inference.

8.4 Bayes' theorem: from prior to posterior confidence

Bayesian case studies: an example

The compensation hypothesis predicts that the government opts for a high level of spending in order to compensate those who have suffered losses owing to globalization and to reassure those employees that feel economically insecure (Walter 2010).[9] Political actors opt for high welfare spending in order to secure the political support of workers and labor unions and other interest groups as they will benefit from a large welfare state. In the following, the compensation hypothesis is specifically referred to as *C*. In this example, I assume that the state of theory justifies the attachment of a prior probability – *p(C)* – of 0.5 to the compensation hypothesis. This automatically implies that the null hypothesis – *p(~C)* – has a probability of 0.5 as well. Substantively, this means that one is uncertain about the veracity of the compensation hypothesis because one considers it as likely to be true as it is to be untrue.

In the second step, one derives as many observable implications as possible that lend credence to the claim that compensatory aims drive high spending. Since this example is illustrative, I limit this step to the specification of two observable implications. The first implication is that representatives of the government meet regularly with labor unions as the unions represent the employees that are harmed or fear harm from globalization. The labor unions can signal to the government what their members expect from the government, while the government can in turn communicate vis-à-vis the labor unions what it intends to do and try to gain union support for its policy-making activities. The second observable implication is related to the government's rhetoric. If the government perceives that economic openness contributes to actual and perceived economic insecurity, it can try to gain credit among the electorate by making clear that it cares about the concerns of employees and intends to guarantee a high level of welfare provision.

In step three, one selects a least-likely case for the compensation hypothesis. With respect to the implications, a least-likely case would be a country with a right-wing government. Right-wing governments, it can be argued, are least likely to cooperate with labor unions and least likely to adopt a rhetoric that endorses high welfare state spending. The choice of a least-likely case makes it possible to specify the conditional prior likelihood of finding both observable implications confirmed, given the theory-related features of the selected case and the prior probability that the compensation hypothesis is true (see Chapter 3). For purposes of presentation, I simply assume that the conditional likelihood *p(E|C & case)* can be set at 0.3. This implies that the likelihood *p(~E|C & case)* is 0.7.

The empirical analysis in step four then serves to collect CPOs for each of the observable implications. As regards the expectation that the government meets with labor unions, possible CPOs are newspaper articles reporting that meetings took place; protocols from multiple meetings demonstrating that the government and the unions talked about the proper response to globalization pressures; protocols revealing that the government would appreciate the support of labor unions at the next election; and so on. Concerning the government's rhetoric, CPOs consist of the analysis of multiple statements addressed to different audiences. If one would only examine a campaign speech of a party's leading candidate, one could dismiss the rhetoric as strategic and lip service. However, the claim that the candidate is indeed in favor of high spending in order to gather support of workers and labor unions gains credibility if he or she makes similar statements in closed-door meetings with other members of the party elite, vis-à-vis party delegates at a convention, and so on.

After having collected the CPOs, step five requires one to assign each observation to one of the two observable implications. The more CPOs belong to one of the implications, the easier it is to claim that one has gathered predicted evidence of the type *E* and that the implication has empirical resonance. This argument runs counter to the assertion that a single observation may suffice to strongly confirm a hypothesis (Beach and Pedersen, 2012; Bennett 2008, 711; 2010, 209). The rationale for the assertion that more CPOs are better than fewer is twofold.

The first is epistemological because reliance on just one or a few CPOs ignores the source coverage problem. For reasons detailed in Section 7.1, it is not appropriate to base causal inferences on a single causal process observation regardless of how theoretically improbable it is to make. Assume that you interview a leading conservative politician from a party that is not known for caring about those who incur loss from globalization. He gives you a statement: '*We keep welfare state spending at high levels because we want to compensate the losers from globalization and hope for their political support in the future*'. Would you discontinue the empirical analysis after one interview because this statement fully confirms your hypothesis? You should not because the politician may have cause to conceal the true reasons for high welfare state spending (Bennett 2010, 219), which is a threat to causal inference unrelated to the status of a case as a least-likely or most-likely case.

Second, the case study literature usually casts the importance of the probative value of CPOs by comparing one or a few unexpected CPOs with a large number of observations having a high certainty (Bennett 2008). If the contrast is like this (ignoring the source coverage problem for a moment), one can contend that the unlikely observations are more valuable. However, the picture changes when one compares a case study that is built on a small number of unexpected CPOs with an analysis that draws on the same CPOs *plus* additional ones. Regardless of the degree to which one expects to collect these observations, the second case study is more convincing than the first one because the empirical basis is more solid. For these reasons, one should not play the certainty of an observation off the number of predictions and should always try to find additional CPOs that all lend credence to the same prediction.[10]

After having subsumed the collected CPOs under observable implications, step six involves their consideration in total so that one can infer whether the working hypothesis is accurate or not. For the empirical example, this involves the determination of whether both observable implications have empirical resonance, highlighting that the relevant criterion is one of frequentism. The specification and confirmation of two predictions makes it easier to conclude that the compensation hypothesis is correct than does the analysis of only one of the two. While a purely frequentist case study would stop here, Bayesian research requires two additional steps. Because of their centrality for Bayesian causal inference, they are addressed in turn in two separate sections. The steps include the calculation of the posterior probability that the hypothesis is correct and the assessment of this probability in light of the posterior probability that the null hypothesis is true.

From prior to posterior confidence: invoking Bayes' theorem

For a demonstration of Bayes' theorem with the empirical example, assume one infers that the evidence supports the compensation hypothesis and that the relevant posterior is $p(H|E)$. This means we are interested in the posterior probability that the compensation hypothesis is true on the basis of the gathered evidence. The question now is how can we use the information we have to determine the *ex post* confidence in the hypothesis? The answer to this question lies in Bayes' theorem, which requires only three probabilities as inputs, two of which have already been introduced above.[11] The first input is the prior probability $p(C)$ that the compensation hypothesis is true. The second ingredient is the conditional likelihood $p(E|H \,\&\, case)$ of finding specific observable implications conditional on $p(C)$ and the theory-specific features of the selected case. Third, one also needs an estimate for the conditional likelihood of gathering observable implications if the working hypothesis is wrong, captured by $p(E|{\sim}C \,\&\, case)$. The latter requirement is interesting because Bayesian case studies force one to generate expectations about the likelihood of collecting specific observations if the working hypothesis is incorrect (Howson and Urbach 2005, 20–1). These ingredients permit the calculation of the posterior likelihood $p(C|E)$ by the following formula:[12]

$$p(C|E) = \frac{p(E|C\&case) * p(C)}{p(E|C\&case) * p(C) + p(E|{\sim}C\&case) * p({\sim}C)} \tag{8.1}$$

Complicated as the theorem may seem at first glance, it is straightforward when presented in the form of 2×2 tables. Table 8.2 tabulates the compensation hypothesis and null hypothesis – captured by the two columns – against the possibility of gathering and not gathering supportive evidence – included by the two rows. The level of prior confidence in the compensation hypothesis is the marginal probability of the left column. The likelihood that the proposition is true and not true must add up to 1,

Table 8.2 Prior probabilities for empirical example

		Compensation hypothesis	
		True (*C*)	*Not true* (*~C*)
Collected evidence	*Supportive* (*E*)	0.3 p(E|C & case)	0.2 p(E|~C & case)
	Not supportive (*~E*)	0.7 p(~E|C & case)	0.8 p(~E|~C & case)
		0.5 p(C)	0.5 p(~C)

meaning that a prior *p(C)* of 0.5 implies a prior of 0.5 for the null hypothesis – *p(~C)* – as well. In Table 8.2, this likelihood is the marginal probability of the right column.

The upper-left cell denotes the probability of confirming the observable implications if the compensation hypothesis is true and given the selected case. The probability *p(E|C & case)* is set to 0.3 for this example (see above). Consequently, the likelihood *p(~E|C & case)* of not gathering the expected observations when the hypothesis is correct is 0.7, which is captured by the lower-left cell.[13]

The probability of finding the observable implications confirmed if the null hypothesis is true is captured by the upper-right cell. This likelihood is fixed at 0.2 for the empirical example. The reason that *p(E|~C & case)* is smaller than *p(E|C & case)* but larger than 0 is as follows. The probability of finding the expected implications confirmed if the compensation hypothesis is wrong is lower because political actors do not respond to demands for higher spending by globalization losers. A case with a conservative government in place has a small likelihood of supporting the observable implications because globalization losers do not tend to be among the constituency of conservative parties. Still, it holds that, when the compensation hypothesis is true, a conservative government should be more likely to conform to the theoretical expectations when it cares about globalization losers, compared with a situation in which it is indifferent to them (that is, the compensation hypothesis is wrong). At the same time, the probability *p(E|~C & case)* is not equal to zero because it is not entirely impossible that a conservative government uses welfare state policy for the compensation of globalization losers in some cases. For these two reasons, in this example, the likelihood of gathering expected evidence, even if the compensation hypothesis is wrong, lies between the other conditional probability and zero.

Table 8.3 Posterior probabilities for empirical example

		Compensation hypothesis	
		True *(C)*	*Not true* *(~C)*
Collected evidence	*Supportive* *(E)*	0.6 p(C|E)	0.4 p(~C|E)
	Not supportive *(~E)*	0.47 p(C|~E)	0.53 p(~C|~E)
		0.5 p(C)	0.5 p(~C)

The formula for Bayes' theorem now permits us to estimate the posterior probability that the compensation hypothesis is true, given that there is supportive evidence for this proposition. Plugging the values from Table 8.1 into the formula produces the posterior likelihoods in the upper row of Table 8.2. The posterior probability *p(C|E)* that the compensation hypothesis is true is 0.6. In other words, the new empirical evidence leads us to have a confidence of 60 percent that the compensation hypothesis is true. On the other hand, we still attach a probability of 40 percent to the null hypothesis that the compensation hypothesis is wrong. Although we assume that the evidence is favorable and that the posteriors in the upper row matter, for reasons of comprehensiveness, the lower row of Table 8.3 also presents the posteriors for a case study where the working hypothesis was not confirmed by the evidence. If the collected CPOs had not supported the compensation hypothesis, the corresponding posteriors *p(C|~E)* and *p(~C|~E)* would have been 0.47 and 0.53, respectively.

The illustrative application of Bayes' theorem produces an interesting insight because the posterior probability that the compensation is true is only 0.6, that is, ten percentage points higher than the prior. This point is important to emphasize because the case study is a least-likely test with a conditional likelihood of just 0.3. According to our intuition and common arguments about least-likely tests, mastering a difficult empirical test should leave us with a higher posterior than 0.6. This finding casts doubt on the advice to choose least-likely cases for process tracing, as passed least-likely tests, it is argued, yield substantial inferential leverage (Levy 2008, 12). The modest increase in our confidence can be understood by taking a closer look at how Bayes' theorem and Bayesian causal inference works, which, in particular, demands a discussion of the role played by the conditional likelihood *p(E|~C & case)* in the next section.

Priors, posteriors, and the likelihood ratio

The small increase in confidence after a passed least-likely test can be clarified by a reconsideration of Bayes' formula and the logic of Bayesian causal inference. The denominator includes probabilities related to the working hypothesis and the null hypothesis, suggesting that the key to the puzzle lies in the conditional likelihood *p(E|~C & case)*. In the empirical example, the probability is set at 0.2 and is thus only slightly smaller than the corresponding probability of the compensation hypothesis. Theoretically, if the compensation hypothesis is true, it would be deemed more likely, but not by much, that one will find supportive evidence than if it is false. Now that process tracing yields CPOs that are in line with the observable implications, the question is what to infer about the compensation hypothesis *and* the null hypothesis in light of the collected evidence.

Given the parameters of the empirical example, it is natural that we would have more confidence in the hypothesis that was more likely to hold true *ex ante* and less trust in the proposition that was deemed to be less likely to be valid prior to the empirical analysis. However, there is no reason to be overly confident in the compensation hypothesis after the empirical analysis because it was only somewhat more likely to be true *ex ante* compared with the null hypothesis. It is then logical to attach a higher posterior confidence to the compensation hypothesis than the null hypothesis but to keep the difference between the posteriors within limits.

This informal reasoning about the role of the conditional likelihoods *p(E|C & case)* and *p(E|~C & case)* can be underpinned with a formal treatment of what is known as the likelihood ratio. The central idea behind the application of Bayes' theorem is to know how likely it is that the working and the null hypothesis are correct. In the example, both probabilities add up to one because they are conditioned on the inference that one has gathered supporting evidence *E*. It is thus possible to build the ratio of the posteriors of the null and of the working hypothesis in order to determine how likely the former is correct *relative* to the latter (Abell 2009a). When it is feasible to build the ratio of posteriors, it is equally possible to bring the two formulas for the respective posteriors into relation. Formally, it then holds:[14]

$$\frac{p(\sim C|E)}{p(C|E)} = \frac{p(E|\sim C\&case)}{p(E|C\&case)} * \frac{p(\sim C)}{p(C)} \tag{8.2}$$

The ratio of the prior conditional likelihoods – *p(E|~C & case)/p(E|C & case)* – is known as the likelihood ratio.[15] The likelihood ratio expresses the change in the priors that is brought about by the confirming empirical evidence. If the priors are equal – that is, p(C) = p(~C) = 0.5 – our posterior confidence in the null and working hypotheses is therefore driven only by the likelihood ratio.[16] The importance of the likelihood ratio is readily illustrated with the empirical example. In the above empirical example, *p(E|C & case)* and *p(E|~C & case)* equal 0.3 and 0.2, respectively, and produce a likelihood ratio of about 0.67. This factor tells us *in advance* of the empirical analysis that the relative posterior confidence in the null and working hypotheses will be 0.67, as well, when we can find confirming evidence. In fact, the posterior is 0.4 for the null hypothesis and 0.66 for the working proposition, which yields a ratio of 0.67.

The illustration of the likelihood ratio points to one instrument with which one can influence the inferential leverage of a Bayesian case study. This instrument is related to case selection and qualifies the intuition and current recommendations that govern the choice of most-likely and least-likely case in theory-based case selection (see above and Chapter 3). The previous discussion shows that the larger the conditional probability of the working hypothesis *relative* to the null hypothesis, the smaller the likelihood ratio is and the greater the confidence in the working hypothesis following a successful test.

A least-likely case for the working hypothesis can meet this criterion but only if the conditional likelihood of the null hypothesis is much smaller. Similarly, a most-likely case for the working hypothesis offers considerable inferential leverage only when the conditional probability of the null proposition is smaller. Thus, the general guideline for Bayesian case studies and theory-based case selection is to *maximize* the difference between the conditional likelihood of the working proposition and the null hypothesis. A case study for which the conditional likelihoods are 0.9 and 0.5 entails more leverage than an analysis with probabilities of 0.3 and 0.1, respectively, though the latter case study would be preferred if one were simply looking for a least-likely case.

The discussion of the likelihood ratio has two more implications for causal inference. First, there is no rationale for doing a case study in which the prior conditional likelihoods are both 0.5. The likelihood ratio is 1 in this instance, and the ratio of the posteriors is equal to the ratio of priors. Such a case study therefore does not change anything about our confidence in the hypotheses. Second, the importance of the null hypothesis puts into the perspective the claim that high certainty is sufficient for causal inference and that low certainty is insufficient (see above and Bennett 2010). This perspective is misleading because the conditional likelihood of the null hypothesis must be considered as well. Both high and low certainty can be sufficient for affirmative causal inference provided that the null hypothesis is much less likely to be found confirmed given the selected case.

Having addressed the role of Bayes' theorem and the likelihood ratio in determining the *ex post* confidence in hypotheses, the next question is: what to do with the posteriors of the working hypothesis and the null? Should we accept the working hypothesis as correct when the posterior is larger than the prior? Or should we consider a hypothesis to be correct only if the posterior reaches a certain threshold? In providing answers to these questions, the following section shows that it is again necessary to take an integrated perspective on the working proposition and the null.

8.5 Bayes factor: relative posterior confidence in hypotheses

In a Bayesian framework, the *ex post* assessment of the working hypothesis can be based on the ratio between the posterior likelihood that the null hypothesis is true and the posterior likelihood that the working hypothesis is true. This is known as *Bayes factor* and applied to the empirical example, the ratio reads:

$$\frac{p(\sim C|E)}{p(C|E)} \tag{8.3}$$

As explained previously in the discussion of the likelihood ratio, the ratio for the empirical example is 0.67 (0.4/0.6). This means that, given the parameters of the case study and the collected evidence, our *ex post* confidence in the working hypothesis is 1.5 (1/0.67) times greater than our trust in the null hypothesis. As is the case with many numeric criteria, there are no unequivocal guidelines for the interpretation of such figures and Bayes factor more generally. The literature on Bayesian causal inference developed conventions about how to interpret the ratio (Jeffreys 1961). A scheme commonly used in Bayesian research is depicted in Table 8.4.[17]

Table 8.4 Interpretation of ratio of posteriors

Range of ratio of posteriors	**Interpretation**
> 100	Extreme evidence for null hypothesis
100 – 30	Very Strong evidence for null hypothesis
30 – 10	Strong evidence for null hypothesis
10 – 3	Substantial evidence for null hypothesis
3 – 1	Anecdotal evidence for null hypothesis
1	Inconclusive evidence
1 – 1/3	Anecdotal evidence for working hypothesis
1/3 – 1/10	Substantial evidence for working hypothesis
1/10 – 1/30	Strong evidence for working hypothesis
1/30 – 1/100	Very strong evidence for working hypothesis
< 1/100	Extreme evidence for working hypothesis

The table shows that inferences are based on ordinal categories and that the smaller the ratio, the more favorable the evidence is for the working hypothesis. The Bayes factor of our empirical example, 0.67, falls into the category of 'anecdotal evidence for the working hypothesis'. Although the empirical insights favor the working proposition over the null, the scheme tells us that the difference in the posteriors is barely worth mentioning. This is different when the posterior of one hypothesis is 0.75 or larger, as the ratio then becomes larger than 3 or smaller than 1/3, depending on whether the posterior of 0.75 is attached to the null or to the working hypothesis.

On a more general level, the scheme emphasizes that information about the mere change in the confidence of a hypothesis conveys little relevant information. For example, confidence in the working hypothesis increases by 0.15 points and 150 percent if the prior is 0.1 and the posterior 0.25. Although impressive, one should not lose sight of the fact that the posterior confidence in the null is 0.75 in this hypothetical example. This means that the null is still three times more likely to be true after the empirical analysis and that the case study yields substantial evidence in favor of the null hypothesis.

A final note on comparative hypothesis testing is in order, that is, case studies that test two substantive hypotheses, as was the case with the compensation and left-wing government hypothesis in the empirical example. When Bayes' formula is invoked, as was done in this chapter, a working hypothesis is always tested against the null to determine whether the former is wrong. If two hypotheses make predictions that are not mutually exclusive, one has to apply Bayes' theorem, as it was introduced above, to each proposition separately. The efficiency hypothesis and the compensation hypothesis of welfare state research are examples of mutually exclusive hypotheses. The latter predicts that high economic openness leads to a high level of welfare expenditure, while the former predicts that globalization is a cause of low spending.[18] Only one of the hypotheses can be correct, which means that an increase in the confidence of one proposition automatically comes at the expense of a decrease in confidence in the other hypothesis. This differs when two hypotheses do not make contradictory predictions because, as explained above, it is possible that both are correct at the same time. In order to allow for both propositions to be correct in Bayesian case studies, one should evaluate them and the corresponding posteriors separately.

8.6 Conclusion

The concluding section draws together insights from this chapter and arguments about case selection, cross-case comparisons, and process tracing made in previous chapters. In a complete perspective, sound causal inference calls for the versatile juggling of multiple tasks. First, strong causal inference begins with strong theory by deriving, in the ideal analysis, multiple contradictory and unique hypotheses from a theory. This recommendation holds true regardless of whether one performs a frequentist or Bayesian case study. When one intends to invoke Bayes' theorem, additional theoretical input is required as one needs to expend some thought on the null hypothesis, too. Second, a distribution-based choice of cases has to consider what the relevant distribution is with respect to which cases qualify as typical, diverse, and deviant (depending on the research goal). Theory-based case selection should be performed with an eye on the priors and the conditional likelihoods – the likelihood ratio in particular– as these drive the conclusions that one draws in light of the collected evidence. Third, the actual empirical analysis needs to master the source coverage problem via the careful collection, evaluation, and triangulation of a broad range of sources. Fourth, reflected counterfactual reasoning is in order when the empirical evidence is overdetermined and one is unable to discriminate between at least two causal inferences.

Fifth, one has to derive causal inferences from the evidence. On the within-case level, one has to decide if one can reasonably infer that a mechanism is operative. On the cross-case level, causal inference means surmising that the observed pattern of scores conforms to a correlational pattern, probably involving an interaction effect or some sort of set relation. For frequentist case studies, causal inference stops at this point by concluding that the hypothesis either received empirical confirmation or did not. Bayesian case studies involve additional steps that have been described in detail in this chapter. One central insight of Bayesian causal inference is the role of case selection and the need to maximize the difference between the conditional likelihood of the null and the working hypothesis in particular. This point exemplifies the close relationship between the various elements of a case studies and the importance of their interplay for causal inference.

9

External Validity and Generalization: Challenges and Strategies

The final task of case study research interested in regularities involves the generalization of causal inferences from those cases that one did examine to those cases that are part of the population and have not been analyzed.[1] In qualitative case studies, one generalizes inferences about causal effects and causal mechanisms because a regularities perspective implies the assumption that both are regular (Kühn and Rohlfing 2010). The small-n literature identifies generalization as a major problem for qualitative case studies (Rueschemeyer 2003) as it lacks tools, such as significance testing, for determining the likelihood that the generated results are due to chance or to systematic cause–effect relationships.[2] The lack of such an instrument gave rise to the assertion that case study researchers have to assume deterministic cause–effect relationships in order to generalize at all (Lieberson 1991; Munck 2005). Against this backdrop, the focus of this chapter is threefold. In Section 9.1, I establish a link between the types of case studies discussed in Chapter 3 and the generalizability of causal inferences. The implications of different types of case studies for generalization were implicitly dealt with in Chapter 3. It is useful to give a more explicit discussion of this topic here in order to highlight that the nature of generalization hinges on the selected type of case.

The other two sections of this chapter focus on scope conditions and their role in the generalization stage of a case study. Section 8.2 relates the multidimensional nature of cases and population to the possibility of reshaping populations in the light of empirical evidence gathered in the case study. It is shown how populations can be extended or reduced on one constitutive dimension while leaving other dimensions untouched. Section 8.3 builds on the idea of multidimensional populations and different types of scope conditions and ties them to the strategy of *layered generalization*. It is shown that layered generalization allows for the analysis of medium or large populations in an iterative procedure while at the same time allowing one to keep the uncertainty of generalization within limits.

9.1 Types of case studies and generalization

The implications of different types of case studies for generalization were partially addressed in Chapter 3, largely by implication. This is not surprising as the rationale behind the choice of some types of cases is centered on the goal of generalizing causal inferences. This section puts to the forefront the scope of generalization that is attached to each of the types of cases introduced in Chapter 3. The discussion follows the distinction between distribution-based and theory-based case selection because the basis for generalization – distribution and theory – differs for the typical, deviant, and diverse case studies, on the one hand, and (failed) most-likely and (passed) least-likely case studies, on the other. As the two-case selection strategies are tied to frequentist and Bayesian case studies (see Chapter 7), the two modes of causal inference in hypothesis-testing case studies also imply different approaches to generalization.

In *distribution-based case selection*, the scope of generalization hinges on the location of a case in the population of interest. A *typical case* is similar to a specific group of similar cases in the population, which means that generalization is limited to those cases that are identified as similar. This is a simple endeavor for case studies that rely on categorical measurement because all of the cases in the same category are qualitatively identical. As discussed in Chapter 3, matters tend to become more complicated when at least one cause or the outcome is measured continuously. Depending on how the cases are distributed, one may be lucky and two or a manageable number of typical cases will cover the entire population. However, this is an empirical matter, and it is more likely that parts of the population are not covered by the generalization of inferences derived from a typical case study. These uncovered cases can be made subject to a subsequent typical case study by choosing cases that are typical for those parts of the population that were not included in the first typical case study.

A *diverse case study* is built on the simple assumption that one can generalize to all other cases in the population that are included in the selected diverse cases. If the independent variable is the gross domestic product (GDP) of a country and one selects the richest and poorest of all the countries in the population, the causal inferences are extended to all other cases because they are embraced by the diverse cases. This sweeping assumption may be right or wrong but is inherent to the notion of a diverse case study (Seawright and Gerring 2008, 297). If, for whatever reason, it is possible to examine only the richest country and the fifth-poorest country, the four countries with a lower GDP are not subject to the generalization of inferences.

The *deviant case study* also entails a specific scope of generalization, though one probably less apparent than that made in typical and diverse case studies. After the analysis of a deviant case, the insights are generalized to all other deviant cases because the causal homogeneity assumption implies that all anomalous cases are deviant for the same reason. Again, this assumption may prove to be incorrect, but this is not known until one performs another deviant case study. On a broader scale, the discussion of hypothesis-building comparisons in Chapter 5 shows that causal inferences are extended to deviant and typical cases alike. If one adds a formerly excluded cause to the analysis in order to remove the anomalies, this cause must be assumed to be operative, in principle, in the deviant cases and the typical cases because both types of cases belong to the same population.

The types of case studies that are tied to *theory-based case selection* follow a partially different logic of generalization. According to one line of reasoning, the likelihood of finding confirming evidence, given the selected case and the assumption that the hypothesis is correct, entails the scope of generalization in hypothesis-testing research. The claim can be best illustrated with the least-likely case study, which is also known as the Sinatra case: if the hypothesis can make it there, it can make it anywhere. This seems to be straightforward because if a hypothesis passes a least-likely test, it is just as likely to have mastered the test had any other case been selected because all of them entailed a higher probability of success. Consequently, a successful hypothesis test based on a least-likely case allows for generalization of inferences to the entire population (George and Bennett 2005, 121–2).

Though this reasoning is intuitively plausible, it is frequentist in nature and breaks with the logic of a Bayesian perspective on causal inference. In the conventional line of reasoning, a passed or failed most-likely or least-likely case study entails the generalization of the inference that the evidence was either confirming or disconfirming. This is the simple binary inference generated in frequentist causal inference, where a test is either failed or passed and where this inference is extended to other cases (see above). However, in Chapter 8, it was shown that Bayesian causal inference goes one step further by calculating the posterior likelihood that a hypothesis is true given the new empirical evidence. Generalization in Bayesian case studies must involve the posterior likelihood and not a true/false inference. In this light, generalization in Bayesian analyses calls for the extension of the posterior likelihood to all other cases in the population.

Passed most-likely and failed least-likely tests that form the basis of hypothesis-modifying case studies do not include a posterior likelihood because of their exploratory nature. As with most-likely and least-likely case studies and deviant cases, though, these types of exploratory case studies imply the generalization of inferences to the entire population of cases. The theory-related reason for a passed least-likely and failed most-likely test is the omission of a case (see Section 4.3). A formerly omitted cause discerned in process tracing is added to a theory that extends to all cases in the population, which means that case studies centered on failed most-likely and passed least-likely cases generalize to all other cases of interest. The arguments on types of case studies and generalization are summarized in Table 9.1.

Table 9.1 Types of case studies and scope of generalization

Selection strategy	Type of case study	Scope of generalization
Distribution based	Typical case	Similar cases
	Diverse case	All cases located between the diverse cases
	Deviant case	All other cases
Theory based	Most-likely case	All other cases
	Least-likely case	
	Failed most-likely case	
	Passed least-likely case	

9.2 Multidimensional populations, cases, and generalization

A case study starts with the specification of a population of cases for which a causal relationship is expected to hold and to which causal inferences are generalized after the empirical analysis. In addition to the generation of inferences, one can use empirical evidence for the reconsideration of the scope conditions that delineate the population. The multidimensional perspective on populations and cases allows one to engage in the disciplined evaluation of scope conditions on single dimensions and relax or tighten them, depending on the collected empirical evidence, while the scope conditions on the other dimensions remain untouched.

The opportunity for a reconsideration of the original population can be illustrated with an empirical example. Suppose that you are interested in an analysis of the hypothesis that strong public pressure accounts for high welfare state spending in OECD countries after World War II. The goal of the case study thus is to explain why some countries maintain high spending. The population of interest involves three dimensions: a territorial dimension (OECD member states); a temporal dimension (the period after World War II until the present); and a substantive dimension (the welfare state).

Now, suppose that you find the hypothesis confirmed but that your case study also suggests that public demand drives welfare state spending regardless of a country's wealth. As a consequence, one relaxes the scope condition 'OECD membership', which served as a proxy for wealth, and extends the population to all countries in the world. At the same time, the substantive and temporal boundary conditions remain: all cases of welfare state spending after World War II.

Another insight of a case study might be that a scope condition needs to be tightened. One may find that lobby groups are particularly well organized in the health care sector which was chosen as a policy field belonging to the welfare domain. Since lobby groups are more powerful in the health care sector than in other policy fields belonging to the domain of the welfare state, one may decide to replace the substantive scope condition 'welfare spending' with 'health care spending' or, more generally, 'spending in policy fields with well-organized stakeholders'. The tightening of the scope condition then eliminates other policy fields, such as child care policy, from the population because stakeholders usually are less well organized in this sector than in the health care sector. The consequence is that the population shrinks on the substantive dimension but has the same scope on the territorial and temporal dimension.

These examples show that empirical evidence can prompt the recasting of the population, but not necessarily on all of the dimensions. The strategy of dimension-specific changes in the population is brought to fruition by tying the empirical insights to a specific type of scope condition and relaxing or tightening that condition. In the previous examples, evidence that wealth does not matter was connected to the territorial scope condition because the wealth of a country is best located on the territorial dimension. In contrast, the insight that well-organized actors matter is subsumed under the substantive dimension because the degree of organization is best related to features of specific policies and policy field.

In addition to its value for the reshaping of populations, the multidimensional view on populations and cases can address the fundamental generalization problem of case studies in ways described in the following section.

9.3 Enhancing external validity: layered generalization

A pertinent problem for case studies is the generalization of causal inferences from the few cases that one examined to cases that one did not examine (Munck 2005, 4). There are two obvious and complementary instruments for confronting this challenge. First, one examines more cases in a given case study. Within common resource constraints, however, there are tight limits on the maximum number of cases that one can choose for an empirical analysis. This constraint undermines the value of this strategy because an increase in the number of cases from, say, two to four or five cases is beneficial only if the population is small. If it includes 25 cases or so, the share of unknown cases would be high even if one is able to study four or five of them instead of two.

Second, one can achieve a more favorable ratio of examined to nonexamined cases by reducing the size of the population via the transformation of causes into scope conditions (see Chapter 5). The transformation of causes into scope conditions addresses the criticism that qualitative case studies cannot handle probabilism and must assume deterministic (i.e., invariant) cause–effect relationships for generalization (Lieberson 1997, 364–74; 1991, 309–12). If the population comprises six cases and all five cases that one examines support the hypothesis, one might assume generalizability without knowing whether the causal relationship holds in the sixth case as well. Imagine a scenario in which only four of the five examined cases support the hypothesis. If the unstudied case fails to corroborate the hypothesis, which may or may not be the case, the ratio of confirming cases would be four out of six. Some scholars might still deem this sufficiently large, while others might consider a success rate of two-thirds too low. The analysis of small populations does not help to determine the proper ratio of successful cases (see Kühn and Rohlfing 2010). But, in principle, it facilitates the generalization of causal inferences in qualitative case studies. Moreover, it brings to the forefront that every case study has to specify the minimum ratio of confirming cases and make that benchmark transparent in order to allow for a debate about generalization.

An additional reason that speaks for initially small populations is the opportunity to gradually increase the scope of generalizations by lifting scope conditions step by step (Lieberson 1997, 376; Mahoney and Rueschemeyer 2003a; Walker and Cohen 1985, 293). If one postulates many scope conditions and finds a hypothesis confirmed, one can relax a boundary statement and examine in a follow-up case study whether the proposition also applies to those cases that are new to the population. With respect to the generalization of causal inferences, I refer to this research strategy as *layered generalization*.[3] The original population of cases is delineated by the set of scope conditions that is postulated in advance of the empirical analysis. With each boundary condition that is relaxed, one adds a layer of cases to the original population. If the empirical analysis can confirm the hypothesis for cases that belong to the new layer, one can conclude that it is justified to drop the scope condition in question, and the size of the population increases.

In order to bring the strategy of layered generalization to full fruition, it is important that the new layer includes only a small number of cases. The rationale for this guideline is the same as for the recommendation to begin with a small population. If one cannot examine all cases that belong to the new layer, a case study on a sample of cases that are taken from this layer again confronts a generalization problem. Consequently, the more cases one can analyze empirically and the smaller the number of cases that remain unstudied, the more credible it is to proclaim external validity with respect to the new layer. Under the assumption that it is justified to relax some of the original scope conditions, the repeated application of the strategy of layered generalization makes it possible to generate causal inferences for a large number of cases within reasonable limits of uncertainty. In a repeated application of the strategy of layered generalization, case studies therefore can be used for the generation of causal inferences for large populations. Of course, this requires performing multiple case studies, each of which builds on the previous ones, in order to gradually expand the population (or not, if it turns out that the original scope condition is warranted).[4] These are the costs one has to incur if one wants to get a hand on the generalization problem with case studies.

The above distinction between the dimensions that constitute cases and populations reenters the scenario here because it shows that layered generalization can be pursued in multiple, but complementary directions. When one aims to add a layer of cases to the existing population, one can do so by increasing the spatial coverage of a hypothesis, by extending the period of analysis, by applying a wider definition of institution, and so on. The distinction between types of scope conditions permits it to engage in layered generalization in a disciplined manner, for example by relaxing a spatial scope condition and leaving the temporal and the substantive scope of the population unaltered.

A simple empirical example from international political economy clarifies the idea of layered generalization and the opportunity to add a spatial, temporal, and substantive layer to a population (other types of scope conditions are set aside here without loss of generality). Suppose that you test the hypothesis that lobbying by powerful exporters can cause the domestic government to negotiate liberalizing trade agreements that benefit the exporters (Dür 2010; Pahre 2008). On the spatial dimension, the analysis covers Belgium and the Netherlands, two small countries that are most likely to benefit from liberal trade. On the temporal dimension, the case study includes two periods: between the first oil crisis and the end of the Cold War, and between the latter event and the present, as these reflect two important events in the realm of international trade. Substantively, the study initially focuses on tariff reductions.

The resulting population comprising the four cases is visualized in Figure 9.1. The spatial dimension is captured by the x-axis and includes the two countries. The substantive dimension (y-axis) refers to the policy field. The z-dimension denotes the temporal dimension and includes the two periods. Each of the four cuboids constituted by the intersection of the three dimensions represents one of the four cases of interest. Assuming that the hypothesis is found confirmed in an analysis of three cases, it seems justified to demand external validity because one generalizes the insights derived from three cases to one additional case.

The strategy of layered generalization now offers a systematic way to discern whether the hypothesis holds beyond the original scope. Figure 9.2 depicts three examples of how this tool can be put into practice. In spatial terms, one might add another country – in this case, Luxembourg – to the population (scenario 1). When the temporal and substantive scope conditions remain unaltered, one adds two more cases of tariff liberalization (or negative cases of tariff nonliberalization) to the original population of four cases. Presuming that the hypothesis receives empirical confirmation and that one can add the layer to the original population,

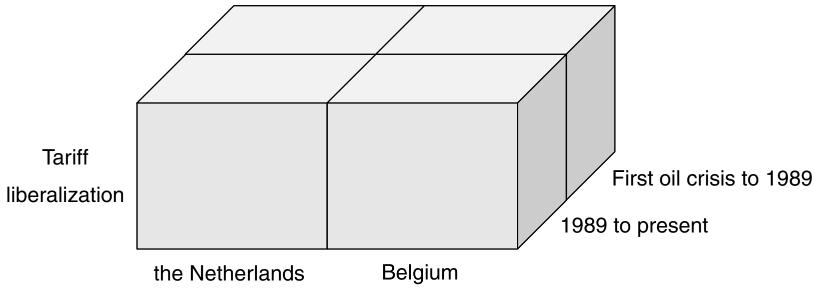

Figure 9.1 Population comprising four cases

the proposition on the effects of exporter lobbying on tariff liberalization now extends across six cases.

An alternative realization of layered generalization focuses on the same countries and substantive matter, but adds a subperiod to the period of analysis (scenario 2). One could add the period between the end of World War II and the first oil crisis so as to test whether the hypothesis applies to the entire post–World War II period, which includes subperiods with different macroeconomic environments. Again, the new layer would be of a manageable size because it would include two cases only. Finally, one could relax the substantive constraint and additionally study liberalization of import quotas. The fact that countries administer quotas and tariffs differently (Kono 2006) potentially has implications for the success of exporter lobbying and its effect on liberalization in this domain. In terms of population size, four additional cases would be included because quota liberalization is added for two periods in the Netherlands and Belgium (scenario 3).

All these examples ease one type of scope condition, probably multiple times (see below) and leave the other two untouched. If one adds one country after another to the original population and does not alter the temporal and substantive constraints, each of the new layers comprises only two cases. Repeated layered generalization gets more demanding if one expands the population in two or all three dimensions because the layers necessarily get larger and larger. Suppose that you start with tariff liberalization in Belgium and the Netherlands in two periods and first add Luxembourg to the population. In the next round of layered generalization, you further expand the population on the temporal dimension via the inclusion of the period between 1945 and the first oil crisis. The increased population now comprises three countries, meaning that the new layer includes three cases as well. Assuming that the population can be extended to all three periods for all three countries, the new population then includes nine cases: tariff liberalization of three countries is observed for three different periods. If

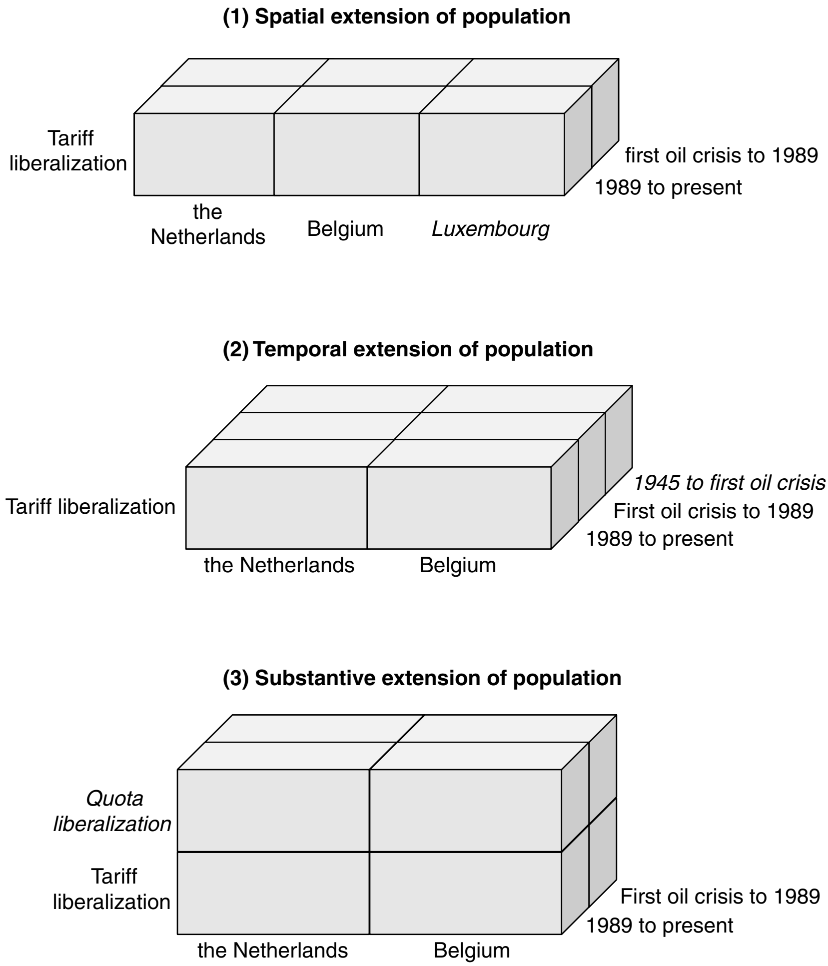

Figure 9.2 Three strategies of layered generalization

one now aims to add another layer by the inclusion of quota liberalization, this layer necessarily includes nine cases. In order to keep the generalizability problem within limits and make a compelling claim for external validity, it would be required to examine about six cases or so from this layer. It therefore needs only two rounds of layered generalization to reach the limit of the number of cases that one can usually handle in qualitative case studies.

Does this mean that the strategy of layered generalization is of limited value, that case studies are condemned to make inferences for rather small populations? No, they are not. The distinction between different types of scope conditions shows the way out of this problem. In the previous example, it was implicitly assumed that layered generalization proceeds simultaneously in all three directions. Luxembourg is added to the population for both periods, not only one. The new period added in the second step includes all three countries. Quota liberalization, when included, is done for all three countries and all three time periods. The generalizability problems that result from what can be now called *even* layered generalization can be mitigated by the strategy of *uneven layered generalization*. Uneven layered generalization means that a scope condition is relaxed for a subset of the population to which it is added with the goal of keeping the new layer of cases sufficiently small.

The rationale behind this can best be illustrated with a constellation wherein quota liberalization is introduced to a population of six cases: tariff liberalization of three countries – the Netherlands, Belgium, and Luxembourg – is examined for the period between the first oil crisis and 1989 and 1990 until present. In combination with the easing of the substantive scope condition, one can temporarily re-invoke spatial and/or temporal constraints in order to diminish the generalizability problem for the new layer of cases.[5] From a spatial point of view, one could test the expansion of the hypothesis to quota liberalization for the Netherlands and Belgium instead of all three countries. Compared with a strategy of even layered generalization, the effect of this step is a reduction of the new layer from six cases to four. This particular manifestation of uneven layered generalization is depicted in Figure 9.3. As can be easily seen in comparison with Figure 9.2, uneven layered generalization produces a smaller layer of cases than one would achieve if one were only to ease to a scope condition. After the case study has scrutinized the link between exporter lobbying and quota liberalization for the Netherlands and Belgium, the layer can be completed

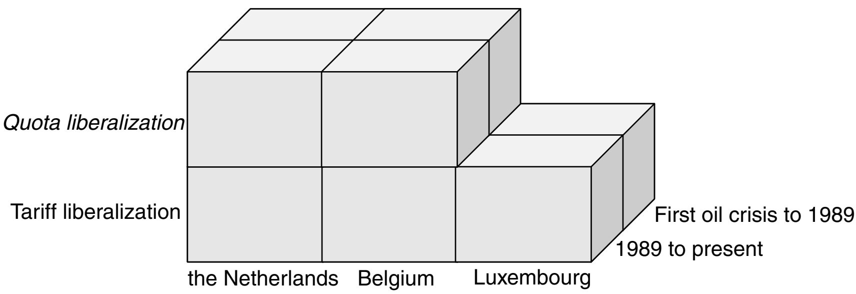

Figure 9.3 Uneven layered generalization

by a subsequent case study on quota liberalization in Luxembourg for the same two periods. The decomposition of a layer into multiple pieces allows one to perform in-depth case studies of manageable size while at the same time increasing the population step by step.

An alternative way to commit uneven layered generalization is to focus on all three countries and to limit the new layer to a single period. It is equally possible to proceed unevenly on the spatial and temporal dimension at the same time. This would be the case if one tests the hypothesis for quota liberalization on one or two countries for one or two periods. With respect to the advancement of knowledge, the strategy of uneven layered generalization entails an even slower speed of progress than layered generalization. As argued before, however, sweeping generalizations are not a value of their own. In case of doubt, the validity of generalizations should take precedence over their breadth.

The strategy of layered generalization as just described builds on and formalizes previous pleas for the need to lift scope conditions and test for the opportunity to expand a population (George and Bennett 2005, 25; Lieberson 1997, 376). The debate within the field of comparative historical analysis (CHA) particularly reflected on the importance of scope conditions (Mahoney and Rueschemeyer 2003a). CHA assumes that the analysis of big phenomena calls for historical specificity, which in turn demands many scope conditions so as to achieve a (usually) small population of rather homogeneous cases (see also Ragin 1987, chap. 2).[6] CHA's and my rationale for many scope conditions are similar, but they are motivated by slightly different reasons. In CHA, the baseline assumption is that cases *are not* comparable and that populations are too large if one stipulates too few scope conditions (Skocpol 2003). For the reasons detailed in this chapter, the concern with small populations keeps the generalization problem within limits.[7] Depending on the size of the population and available research resources, it might even be possible to examine all cases in the population, thereby eschewing generalization.

While CHA centers on the premise that historical specificity demands small populations, it is also acknowledged that it might be possible to lift scope conditions and broaden it. The strategy of layered generalization proposed in this chapter provides CHA with tools for testing the need of scope conditions. Equally important, the rationale for layered generalization extends beyond CHA. Even if one believes that historical specificity is not warranted, one might not be able to construct an optimal comparison simply because nature does not provide one with the appropriate cases. The generation of causal inferences can be enhanced by temporarily downsizing the initial population via the transformation of causes into scope conditions. In the previous trade example, layered generalization was illustrated with a population that started very small at the outset. Imagine now that one makes an argument about the influence of lobby groups that is not limited to any specific period in time such as the post–Cold War period. Suppose it also turns out that an ideal comparison is not available because too many potential causes vary over time. The population of theoretical interest, which can be called the *target population*, then can be cut down via the stipulation of a scope condition imposing a temporal constraint. The specification of a scope condition such as 'post–Cold War period' produces what can be coined the *working population*. It is a working population because it is a subset of the target population and forms an intermediate step toward the analysis of the target population via the repeated application of the strategy of layered generalization. This example points to the difference between the strategy of layered generalization in CHA and case studies not rooted in CHA. In CHA, the target population is small at the beginning in order to do justice to historical specificity and is gradually increased afterwards. In non-CHA case studies, in contrast, the target population can be large and the small population that the case study starts with is a working population. While the perspectives on the size of populations differ in both instances, layered generalization is equally useful for testing the need of scope conditions and assessing the opportunity to broaden the population.

9.4 Conclusion

The generalization of causal inferences has been identified as a significant problem in qualitative case studies. This chapter took a more nuanced perspective on generalization, shedding a less pessimistic light on the extension of causal inferences in small-n research. First, one should understand that different types of case studies imply different scopes of generalization. For instance, typical cases tend to make more limited generalizations than do most-likely case studies. Second, the empirical insights can be used to recast the size of the population by tightening or relaxing the scope condition that is affected by the evidence. Third, the diagnosis of a generalization problem implicitly presumes that the population is much larger than the number of examined cases. This might be so, but it does not have to be; a case study researcher can manufacture the population via the specification of scope conditions. Tightly delineated populations might seem of little substantive relevance, but this is not a matter of the case study method. From the perspective of methods, though, the generation of valid causal inferences and their generalization is central, which can be best achieved if the population is small. Moreover, this chapter shows that case studies are not confined to the analysis of small populations. It is true that every small-n analysis must be confined to a small population or layer of cases in order to keep the generalization problem within limits. In an iterative perspective, however, the disciplined and gradual relaxation of scope conditions can achieve the analysis of large populations via qualitative case studies.

10
Conclusion: Guided by Theory, Moving Forward Step by Step

This book started with two main assertions. First, a case study crucially though not exclusively depends on the research goal, the level of analysis, the nature of the causal effect, and the mode of causal inference in hypothesis-testing case studies. The chapters that followed confirmed this assertion and demonstrated that it is important to grasp the interplay of the dimensions in various parts of the research process. The second major argument was that case studies can be used for the building, testing, and modification of hypotheses on both the cross-case and the within-case level. The remainder of the chapter gives a broader evaluation of the degree to which case studies are suitable for the generation of causal inferences on regularities.

The previous chapters showed that I generally concur with what has been identified as the two major problems of case studies: an indeterminacy problem that bedevils internal validity and a problem of external validity deriving from the generalization of causal inferences from few cases to (probably) many. The problem of external validity directly follows from the adoption of a regularities framework. The assertion that case studies face a pertinent indeterminacy problem has been previously diagnosed for the cross-case level (Lieberson 1991; Zelditch 1971) and has been countered with the claim that process tracing can diminish or even eliminate it (Bennett 2008; Goldstone 1997). In Chapters 6 to 8, I argued that the extent to which process tracing can reduce indeterminacy is an empirical question because it depends on whether an outcome is due to monocausation or to more elaborate forms of causation. If one believes that phenomena are characterized by equifinality and interaction effects, a belief that reflects insights gained in many fields of research (Franzese 2008) and that is common in the small-n literature (Bennett and Elman 2006, 457; Checkel 2008, 126; Gerring 2007a, 61; Mahoney 2007b, 135), one should be more skeptical about the value of process tracing for the reduction of indeterminacy.

The two challenges to internal and external validity hold regardless of the characteristics of a case study, that is, the research goal, the level of analysis, and so on. This claim has an interesting implication because it qualifies the widely held view that case studies are highly suitable for exploratory purposes and far less useful for confirmatory analyses (Gerring 2004; Odell 2004). The evaluation of hypotheses is more preliminary in case studies in which the goal is the formation and modification of propositions rather than the testing of hypotheses. Nevertheless, it was shown in Chapter 4 that case studies are likely to confront the same problem of indeterminacy regardless of the underlying research goal. Exploratory small-n research allows one to generate hypotheses inductively, but they are likely to offer less specific guidance for subsequent hypothesis tests than is suggested by the literature that emphasizes the merits of exploratory case studies.

In this view, my perspective on the case study method seems to be rather pessimistic and a reinforcement of similar criticisms that have been raised before (Goldthorpe 1997a, 1997b; Lieberson 1998; Steinmetz 2004). However, the previous chapters highlight that this interpretation would be wrong. In particular, Chapters 5 and 9 showed that strong theory and the specification of scope conditions are two viable instruments with which one can enhance both internal and external validity. These two tools promote causal inference *within the context case studies interested in regularities and with instruments that are readily available to case study researchers*. This is important to underscore in light of two alternative conclusions that have been reached as a consequence of the indeterminacy and generalization problems. Both recommendations target the assumption of regular cause–effect relationships but resolve it from fundamentally different angles.

The first position maintains that a regularities framework is not suitable for case studies and that it is more promising to follow a philosophy of science that does not share this premise (Chatterjee 2009). Technically, a change of the philosophical perspective eliminates the problems that derive from a regularities perspective. However, it is odd to make the philosophical premises dependent on the number of cases that one can examine. The recommendation to switch to another philosophy of science means, at least implicitly, that such a change is not needed when one has enough cases for the application of large-n techniques because they make it easier to generate inferences (at the cross-case level only, though). If one believes that the social world is governed by regularities, one should stay true to this premise regardless of how many cases are examined empirically. Unsatisfactory as it seems, the problems that result from a regularity perspective should be diminished as much as possible with the available instruments and clearly acknowledged when generating causal inferences.

The second line of reasoning evolved out of the recent debate about the pros and cons of small-n and large-n methods and recommends the pursuit of multimethod research (MMR) (Bäck and Dumont 2007; Brady et al. 2006; Lieberman 2005). Notwithstanding that MMR introduces new problems, many of which have not been fully resolved by now (Dunning

2007; Rohlfing 2008), MMR is valuable because it combines the strengths of a large-n method such as regression analysis or Qualitative Comparative Analysis (QCA) with process tracing. In short, the large-n method serves to discern the causal effect, which is correlational in regression analysis and set-relational in QCA, whereas process tracing is done so as to shed light on the underlying causal mechanism. In comparison with case studies, MMR has a more solid basis for the analysis of causal effects because it usually draws on more cases. However, MMR does not solve the problem of generalizing insights on causal mechanisms because they are derived from a small number of in-depth case studies (Kühn and Rohlfing 2010). Besides this issue and notwithstanding that there are good reasons for performing MMR (Lieberman 2005), it accepts, at least implicitly, the downsides of case studies and therefore ties them to a large-n method. Although there is a huge difference between these two positions, they agree in their diagnosis that case studies find it difficult, if not impossible to overcome the two major challenges to internal and external validity. In contrast to this, I argued in the previous chapters that case study researchers can address both challenges via more elaborate theory and the transformation of potential causes into scope conditions.

To conclude this discussion and the book more generally, case studies are suitable for the building, testing, and modification of hypotheses on the cross-case and within-case level. However, case studies can fulfill this promise only if they are carefully crafted with an eye on their two most significant challenges. The strategy of layered generalization developed in Chapter 9 particularly calls for patience because the scope of generalizations can be increased only by doing one small and theory-guided step after another. However, the need for patience can hardly be invoked as a downside of case studies. Every method works better with a strong theory than a weak one (Achen 2002), and a sweeping generalization of causal inferences is not a virtue in itself. Most importantly, patience and a disciplined, gradual approach to causal inference and generalization are the prerequisites for bringing to fruition the case study method's inferential potential.

Notes

1 Introduction

1. The philosophies of social science that do not share the belief in the existence and detectability of regularities (see Jackson 2010, chap. 3) are largely ignored in this book. While some researchers argue that this view is doomed to fail or even that it has already been shown to have failed (Flyvbjerg 2001; Friedrichs and Kratochwil 2009), I regard this as an indeterminable ontological matter for which each researcher must make his or her own call (Hay 2006, 82). Ontologically, a regularities perspective is built on the belief in mind-world dualism and phenomenalism (Jackson 2010, 37).
2. Throughout the book, I do not make a sharp distinction between case studies and the case study method. The case study method is naturally tied to empirical small-n research, while the latter, in turn, must rely on the case study method for the generation of inferences.
3. This distinction generally mirrors the different views that (neopositivist) social scientists and historians take on cases and case studies (see Elman and Elman 2001).
4. Generalization is committed unless one is able to study all cases in the population.
5. In the social sciences, the democratic peace phenomenon is arguably one of the most famous examples of a causal effect that lacks, at least at present, the underpinning of a well-understood causal mechanism. Although it is not entirely undisputed (Gowa 2011; Rosato 2003; Schneider and Gleditsch 2010), it is widely agreed that any two democracies – a democratic dyad – are almost always at peace with each other and that a democratic dyad is the cause of this peace. What is missing from a thorough explanation of democratic peace is the causal mechanism that elucidates why two democracies maintain peaceful relations (George and Bennett 2005, chap. 2).
6. It seems that the understanding of causal explanation that I present in the text represents an emerging consensus (Cartwright 2004; Johnson 2006; Runde and de Rond 2010, 432).
7. Multimethod research focuses on the causal effect and causal process as well, but through the integration of at least two methods such as case studies and regression analysis.
8. Concept formation (Goertz 2006; Sartori 1970), the specification of the population (Mahoney and Goert 2004; Ragin 2000, chap. 2; Walker and Cohen 1985), and theory building (Geddes 2003; Rueschemeyer 2009) are equally important tasks. The implementation of these issues does not (at least should not) depend on the method that is applied. For this reason, they are not discussed in great detail or not at all.
9. Thomas (2011) proposes a typology of case studies that seems to bear resemblance to mine. However, it is less a typology, but a framework that brings together different elements of a case study such as the research purpose and the logic of comparison. Besides, it ignores important issues that are covered in this book.

10. 'Correlation' and 'covariation' are used synonymously in the following, though the two terms are not identical. Correlation measures the strength of an association without an underlying metric. Covariation is a measure for the association between two variables in terms of their respective units. The two concepts are closely related because one obtains the correlation between two variables by dividing their covariance with the product of their respective variances. In qualitative case studies, one cannot literally apply either of the two because the number of cases is too small. Heuristically, however, the terms have merit because they denote a specific conception of causal effects (see Section 2.4).
11. Gerring (2004, 343) contends that a single-case study is useless for causal inference because establishing a correlation between cause and effect requires two cases at minimum. However, single-case studies have merit if one believes in set-relational causation (see Section 2.4).
12. Necessary conditions come close to the idea of monocausation because the absence of a condition ensures the absence of the outcome. However, even here it is possible to have two (or more) necessary conditions that serve as functional equivalents (Ragin 2000), meaning that more than one condition qualifies as necessary.
13. The broadening of social science theory mirrors Sil and Katzenstein's (2010a, 2010b) plea for analytic eclecticism.
14. Some small-n studies may aim at the building, testing, or modification of an entire theory. However, most case studies pursue more modest goals and are only concerned with specific hypotheses. Hypotheses are necessary elements of a theory, but a theory entails more than a collection of propositions (Stinchcombe 1968).
15. If no hypothesis exists for the phenomenon of interest, one can choose between a hypothesis-building and a hypothesis-testing case study. The decision between the two is a matter of personal preference as both have advantages and disadvantages.
16. Researchers who test a hypothesis and find it at least partially disconfirmed should also speculate about potential reasons for the failure and possible avenues for modifying the proposition. Genuine hypothesis-modifying research centers on this goal and strives to reformulate a proposition on the basis of an exploratory case study.
17. The research process is rarely as linear as it is described here but can be represented accordingly in a stylized perspective.
18. Gerring (2007a) distinguishes between within-case studies, case studies, and cross-case studies, the last term being reserved for large-n comparative research. It is not clear why a qualitative macro comparison of two cases should not count as a cross-case study (see, for example, Ragin 1997, 36). I distinguish between the cross-case and the within-case level of analysis and use the notion of a case study as a generic term for small-n research operating on one or both levels.
19. It might be tempting to equate the cross-case level with the macro level and the within-case level with the micro level. However, this is not always feasible as the substance of the cross-case and the within-case level depends on the case study at hand. If one wants to explain the ideological behavior of parties in party competition, the cross-case level would refer to organizations – the parties – and not to countries. Moreover, if partisan action could be explained by the behavior of factions as units within parties, the within-case level would not be about individuals.

20. If the cross-case pattern is well-established, the case study also contributes to the establishment of a causal explanation because it helps to explain the cross-case pattern (Hedström and Ylikoski 2010).
21. This type might entail the semantic problem of talking of the cross-case level in the analysis of a single case. However, the more established notion of a single-case study is ambiguous because it does not say anything about the level of analysis that represents the theoretical end. Besides that, one can do a single cross-case study in a set-relational perspective on cross-case causation (see Section 2.4); it involves a cross-case element to the extent that the case belongs to a population of cases forming the background of the single-case study (Gerring 2007a, 22).
22. In the context of large-n research, the discussion takes place between advocates of regression analysis as a correlational method and proponents of Qualitative Comparative Analysis as a set-relational technique (Achen 2005; Grofman and Schneider 2009; Schneider and Wagemann 2012).
23. I think this qualifies the second point and a difference between ordinary frequentist significance testing in quantitative research and frequentist qualitative research. In significance testing, one calculates the likelihood of obtaining the result from the sample at hand (a correlation coefficient, an estimated marginal effect, and so on) if there is no association in the population from which the sample was drawn. If the result is statistically significant at a given level of significance, the hypothesis that there is no association in the population is rejected. What the association is in the population is still unknown because one only tests whether there is no association. In frequentist case studies, the basic reasoning is similar because one uses the evidence for judging how likely it is to obtain this evidence if the null hypothesis is wrong. When one concludes that it is too unlikely (an informal assessment of statistical significance), one can first reject the null hypothesis stating that there is no association in the population. The salient point now is that case studies pit a substantively meaningful hypothesis against the null hypothesis. In the transposition example, the proposition states that misfit and the ensuing adaptation costs account for delays. The corresponding null hypothesis stipulates that misfit and adaptation costs do not drive transposition failures. In this setting, the rejection of the null hypothesis on the basis of empirical evidence thus implies the confirmation of the misfit hypothesis.
24. The reverse is not true because frequentist case studies are void of any reference to the theoretical implications of observations.
25. There is no single best way to structure the discussion of the case study method. Among other things, this is due to the variety of means-ends relationships between the cross-case and the within-case level because they imply different orders in which the two levels come into play.

2 Case, Case Study, and Causation: Core Concepts and Fundamentals

1. Throughout the book, I use 'causality' and 'causation' synonymously.
2. The understanding of what the relevant case is may change over the course of the case study. Acknowledging this, at any given point of the case study, one should be able to state what the case of interest is.
3. See Mahoney and Goertz (2004) and Goertz (2006) for negative cases.

4. At least unless one tests for the possibility of dropping a scope condition and expanding the population (Bennett 2005). In this instance, at least one of the selected cases should not be a member of the original population.
5. On the cross-case level, though, the quantitative case study would be similar to a qualitative case study (see Chapter 4) because one would simply link a specific type of classification strategy to the campaign outcome.
6. There is some debate about what causal inference requires in case studies and empirical research more generally (Goertz and Mahoney 2012). The debate follows the large-n versus small-n divide. Large-n researchers infer causation from a cross-case pattern. Small-n researchers raise the criticism that an association is not causation and thus generate inferences about causal mechanisms and causal processes from within-case analyses. Given my integrative perspective on causal relationships, causal inference is strongest when it builds on both levels of analysis and can make a credible claim for the presence of a causal effect and a causal mechanism. Consequently, causal inference is weaker if it is concerned with one level only. Since causal inferences are made on both levels, I use this general term in the context of cross-case and within-case analyses alike.
7. There is no clear-cut benchmark separating case studies from other types of designs, which is why I leave it open what 'small' means. Still, this view on the case study contradicts Tight's (2010) assertion that the notion of 'case study' is superfluous because everything can be taken as a case study.
8. The term CPO is contested. Beck (2010, 2006) argues that it is an oxymoron because causal processes and causation more generally are unobservable. Brady, Collier, and Seawright (2006, 2010) concur with the claim that causation is unobservable. However, they also emphasize that they wanted to coin a label signifying the different types of observations that one uses in qualitative and quantitative research. The distinction between DSOs and CPOs certainly achieves this goal. However, some uneasiness remains and it extends to the term causal-process tracing (Blatter and Haverland 2012). Process tracing is about tracing an empirical process and causation can only be *inferred* on the basis of the gathered evidence.
9. The cross-case level refers to electoral districts and not to the national level, highlighting that the cross-case level is not necessarily the same as the macro level (i.e., country level).
10. It is also known as a *closed-systems perspective* (Kurki 2007, 47). Kemp and Holmwood (2003, 176–7) invoke the notion of a spontaneous regularity. I do not use this term because it strikes me as an oxymoron. 'Spontaneous' contains an element of chance and contradicts the notion of a regularity, which is not spontaneous but regularly occurs, on average, under certain conditions.
11. These are instances of type-level causation (also known as general causation), which is integral to some philosophical theories of causation (Hitchcock 1995). The first statement links two types, or classes, of phenomena to each other, namely, exporter lobby groups and a country's trade policy. Type-level causation contrasts with token-level causation (also known as singular causation). An example of token-level causation is 'In 1920, the United States pursued a protectionist trade policy because exporters lobby groups in the USA were weak'. In contrast to the type-level statement, token-level claims refer to specific places and points in time, involve particular institutions, and so on. Case studies as I discuss them in my book look at instances of token-level causation in order to infer something about type-level causation.

12. Although it is now of considerable age, Hempel's (1965) covering law model of explanation still serves as a point of reference in the social sciences (for example, George and Bennett 2005, chap. 7). The covering law model also alludes to regularities but is very different from the perspective underlying this book. In an analysis based on covering laws, individual cases are 'explained' by subsumption under statements expressing a regular causal relationship. In this book, in contrast, cases are used to build, test, or modify general hypotheses and are not simply subsumed under them.
13. Hume argues that the effect should always follow the cause, which is equivalent to a deterministic causal relationship. Currently, this requirement can be and is relaxed by allowing for probabilistic causation; that is, the outcome mostly follows the cause (leaving open here how often 'mostly' is).
14. In contrast to this position, Gerring (2005) and Pierson (2004, chap. 3) make a case for distant causes because they are less obvious than proximate causes.
15. Probability-raising conceptions of causation are inherent in most quantitative research, which is often criticized because a cross-case association does not mirror causation (Abbott 1998). This criticism, which is justified, is a major argument for process tracing and the analysis of causal mechanisms (Dessler 1991; George and Bennett 2005, chap. 10). However, a change in the level of analysis is no reason to abandon a given conception of causation (Woodward 2011), as this is an epistemological decision that is independent of whether one is concerned with the cross-case or the within-case level. In addition, I do not claim that a probability-raising conception of causation is the best epistemological view on causation. However, all conceptions, including mechanismic accounts, have their share of problems (Waskan 2011). Finally, as many have noted before, a probability-raising conception of causation is *not* incompatible with the ontological view that an empirical entity has the power to bring about another empirical phenomenon. But since this power is unobservable, one needs some epistemological criterion for causal inference.
16. If one does process tracing in a case without generalizing the insights, one is simply formulating an explanation invoking a mechanism. Strictly seen, neither the singular explanation nor a singular mechanism qualifies as causal when one is operating with a probabilty raising view on causation (Russo and Williamson 2007).
17. Some small-n researchers, who are in my view not anchored in scientific realism, borrow the idea of causal mechanisms from scientific realism (for scientific realism, see Bhaskar 1975; Demetriou 2009; Lane 1996; Manicas 2006; Sayer 2010). Many scientific realists argue that one cannot detect empirical regularities in the social sciences (for a minority position, see Downward et al. 2002; Kemp and Holmwood 2003). A researcher interested in inferences about regularities should be careful about turning to scientific realism in order to avoid producing a philosophical hodgepodge.
18. The value of mechanism analyses is disputed. An archetypical case for a critical perspective can be found in Gerring's discussion of mechanisms (2010). Gerring argues that mechanismic thinking is bedeviled with definitional ambiguities, that it is not new and not needed, and that process tracing is too demanding empirically. It's true that the literature is conceptually ambiguous, but many definitions share an intelligible core meaning of causal mechanism that one can build on (Gerring 2008, 161). The fact that it is also true that mechanismic thinking is not new is unrelated to its relevance. On relevance, Gerring argues that one can infer causation in experiments without knowing why a causal

relationship is in place. But experiments are not always available and there are also threats to experimental causal inference (Morton and Williams 2010). In addition to his, consider an experiment in medicine demonstrating the success of a new drug against cancer. The problem is that it is not known why this drug works and it has severe side-effects. If one would know how the drug fights cancer, one could try to design a drug that is as effective but lacks the side-effects. Finally, no serious proponent of process tracing would deny that it is demanding to implement. However, if this were the criterion for the appropriateness of methods, social scientists could not apply any method at all because no method is foolproof.

19. Causal mechanisms are sometimes also referred to as social mechanisms (Hedström and Swedberg 1996), denoting mechanisms that are operative in the social world.
20. There is an inconsistency between Figure 2.2 (and 2.3 and 2.4) and Ziblatt's cross-case hypothesis stipulating that greater landholding inequality leads to an increased likelihood of electoral fraud. The figures put landholding inequality, electoral fraud, and capture into relation without establishing a positive correlation. Figures 2.2 to 2.4 therefore should not be mistaken as visualizations of the theoretical argument. However, the figures are correct insofar as the mechanism is 'capture of the local administration' and not higher levels of capture of the local administration (in response to higher levels of landholding inequality). At the same time, the expectation of a positive correlation and varying levels of capture, depending on the extent of landholding inequality, is compatible with the visualizations in Figures 2.2 and 2.4 because they are open to the imputation of any kind of causal effect.
21. Machamer, Darden, and Craver developed their conception of mechanism for the life sciences, in which it is appropriate to speak of molecules as entities, and these entities are described as engaging in activities, such as connecting to other molecules.
22. Waldner (2012) argues that the mechanism rests in the arrow connecting two intervening steps, which relates back to the question of whether mechanisms are observable or unobservable. This distinguishes Waldner's take on mechanisms from that of Machamer, Darden, and Craver (2000), who take the arrows as indications for an activity, and Mahoney and Goertz (2012) and George and Bennett (2005, chap. 10), who seem to consider intervening steps to be manifestations of a single mechanism. This aspect is discussed below in the main text.
23. The question of the detail of information is not related to the problem of infinite regress in mechanismic reasoning. Skeptics of mechanismic explanation argue that one can always try to discern lower-level mechanisms (Gerring 2010). Instead of treating individual behavior as constitutive for a mechanismic explanation, one could rely on neurological factors. Addressing this claim, Machamer, Darden, and Craver (2000) convincingly argue that theory determines the level at which a mechanismic explanation bottoms out. If we want to explain a cross-case outcome via individual behavior (or the behavior of collective and corporate actors), it is perfectly legitimate not to delve deeper. In the Ziblatt example, the question of whether to specify one, three, or more mechanisms is therefore always answered on the individual level and does not imply that, for instance, a triple-mechanism explanation is located on a lower level of analysis than a single-mechanism account.

24. Considering the plethora of definitions of mechanisms, the claim that my discussion of the case study method is compatible with all of them might be wrong. However, I have not encountered such a definition thus far. (The reader is invited to get in contact with me in order to discuss definitions invalidating my claim.)
25. What might strike one as conceptual imprecision mirrors admonitions from philosophers of science that a uniform definition of mechanism might not be appropriate across the disciplines of biology, physics, and the social sciences because these disciplines refer to different objects when talking about mechanisms (Machamer et al. 2000; Woodward 2011). What holds across disciplines might also hold within a discipline because, as explained in the main text, different forms of mechanismic explanation put different aspects in the theoretical forefront. The goal then would not be to find and impose a single conception of mechanism on every empirical analysis but to adopt (and make transparent) a definition of mechanism that meets the theoretical goal.
26. Kittel (2005), referring to Casti (1989), invokes the term operation science, which is similar to X-centered research.
27. In this sense, they are also Y-centered. Kittel (2005), referring to Casti (1989), additionally invokes the term 'origin science'.
28. See Nassmacher (2010) for a general argument against the latter distinction.
29. Before the term CHA was coined, debates often took place under the rubric of macro-sociological research (see, for example, Goldstone 1998; Goldstone 1997; Kiser and Hechter 1991; Nichols 1986; Quadagno and Knapp 1992; Skocpol 1986; Skocpol and Somers 1980).
30. I would add that nothing speaks against the formulation of all-encompassing explanations in combination with the belief in correlational causal effects.
31. I ignore that there is another line of cleavage between frequentist and Bayesian quantitative researchers (Howson and Urbach 2005).
32. Correlational and set-relational cause–effect relationships can be deterministic and probabilistic (Mahoney 2004; Mahoney and Goertz 2006). I do not address this issue in more detail here.
33. For all examples that follow, I assume that the effect holds on average.
34. One can take any two cases with different values on a continuous variable to establish a correlation. However, it may be that the difference between two cases is not statistically significant. This would open the floor to the argument that one explains differences in degree that are likely to be the result of chance. This criticism can be prevented by picking cases that take scores where the differences are statistically significant.
35. This measurement strategy mirrors the calibration of variables in QCA (see Ragin 2006b). Second-best alternatives to theory and concepts are conventions and simple statistical criteria, such as taking the mean of a variable as the threshold (Rihoux and De Meur 2008).
36. The opposite strategy of explaining differences in kind with differences in degree is not viable. If one observes that differences in degree produce a difference in kind, one has found evidence for a relationship between two differences in kind.
37. This discussion presumes *crisp sets*; that is, cases are a member of a set or not (Ragin 1987, chap. 6). *Fuzzy sets* additionally allow one to measure the degree to which a case is in or out of a set (Ragin 2000, chap. 6). For ease of presentation and without loss of generality, I limit the discussion to crisp sets.

38. Another logically equivalent definition of necessity states that the outcome is absent whenever the condition is absent. In formal terms, this reads $\sim X \rightarrow \sim Y$, where the symbol ~ denotes the negation of a set. The understanding of necessity that is presented in the main text is more widely used in the social science literature.
39. All other set relations derive from necessity and sufficiency. Consequently, asymmetry is a characteristic of all types of set relations and is not addressed any further in the following paragraphs.
40. It is possible to join two or more necessary conditions by the OR operator. This means that $X1 + X2 \leftarrow Y$ and that whenever the outcome is present, one observes either X1 or X2 or both. The logical OR signals that the two necessary conditions are indicators of the same higher-order concept (Ragin 2000). If one refers to this higher-order concept as C, one can substitute X1 + X2 with C; the resulting formula then reads $C \leftarrow Y$.
41. Individually sufficient conditions can be simultaneously present. In this instance, the outcome is *overdetermined* because there are several sufficient conditions, each of which can produce an outcome of its own (Schaffer 2003).
42. This implies that the idea of a SUIN cause is not applicable to conditions defined by a single attribute. Furthermore, SUIN causes do not pertain to necessary/sufficient concepts (see Goertz 2006, chap. 2).
43. I leave open here how 'frequent' is defined.

3 Types of Case Studies and Case Selection

1. See Eckstein (1975), George and Bennett (2005, chap. 5), Levy (2008), Lijphart (1971), Seawright and Gerring (2008), and Yin (2008) for discussions of various types of case studies.
2. See Klotz (2008) and Seawright and Gerring (2008) for other treatments of case selection.
3. Some types of case studies require the choice of two cases. Keeping this in mind, I use the singular when talking about types of case studies. This means, for example, that I speak of a typical case even if the correlational variant of the typical case study includes the comparison of two cases.
4. Chapter 2 dealt with additional variants of set relations. A detailed discussion of all forms of set relations is beyond the scope of this chapter.
5. *Plausibility probes* are sometimes presented as a separate type of case study (Eckstein 1975, 108–13; George and Bennett 2005, 75; Levy 2008, 6–7). I decided against this type because plausibility probes do not yield added value when compared with the types of case studies presented in the main text (David Kühn eventually convinced me of this point). On the one hand, plausibility probes are presented as first tests of a hypothesis in order to see whether a more thorough test is warranted. On the other hand, some argue that a plausibility probe should not have lower standards of collecting and interpreting evidence than genuine tests (George and Bennett 2005, 75). (This indicates that plausibility probes might be better conceived of as a fourth type of research goal rather than a type of case study, as they are located between exploratory and confirmatory research.) As regards the latter point, it is not obvious what the difference between a plausibility probe and a conventional hypothesis-testing case study is and why plausibility probes are needed. If one adopts lower standards, however, a plausibility probe is of very limited value because it is not too difficult to come

up with a case that lends superficial support to a given hypothesis (Fearon and Laitin 2008).

6. Since one selects cases with respect to their utility for the generation of causal inferences and their generalizability, there is some overlap between this chapter and the next chapters. In order to achieve a minimum of redundancies, I limit the arguments on causal inference to those points that are indispensable for an elaboration of case selection principles.
7. This problem qualifies the advice about relying on theoretical sampling in hypothesis-building research (Eisenhardt 1989, 537). Theoretical sampling is understood as using theory for the identification of cases that are expected to deliver the most promising insights. Theory is available when one intends to develop a within-case hypothesis but is necessarily weak for the cross-case level.
8. Invariant causes cannot be discarded as potential independent variables. Since they are invariant, it is neither possible to make a confirming causal inference nor to generate a disconfirming inference.
9. The reason for this lies in the definition of sufficiency. Assume you are interested in the occurrence of revolutions and select two countries that experienced a revolution. In country A, a famine preceded the revolution but not in country B. What does the no-variance-on-Y comparison tell us now? Nothing. If a famine is sufficient, country A is in line with this pattern while country B is irrelevant. If no famine is sufficient, country B is in accord with a pattern of sufficiency and country A is irrelevant. In practice, of course, the cross-case analysis is supplemented with process tracing that might indicate whether a famine or no famine is sufficient. Nevertheless, one can increase the inferential leverage on the cross-case level by crafting a variance-on-Y design.
10. One issue to be taken into account is that an actually sufficient condition might be absent in all selected cases. Building on the example in the previous note, suppose a famine is sufficient for a revolution but that all selected cases are characterized by the absence of a famine. The cases then fail to deliver any evidence for the sufficiency of a famine. For this reason, one should always be aware of the possibility that an exploratory comparison does not point to all sufficient conditions of an outcome (George and Bennett 2005, 27). Moreover, the problem is that the absence of a famine seems to be sufficient though it is not in fact.
11. A classic sampling bias is introduced if one selects only cases with scores above or below a specific threshold on the outcome and if this selection rule is correlated with an independent variable of interest (King et al. 1994, 128–49).
12. Collier and Mahoney (1996) correctly assert that the actual problem Geddes and King, Keohane, and Verba are referring to is one of having no variance on the outcome, which they term no-variance design. I go beyond this term and distinguish between no variance on the outcome and on a cause.
13. The reason that the correlational diverse case study is limited to continuous and multi-ordinal measures can be highlighted by considering the two remaining alternatives (for purposes of illustration, I assume that one selects cases on the outcome). First, the outcome can be binary. In this instance, a correlational case study is necessarily diverse because it covers both categories. Second, the outcome could be multinominal. Seawright and Gerring (2008) then recommend the selection of cases from each category (their advice extends to multicategorical measures in general). Cases could be selected accordingly, but this effectively means analyzing multiple typical cases because each selected case is typical for

all other cases from the same category. A diverse case study thus is the same as a comparative case study on typical cases belonging to different categories.

14. Intentional case selection on the cause and the outcome seems to be in discord with King, Keohane, and Verba's admonition not to rely on this strategy. However, this contradiction can be easily resolved because King, Keohane, and Verba (1994) are exclusively concerned with hypothesis testing on the cross-case level.
15. When all other cases are closely located to one of the two extreme cases, the diverse case study is equivalent to the typical case study.
16. The set-relational diverse case study is equally applicable when one calibrates a multi-ordinal variable and the resulting set includes two or more categories. In this instance, one should pick one case from the highest-ranking category and one case from the lowest-ranking category belonging to the same set.
17. More precisely, one gathers evidence that the set does not include too many countries. Still, it could be that the set is too narrow and that countries with an even lower GDP display the outcome as well.
18. The argument could be that the negative effect of economic openness on spending levels decreases as the unemployment rate increases because more unemployment creates higher demand for social spending.
19. The choice of cases with similar levels of unemployment means that one is not performing a comparative test. Comparative testing quickly runs into inferential problems in correlational two-case comparisons (Tarrow 2010). In correlational case studies, the essence of comparative testing is *not* to try to test all hypotheses in one analysis (see Chapter 4). In set-relational case studies, parallel testing is possible in two-case comparisons when no conjunctural causation is involved. The two cases then should be selected accordingly.
20. This strategy reflects Lijphart's (1971, 687) long-standing recommendation to maximize the variance on the variable of interest and to minimize the variance on control variables.
21. This case selection strategy is different from what is known as theoretical sampling in grounded theory (Wasserman et al. 2009).
22. A most-likely case is also referred to as an *easy case* because it is easy for the hypothesis to pass the test. A least-likely case is also known as a *hard case* or a *Sinatra case*: if the hypothesis can be confirmed with a hard case, it is likely to be confirmed by all other cases in the population as well.
23. Sometimes, the *extreme case study* is proposed as a distinct distribution-based type (for example, Seawright and Gerring 2008, 301–2). A case is defined as extreme when it is distant from the bulk of other cases on one or more causes *or* on the outcome. An extreme case study is appraised as suitable for exploratory case studies in order to better understand the extreme case. The problem that I see is that a case that is extreme on an independent variable may be typical or deviant once we take the dependent variable into account. Since we do not know which of the two scenarios holds true unless we consider the independent and the dependent variable, we do not know what case we are dealing with as regards the causal relationship of interest. However, this is what matters above all. An alternative would be to perform a diverse case study wherein the extreme case constitutes one of the two diverse cases. But such a case study would resemble a least-likely case study because its extremeness causes us to assume that it is unlikely that the causal process is similar in an extreme and nonextreme case. For these reasons, I see little justification for the extreme case study as a distinct type.

24. If one deems it too demanding to give a point estimate, one can specify a range of probabilities (say, 70–90 percent instead of 80 percent) or a distribution of priors. For reasons of presentational convenience, I simply speak of probabilities in the following.
25. Sometimes, the literature only refers to the probability of E on the condition that H_{DP} is true as $p(E|H_{DP})$. In case study research, this likelihood is not as informative as it could and should be because the probability of observing the outcome can only be determined by additionally considering case selection in ways to be detailed below. (I owe this insight to Christina Zuber.) For instance, Dion (1998) uses the case-free conditional likelihood in his discussion of case studies for the assessment of necessary condition hypothesis. Implicitly, it is then argued that the hypothesis should be confirmed independently of what case is chosen. This assumption can be made (in particular when one is interested in single cases and singular causation (see Abell, 2009a), but it is a strong assumption because it is more compelling that different cases entail a different likelihood of confirming a hypothesis. Sometimes, one also finds the notation $p(E|H \& B)$, where B represents background knowledge (Howson and Urbach, 2005). In short, background knowledge subsumes issues that, in principle influence the likelihood in question (such as the quality of sources used in empirical research), but that are not explicitly considered in the formulation of the conditional probability. In order to keep the notation simple, I leave the parameter aside here.
26. Constantelos performs a quantitative analysis. This underscores Levy's claim that the general idea behind most-likely and least-likely sampling extends to quantitative research (2007).
27. For the cross-case level, an additional type of case invented recently is the *pathway* case (Gerring 2007b). Because the pathway case is the same as a typical case when controlling for rival explanations (see Chapter 4), the pathway case is not introduced as a separate type here.
28. George and Bennett (2005, chap. 9) propose the *congruence method* for qualitative case studies. In my view, it follows the logic of ordinary hypothesis-testing. In a case study on sufficiency, the congruence method involves the selection of cases on the cause and to test for the presence of the outcome (George and Bennett, 2005, 181). They elaborate other aspects of the congruence method such as the search for spuriousness. These issues are important, but not special to the congruence method because they reflect the good practice of multiplying the number of observable implications that are related to a hypothesis.
29. It is possible to examine the same case in the exploratory and confirmatory part. However, it is more likely that the focus will be on different observable implications, which usually implies different cases that belong to the same units (see Section 5.1). The arguments that are made in the main text extend to analyses centered on the same case or the same unit.
30. Chapter 4 elaborates additional sources of anomalies.
31. This can be framed as a sufficiency case study when the interest lies in the effects of a democratic dyad (failed most-likely case). If the case study is interested in the outcome war, it is about the presence of the outcome in the absence of the necessary condition (passed least-likely case).
32. Populations tend to be small in comparative historical analysis (CHA) because it relies on concepts with high intensions and many restrictive scope conditions (Skocpol 2003). Consequently, the number of cases one has to screen in CHA tends to be smaller than in ordinary case studies.

4 Forms and Problems of Comparisons

1. Causal inferences can be confined to the statement that a cause matters, without an additional inference about the type of causal effect. This avoids engagement with the issues discussed in this and the next chapter, but there remain some pertinent problems that will be the subjects of Chapters 6, 7, and 8.
2. Since this chapter is concerned with theory and the accumulation of knowledge, I presume that all comparisons follow the idea of structured, focused comparisons (George 1979). A comparison is structured when one asks the same questions as regards all cases under scrutiny. A comparison is focused if the empirical analysis is guided by theory and concerned with particular aspects of a case (George and Bennett 2005, 67).
3. Sometimes this is not fully apparent because Mill's methods may come in the guise of an experimental template (Gerring and McDermott 2007), the most-similar design for the method of difference (Zangl 2008), or the most-dissimilar/most-different design for the method of agreement (De Meur and Berg-Schlosser 1996); in addition, the underlying logic of inference is implicitly used only in cross-case comparisons (Slater 2009).
4. I say 'largely free of theory' because experiments (and Mill's methods) require theoretical input as well (Cohen and Nagel 1934, 252). But according to Mill, less theory is needed in experiments than in observational research. One might find it odd to speak of theory-free observations in empirical research, but that's the context in which Mill discusses his methods.
5. As regards the method of difference, Mill (1874, 610–1) additionally argues that one will never find two cases that display the required cross-case scores (see below). This is an empirical matter and may or may not be the case. Moreover, social scientists rarely have the nice data that they need for their methods to work properly, and one can take this into account by framing causal inferences accordingly (King et al. 1994, 28).
6. Before a test is performed, one must decide between a correlational and set-relational perspective and formulate the hypotheses that are to be tested accordingly. Unfortunately, empirical evidence can hardly be a guide in making this decision.
7. Set relations are about invariance, meaning that a single case would suffice for building a hypothesis from scratch. But set-relational designs can also benefit from comparisons in ways described in the following.
8. Zelditch (1971) and Collier, Brady, and Seawright (2004b, 238) prefer the term 'interpretable' to 'indeterminate'. With an eye on statistical analysis, they argue that a design can be determinate but that the results are not interpretable nevertheless. The notion of an interpretable design is valuable, but I prefer the term 'indeterminate' (in philosophy of science, it is referred to as contrastive or empirical underdetermination because theory is underdetermined by evidence (Park 2009). Many issues have the potential to undermine the interpretability of an analysis. The inferential problem that is at the core of this chapter is one of indeterminacy and should be labeled as such instead of being obscured by the more general term 'interpretability'.
9. Indeterminacy is sometimes also referred to as a *degrees-of-freedom problem* (Campbell 1975). In quantitative research, one has insufficient degrees of freedom when the number of parameters one has to estimate is equal to or larger than the number of cases. The consequence is that the model perfectly fits the data at

hand or cannot be estimated at all. I prefer indeterminacy to degrees-of-freedom problems because the notion of degrees of freedom cannot be easily transferred to the realm of qualitative case studies and tends to be more confusing than illuminating (George and Bennett 2005, 28).

10. See Lazarsfeld (1937) and Barton (1955) for early discussions of the property space in social research. The idea of a property space is usually invoked in typological analysis and typological theory where one measures differences in kind (George and Bennett 2005, 234). However, it can be generalized to correlational qualitative case studies that rely on differences in degree. Imagine a case study that is interested in the effects of economic openness and labor union strength on welfare spending. This analysis involves four types: globalization and labor union strength can be equal in both countries, globalization can be higher in one country and union strength similar, globalization can be similar and labor union strength higher in one country, and both can be dissimilar in both countries.
11. In QCA, this problem is also known as one of *limited diversity* (Ragin 1987, 104–5).
12. The role of case selection for cross-case inferences in hypotheses tests should be recalled. As explained in Chapter 3, a test of correlational propositions requires establishing variance on the outcome or the cause of interest. The distinction between differences in degree and differences in kind, which is salient for case selection, does not matter here. In set-relational research, causal inference should be limited to claims of necessity when the two cases are chosen with the outcome present and sufficiency when cases are selected on the conditions. This information cannot be communicated via the stylized visualizations of comparisons in this chapter. In the discussion of set-relational causal inferences, I therefore presume that the cases were chosen according to the principles laid down in Chapter 3.
13. The MoA is also referred to as the *most-dissimilar design* (Berg-Schlosser and De Meur 1994; Berg-Schlosser and Quenter 1996), which is attributable to Przeworski and Teune's discussion of country comparisons (1970, chap. 3). They introduce their design as the most-dissimilar systems design (MDSD) because of their preoccupation with cross-national research. (The logic of inference does not mandate the analysis of nation states as long as one can distinguish a cross-case and a within-case level.). Treating the MoA and the MDSD as synonymous is misleading because the MDSD follows a different logic of causal inference (see Przeworski and Teune 1970, 34–5).
14. Cohen and Nagel argue that Mill's methods are useful only for elimination of causes, not for confirmatory causal inferences (e.g., 1934, 252–4). They make the principled claim that one can never be sure that there will be future cases that contradict the inferences derived from Mill's methods. This is true, of course, but for the time being – that is, unless new cases enter the picture – Mill's methods do permit the making of confirmatory inferences, not just the elimination of causes.
15. This reading of the MoA contrasts with DeFelice's argument that the observation of an invariant factor in the MoA allows one to infer covariational causation (1986, 421). If covariation is the benchmark, the MoA does not allow one to make a hypothesis-confirming causal inference.
16. An additional problem occurs when a cross-case cause is a composite measure indicating that multiple constituent factors are jointly present. Savolainen (1994, 1221) argues that aggregate measures are a way to handle interactions

when using Mill's methods. But the aggregation of multiple causes into one only conceals the problem of inferring an interaction effect and is not a solution to the actual problem (Nichols 1986). In the Hendriks and Michels example, nothing would be gained in inferential terms if the institutional variables were integrated in the composite causes 'consensus democracy' and 'majoritarian democracy'. The problems can be avoided only if the theoretical interest lies in a composite measure such as consensus democracy and not in the constitutive dimensions.

17. If one takes into account that multiple necessary conditions can be connected via the logical OR operator (see Chapter 2), one cannot automatically discard these three conditions as necessary. While one could make a claim of substitutability in this example, I do not address this issue in more detail here.
18. Typological theory takes a different set-relational perspective than the one adopted here (Elman 2005; George and Bennett 2005, chap. 11). In short, typological theory takes configurations of conditions as types. In this view, the United Kingdom and the Netherlands are two different types out of 16 possible types (four dichotomous conditions produce 16 logically possible types). All countries that display the same configuration of conditions as the United Kingdom and the Netherlands are of the same type. The default assumption is that the entire conjunction of conditions is sufficient for the outcome. There is no aim to single out one condition or a subset of conditions as sufficient. Typological theory and the set-relational interpretation adopted here therefore take exactly the opposite perspective; the standard view focuses on the sufficiency of a single condition, while typological theory presumes that the largest possible conjunction is sufficient. As the discussion in the main text shows, both inferences can be made, but only in the presence of strong theory or demanding counterfactual assumptions about what the outcome is in cases described by configurations of conditions that do not exist empirically.
19. This aligns with Savolainen's assertion that a comparison should be constructed according to what one aims to find out (1994, 1220–1).
20. Both inferences are made under the hypotheses that the presence of a debate, a majoritarian electoral system, a single-party government, and unitarism are individually sufficient.
21. The MoD is also referred to as the *most-similar design* because the cases are very similar with respect to the causes. The notion of a most-similar comparison is related to Przeworski and Teune's most-similar systems design (MSSD) (1970, chap. 3). The MoD and the MSSD can be considered equivalent.
22. The substance matter is treated as a shortcut for factors that are related to hormone-treated beef.
23. Gerring and McDermott (2007, 691) call this a spatial comparison. According to them, the comparison that mirrors an experiment most is a dynamic comparison drawing on variation across cases and for the same case over time. The notion of a spatial comparison does not apply to Zangl's study involving international institutions, but it can be more generally put as a post-treatment-only comparison of the treatment and control case.
24. This applies to all kinds of designs that mimic an experiment. For this reason, I see little ground for arguing that one form of observational comparison comes closer to the ideal of an experiment than another (Gerring 2007a, chap. 1; Gerring and McDermott 2007).
25. Tilly (1997) recommends not to implement small-n comparisons for theoretical reasons. He emphasizes that societies and nations are not necessarily useful

entities for comparisons (see also Ragin 1981). This point is correct to the extent that every researcher must ensure that the entities of interest are amenable to a comparison (Sartori 1991), which might or might not be the case for countries and societies.

26. I leave aside here that the cause–effect relationship could also be set-relational. Similar problems apply here.
27. Here, 'probabilism' means that causal relationships are inherently probabilistic and cannot be turned deterministically by reshaping a concept, acquiring better data, and so on (see Salmon 1998, chap. 2).
28. For purposes of illustration, I focus on puzzles including single causes.
29. In contrast to Lange, I use dichotomous measures for the outcome (see also Chapter 5).
30. Columns with entries in italics include causes that are added to the comparison after exploratory process tracing has been performed.
31. Left governments might prefer centralized solutions, while right-wing governments prefer a decentralized approach.
32. For example because left governments attach more salience to low levels of inequality than right-wing governments.
33. If early childhood education is not given, one again faces a puzzle that can be approached in the same manner.
34. Oliver is more specifically interested in the fact that inequality is rising in Sweden and declining in Italy. In the later 1990s, the levels of inequality reached similar low levels, and this is the point I am focusing on.

5 Enhancing Causal Inference in Comparisons

1. Elman's (2005) formidable discussion of explanatory typologies, which underlies parts of the following paragraphs, includes five related strategies for diminishing the size of the property space. Some of the strategies do not extend to cross-case comparisons, where the notion of a 'property space' has a slightly different meaning because a logically possible combination of cross-case scores does not automatically represent a type. Moreover, Elman does not relate the strategies for reducing the space to the viability of cross-case inferences and indeterminacy because of his concern with typological theory.
2. Gerring addresses the two dimensions in the context of a covariational typology of research designs (2004, 343), yet without arguing that they are characterized by a trade-off. Moreover, the two dimensions can be at least partially extended to set-relational comparisons.
3. If the comparison of two sectors in the same country is a spatial within-unit analysis, it follows that a substantive within-unit comparison would involve the comparison of identical sectors.
4. Because of the preoccupation with experiments, discussions of this design always presume that one observes changes on a cause. In correlational and necessary condition case studies, it is equally possible to observe change of the outcome over time and to determine whether one cause changes, too.
5. Bartolini (1993) takes a more skeptical perspective in the context of a discussion of developmental theory.
6. See Lieberman (2001), Katznelson (2003), and Grzymala-Busse (2011) for a general discussion of periodization in Comparative Historical Analysis, that is, the delineation of the relevant period of analysis for case studies.

7. One may dispute these qualifications because the literature on interrupted time-series (ITS) designs, which is a template for longitudinal comparisons (Collier 1993; George and Bennett 2005), recommends focusing on broad periods of time before and after the treatment (Cook and Wong 2008). However, the analogy is misleading because ITS designs include continuous outcomes, such as traffic fatalities, that are subject to natural fluctuations (Campbell and Ross 1968). In order to focus on the effect of a treatment, it is necessary to extend the period of analysis as nonsystematic fluctuations can cancel each other out on average. This point does not extend to qualitative case studies that do not center on continuous outcomes and are interested in phenomena that rarely change. An additional reason that the ITS design might be appealing is that it is one legitimate variant of a quasi experiment (Cook and Wong 2008). However, if one wants to transfer quasi-experimental designs to the field of case studies, one would better follow the idea of a differences-in-differences design. Without going into the details here, that design mirrors the structure of a combined within-unit and cross-unit comparison – or a dynamic comparison in terms of Gerring and McDermott (2007).
8. I do not complicate the synthesis by additionally distinguishing between the two variants of cross-section comparisons. The differentiation between both cross-section comparisons would call for the splitting up of the upper row of Table 5.1.
9. Bartolini (1993) recommends a similar design in the course of his discussion of methods for the improvement of developmental theory.
10. Lijphart (1971) makes this recommendation by stating that one should increase the number of cases 'as much as possible'.
11. The arguments that I make in this section extend to case studies that rely on fuzzy sets (see Ragin 2000).
12. Mahoney (1999, 2003a) gives one of the few laudable treatments of nonbinary causes and outcomes in comparative case studies. He refers to the two measurement approaches as nominal and ordinal comparison, the former referring to binary causes and outcomes and the latter to multicategorical ones. This distinction is misleading because it mixes the measurement level – nominal, ordinal, interval, and metric – with the level of measurement aggregation. Note that Mahoney (2003a, 338) uses the term 'level of aggregation' for distinguishing between what I call the cross-case and within-case level of analysis.
13. Some phenomena (e.g., membership in an international organization) are inherently binary. Other phenomena (e.g., marital status) are inherently multicategorical. All other phenomena can be measured continuously, but not necessarily. One may prefer to calibrate continuous variables, such as growth, so as to obtain a binary or multicategorical cause.
14. Elman (2005, 302) refers to a change in the level of measurement aggregation as rescaling.
15. The variable 'capitalist system' can be considered ordinal, taking the degree of market orientation as the underlying criterion.
16. There are limits to this strategy because the more categories one distinguishes on the outcome (and the causes), the more cases one needs for a thorough assessment of additive causation.
17. The insight that social science theory is weak is not new (Bohrnstedt 1980, 785–6). See Leavitt, Mitchell, and Peterson (2010), Lieberson and Horwich (2008), and Achen (2002) for general pleas for better theory.
18. Elman (2005, 305–6) calls this logical compression.

19. Hypothesis-building and hypothesis-modifying case studies do not draw on established bodies of research. Consequently, elaborated theory can come into play only after the empirical analysis if the case study is centered on one of the two research goals.
20. This example is taken from Dür (2007a, 191).
21. There is no need to explicitly think about equifinality. If one does not declare a cause to be necessary, one implicitly argues that multiple causes can influence the outcome.
22. All causes are of some interest because, were they not, they would not be included in the analysis. Because of the original logic of the MoA and MoD, however, one can say that the invariant cause (MoA) and varying cause (MoD) are at the focus of both comparisons.
23. Smelser (1973, 44–5) distinguishes between variables and parameters in his discussion of cross-case comparisons (see also Lijphart 1971, 687). The variables vary and are at the heart of the empirical analysis, while parameters are held constant, the implication being that they are included in the comparison. In contrast to scope conditions, parameters thus are part of the inferential problem and not of the solution because they tend to increase the number of viable causal inferences.
24. This is the case unless one explicitly aims to test whether a scope condition can be relaxed (George and Bennett 2005, 25). In this instance, the boundary condition is treated as a necessary condition for a causal relationship to hold. In all other instances, one is not interested in what is going on outside of the population (Walker and Cohen 1985).
25. A high trade volume could be a scope condition, as well, with different consequences for the composition of the population.
26. This point relates to Caramani's (2010) argument that one will not be able to establish variance on a cause when there is a uniform trend in the international system that affects all countries of interest (in the case of country research).
27. It is assumed that the presence of a debate, a majoritarian electoral system, a single-party government, and unitarism are hypothesized to be individually necessary.
28. It is assumed that the presence of a debate, a majoritarian electoral system, a single-party government, and unitarism are hypothesized to be individually sufficient.

6 Process Tracing: Theory, Temporality, and Method

1. See Hall (2003) for an intriguing discussion of problems in aligning theory and method.
2. Beach and Pedersen (2012) make such a claim in relation with causal mechanisms.
3. The distinction between a narrow and broad conception of pattern matching is tied to theorizing about causal mechanisms and processes. In the narrow perspective, a specific sequence of steps is closely linked to causal inference about a mechanism. When the process is not realized in the same way as it was theorized, the corresponding hypothesis about the mechanism is found disconfirmed (Beach and Pedersen, 2012; George and Bennett 2005, chap. 10). In the broad version of pattern matching, inferences about mechanisms are independent of the empirical sequence of steps. This shows that while the focus of this chapter is on processes, arguments about processes and mechanisms are intertwined if one adopts the narrow view on pattern matching.

4. One could also look at noncrisis situations, but it is easier and more plausible to look at crises because the purported cause of democratic peace then becomes more visible.
5. I acknowledge that there is some reluctance to use causal vocabulary and to speak of patterns when talking about ideas, identities, and so on (Kurki 2007). But whether ideas can give rise to regular causal relationships or not is an ontological assumption that is, as many others, unresolvable (Hay 2006).
6. Moreover, one should not derive the leader's motivation for his or her decision from the subsequent empirical development. As has been aptly described elsewhere, it is risky to assume revealed preferences (Frieden 1999; Pierson 2000); that is, that the empirical events that ensue from an actor's choice fully reveal the reasons for making the decision.
7. This argument extends to case studies where the macro outcome is the result of aggregated micro behavior; for example, the vote share of a party is the consequence of a multitude of individual voting decisions. Before casting a vote, each individual weighs the consequences of voting for a different party and makes a voting decision according to the expected consequences of the vote.
8. Decision-making processes are more complex, and one can think of more specific steps, such as 'first meeting of government', but complexity and specificity of steps is not the issue here.
9. If one wants to emphasize that the process is central for causal inference, one should speak of systematic process analysis (Hall 2008).
10. The example would also work when identity or ideas are the mechanism.
11. In the following chapter, I question the goal and possibility of singling out one explanation more generally.
12. Goertz (2003) brings up the possibility that a process includes intervening steps where each step is necessary for the following step. Although possible in principle, it is demanding to specify a sequence where the absence of a single step could block the occurrence of the following steps.
13. The population additionally includes relevant negative cases.
14. One might argue that without the liberalization in the 1980s, there would not have been a financial crisis at all. Without going into the details of the crisis, it is important to note that multiple factors worked together to produce the mortgage crisis that preceded the financial crisis, including a low-interest rate policy of the US Federal Reserve and a US federal government that encouraged banks to give credits to home buyers. It is not easy to say whether without liberalization there would have been no crisis. Besides, less liberalization might have led to a crisis as well and also to less influential lobbying actors, therefore paving the way for large-scale change in the finance sector.

7 From Evidence to Inference: Use of Sources and Counterfactuals

1. Campbell (1975) makes this argument without invoking the terminology of CPOs and DPOs as these terms were invented 30 years later.
2. Whether one is theorizing realized or anticipated processes does not matter.
3. If it is argued that the outcome occurs in the observed and counterfactual case, the inferences that follow depend on the design at hand. The occurrences could

indicate that the outcome is due to equifinality, the implication being either that the manipulated cause is relevant or that the cause is truly irrelevant.

8 Frequentist and Bayesian Causal Inference in Tests of Hypotheses

1. See, for example, Bennett (2010), Checkel (2008), Mahoney (2003a, 2007), McAdam et al. (2008), Falleti (2009), and Tilly (2001, 2004).
2. One could also think of the two dimensions as continuous, though it is not immediately apparent what a continuous measure for uniqueness could be. In any case, this point is unrelated to the arguments that follow.
3. Van Evera (1997, 30–2) introduces the tests without reference to a particular level of analysis.
4. The hypothesis is of the cross-case type. As noted before, all arguments easily extend to within-case hypotheses on mechanisms and processes.
5. The argument extends to correlational hypotheses because they can be formulated as 'Only X correlates with Y'.
6. I presume that spending either decreases or increases and does not remain at exactly the same level.
7. Abell (2009a, 2009b, 2004) relies on Bayesian tools in his elaboration of Bayesian narratives. Simplifying somewhat, Bayesian narratives are presented as an alternative to causal inference with the 'orthodox statistical model' (Abell 2009b, 561) and as a tool for generating causal inferences in the analysis of unique phenomena. This means that Bayesian narratives are explicitly separated from an interest in general causation, which is possible because Bayesian tools can be applied to single cases.
8. One aspect I do not focus on here is that one needs to determine for each observable implication a prior and the likelihood of finding it confirmed. On the cross-case level, this is not a problem because one hypothesis entails only one prediction (a democratic dyad coincides with peace; a low parliamentary seat share coincides with a low cabinet share, and so on). Matters become more complicated for two reasons. First, the observable implications are located on the within-case level and need to be aggregated to make a single inference as to whether a mechanism or process is in place. Second, one can theorize auxiliary outcomes that should be found confirmed if the hypothesis of main interest is correct. In either scenario, this can be done by dealing with each implication singly or all at the same time. Since this is an issue that requires some technical elaboration, I leave it aside here in order to concentrate on the distinction between frequentism and Bayesianism. The arguments made in the main text also apply when one allows multiple observable implications to have different conditional likelihoods of being found confirmed.
9. Without loss of generality, I limit the discussion to a within-case hypothesis that underpins a pattern of sufficiency.
10. At the margin, additional CPOs lose value, and there is little to gain from gathering more. If 200 politicians confirm that spending is high in order to compensate the losers from globalization, an additional interview is of little value. If only two politicians were interviewed, one more interview has more to offer and should be done.

11. Bayes has formulated multiple theorems. A general reference to Bayes' theorem usually refers to the formula presented in this text.
12. Bayes' theorem can be written in different ways (Howson and Urbach 2005, 20–2). This one is most appropriate for the arguments I make in the following. As explained in Chapter 3, one can take background knowledge – *B* – into the formula by writing, for example, *p(E|C & B & case)* instead of *p(E|C)*. What background knowledge is depends on the research question and hypothesis under scrutiny. For instance, background knowledge could be that the availability of sources is good so that the source coverage problem and source coverage bias should be small.
13. The conditional probabilities *p(E|C)* and *p(~E|C)*, and *p(E|~C)* and *p(~E|~C)* should add up to one per column, implying that the specification of three parameters – one marginal probability of a column and one conditional probability per column – allows it to complete the table and invoke Bayes theorem. Three parameters therefore also permit it to determine all four possible posteriors.
14. The denominator of Bayes' theorem is the same for the working hypothesis and the null hypothesis and drops out of the equation.
15. The relation between the likelihood ratio and Bayes factor is discussed below.
16. If the conditional likelihoods are not identical, the ratio of posteriors depends on the likelihood ratio weighted by the ratio of the prior probabilities.
17. The table is adopted from Wagenmakers et al. (2011). Their labels differ slightly from the original ones proposed by Jeffreys (1961). Wagenmakers refer to a ratio of one as 'no evidence'. I prefer inconclusive evidence because one has gathered evidence, but it equally supports the working and the null hypothesis.
18. Again, I assume that constant levels of spending are not a third option.

9 External Validity and Generalization: Challenges and Strategies

1. Blatter and Blume (2010, 336) distinguish statistical generalization, contingent generalization, and abstraction. They prefer abstraction to the two other modes of generalization. In my reading, however, abstraction it is not about external validity but internal validity. Abstraction is characterized by subsuming observations under theories and evaluating the latter in light of the 'likeliness' of the observations. This is what I discussed in Chapter 8 under the rubric of Bayesian case studies. Moreover, Blatter and Blume (2010, 341–8) distinguish between horizontal generalization, subsuming contingent and statistical generalization, and vertical generalization (which is equivalent to abstraction). The notion of contingent generalization, which means to generalize insights from cases to configurations of conditions, seems to ignore the relationship between the intension of a concept and its extension (Sartori 1970). By making inferences about configurations of conditions, each of which is a manifestation of a properly specified concept, one implicitly commits (horizontal) generalization because the concept has a certain extension and encompasses more cases than those one did examine (unless the entire population is analyzed empirically). Statistical generalization is the mode of generalization that is discussed in this section although I do not use this term. As will become clear in this chapter, there is nothing in the generalization of causal inferences that would qualify as 'statistical', rendering this term more confusing than illuminating.

2. There are tools for estimating the statistical significance for a small number of cases (Mahoney 2003a, 350–1). However, these are techniques for estimating the likelihood that the causal relationship is systematic for the cases under analysis (which is disputed if the cases constitute a population; see Berk et al. 1995). This is different from significance testing for generalization from a sample to a population. Furthermore, it is sometimes argued that cross-case determinism is compatible with the notion of a probabilistic world. The claim is that one can have invariant patterns on the cross-case level, while the behavior of individuals that produce the cross-case relationship is probabilistic. Suppose that you are interested in the occurrence of famines and the outbreak of revolutions. The simple line of reasoning goes that a famine makes people desperate and that desperation leads to an individual decision to revolt against the political elite. In order for a famine to result in a revolution, it is not necessary that every citizen in a country be desperate and that every desperate person rebels against the elite. The outbreak of a revolution requires only a critical mass of people to behave as theorized. We thus have a stable cross-case pattern, although the effects are probabilistic on the within-case level. Though appealing, this line of reasoning does not seem to be promising at second glance. The point is that, *ex post*, we can always say that the critical mass was reached when we observe the outcome and not reached when we do not observe it. But then we are engaging in circular reasoning and inferring the causal mechanism from the outcome in order to explain the outcome. Unless one has a very strong theory that predicts specific probabilities and thresholds for every intervening step – how many people are needed for a revolt, how many people become desperate in response to a famine, etc. – cross-case determinism cannot be credibly explained with within-case probabilism.
3. Testing the need for specific scope conditions is sometimes presented as an advantage of case studies (George and Bennett 2005, 27). However, this strategy has not been elaborated in detail so far. In terms of the three research goals discussed in Chapter 1, this is a hypothesis-testing case study because one tests the hypothesis that a scope condition can be skipped.
4. Layered generalization is different from the building-block approach proposed by George and Bennett (2005, 78). Their building-block procedure is specifically tied to the idea of typological theory (chap. 11) and denotes the analysis of a single type (block) one after another. The cumulative analysis of types is not related to generalization because causal inferences are not extended from one type to another (112).
5. I say 'temporarily' because until the entire layer of new cases has been examined, the strategy of uneven layered generalization produces only intermediate steps toward the analysis of an entire layer. If it turns out that only some of the cases in the new layer confirm a hypothesis, one has to make theoretical sense of these results and try to formulate theoretically intelligible scope conditions instead of creating an oddly shaped population in an ad-hoc fashion.
6. In their comparison of a quantitative and qualitative research culture, Mahoney and Goertz (2006, 237–8) make this argument for case studies in general, that is, regardless of whether they are anchored in CHA or not. Pending a review of published case studies that shows how scope conditions are handled in practice, I share Levy's (2007, 203–4) skepticism that, *in general,* qualitative researchers are more attentive to scope conditions than are quantitative researchers.

7. Goldstone (2003) argues that one should not make any homogeneity assumption at all. Instead, the comparability of cases should be made subject to an empirical analysis. In practice, this might not always be feasible and essentially eliminates any problem of generalization by refusing to generalize inferences in the first place.

References

A. Abbott (1998). ‘The Causal Devolution’, *Sociological Methods & Research* 27:2, 148–81.

P. Abell (2009a). ‘A Case for Cases: Comparative Narratives in Sociological Explanation’, *Sociological Methods & Research* 38:1, 38–70.

—— (2001). ‘Causality and Low-Frequency Complex Events – the Role of Comparative Narratives’, *Sociological Methods & Research* 30:1, 57–80.

—— (2009b). ‘History, Case Studies, Statistics, and Causal Inference’, *European Sociological Review* 25:5, 561–7.

—— (2004). ‘Narrative Explanation: An Alternative to Variable-Centered Explanation?’, *Annual Review of Sociology* 30, 287–310.

C. H. Achen (2002). ‘Toward a New Political Methodology: Microfoundations and ART’, *Annual Review of Political Science* 5:1, 423–50.

—— (2005). ‘Two Cheers for Charles Ragin’, *Studies in Comparative International Development* 40:1, 27–32.

R. Adcock and D. Collier (2001). ‘Measurement Validity: A Shared Standard for Qualitative and Quantitative Research’, *American Political Science Review* 95:3, 529–46.

M. Alvesson and D. Karreman (2007). ‘Constructing Mystery: Empirical Matters in Theory Development’, *Academy of Management Review* 32:4, 1265–81.

C. Anckar (2008). ‘On the Applicability of the Most Similar Systems Design and the Most Different Systems Design in Comparative Research’, *International Journal of Social Research Methodology* 11:5, 380–401.

H. Bäck and P. Dumont (2007). ‘Combining Large-*N* and Small-*N* strategies: The Way Forward in Coalition Research’, *West European Politics* 30:3, 467–501.

T. Bartelborth (2004). ‘Wofür Sprechen die Daten?’ *Journal for General Philosophy of Science* 35:1, 13–40.

S. Bartolini (1993). ‘On Time and Comparative Research’, *Journal of Theoretical Politics* 5:2, 131–67.

A. H. Barton (1955). ‘The Concept of Property Space in Social Research’, in P. F. Lazarsfeld and M. Rosenberg (eds), *The Language of Social Research*. New York: Free Press, 40–53.

R. H. Bates (2007). ‘From Case Studies to Social Science: A Strategy for Political Research’, in C. Boix and S. C. Stokes (eds), *The Oxford Handbook of Comparative Politics*. Oxford: Oxford University Press, 172–85.

D. Beach and R. B. Pedersen (2012). *Process Tracing Methods*. Ann Arbor: University of Michigan Press.

N. Beck (2010). ‘Causal Process “Observation”: Oxymoron or (Fine) Old Wine’, *Political Analysis* 18:4, 499–505.

—— (2006). ‘Is Causal-Process Observation an Oxymoron?’ *Political Analysis* 14:3, 347–52.

N. Beck and S. Jackman (1998). ‘Beyond Linearity by Default: Generalized Additive Models’, *American Journal of Political Science* 42:2, 596–627.

A. Bennett (2005). ‘The Fallacy of Fallacies’, *Qualitative Methods – Newsletter of the APSA Organized Section on Qualitative Methods* 3:1, 5–9.

—— (2008). ‘Process-Tracing: A Bayesian Perspective’, in J. M. Box-Steffensmeier, H. Brady and D. Collier (eds), *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, 702–21.
—— (2010). ‘Process Tracing and Causal Inference’, in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Md.: Rowman & Littlefield, 207–19.
—— (2006). ‘Stirring the Frequentist Pot with a Dash of Bayes’, *Political Analysis* 14:3, 339–44.
A. Bennett and C. Elman (2006)‘Qualitative Research: Recent Developments in Case Study Methods’, *Annual Review of Political Science* 9, 455–76.
D. Berg-Schlosser and G. De Meur (1994). ‘Conditions of Democracy in Interwar Europe: A Boolean Test of Major Hypotheses’, *Comparative Politics* 26:3, 253–79.
D. Berg-Schlosser and S. Quenter (1996). ‘Macro-quantitative and Macro-qualitative Methods in Political Science: Advantages and Deficits of Comparative Procedures Using the Example of Theories of the Welfare State’, *Politische Vierteljahresschrift* 37:1, 100–17.
R. A. Berk, Bruce Western, and Richard E. Weiss (1995). ‘Statistical Inference for Apparent Populations’, *Sociological Methodology* 25, 421–58.
J. N. Bhagwati (2002). ‘Introduction: The Unilateral Freeing of Trade versus Reciprocity’, in J. N. Bhagwati (ed.), *Going Alone: The Case for Relaxed Reciprocity in Freeing Trade*. Cambridge, Mass.: MIT Press, 1–30.
R. Bhaskar (1975). *A Realist Theory of Science*. New York: Routledge.
J. Blatter and T. Blume (2010). ‘In Search of Co-variance, Causal Mechanisms or Congruence? Towards a Plural Understanding of Case Studies’, *Swiss Political Science Review* 14, 315–56.
J. Blatter and M. Haverland (2012). *Designing Case Studies*. Basingstoke: Palgrave Macmillan.
J. Blatter, M. Kreutzer, M. Rentl, and J. Thiele (2010). ‘Preconditions for Foreign Activities of European Regions: Tracing Causal Configurations of Economic, Cultural, and Political Strategies’, *Publius* 40:1, 171–99.
G. W. Bohrnstedt (1980). ‘Social Science Methodology: The Past Twenty-Five Years’, *American Behavioral Scientist* 23:6, 781–7.
H. A. Brady (2008). ‘Causation and Explanation in Social Science’, in J. M. Box-Steffensmeier, H. Brady, and D. Collier (eds), *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, 217–70.
H. E. Brady and D. Collier (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (1st edn). Lanham, Md.: Rowman & Littlefield.
H. E. Brady, D. Collier, and J. Seawright (2004). ‘Refocusing the Discussion of Methodology’, in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Md.: Rowman & Littlefield, 3–20.
—— (2006). ‘Toward a Pluralistic Vision of Methodology’, *Political Analysis* 14:3, 353–68.
B. F. Braumoeller and G. Goertz (2000). ‘The Methodology of Necessary Conditions’, *American Journal of Political Science* 44:4, 844–58.
B. Bueno de Mesquita (2003). *The Logic of Political Survival*. Cambridge, Mass.: MIT Press.
V. Bunce (2000). ‘Comparative Democratization: Big and Bounded Generalizations’, *Comparative Political Studies* 33:6–7, 703–34.
M. Bunge (1997). ‘Mechanism and Explanation’, *Philosophy of the Social Sciences* 27:4, 410–65.

M. Bunzl (2004). 'Counterfactual History: A User's Guide', *American Historical Review* 109:3, 845–58.

M. Burawoy (1989). 'Two Methods in Search of Science: Skocpol versus Trotsky', *Theory and Society* 18:6, 759–805.

S. S. Bush (2011). 'International Politics and the Spread of Quotas for Women in Legislatures', *International Organization* 65:1, 103–37.

T. Büthe (2002). 'Taking Temporality Seriously: Modeling History and the Use of Narratives as Evidence', *American Political Science Review* 96:3, 481–93.

D. T. Campbell (1975). 'Degrees of Freedom and the Case Study', *Comparative Political Studies* 8:2, 178–93.

D. T. Campbell and H. L. Ross (1968). 'The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis', *Law & Society Review* 3:1, 33–53.

G. C. Capoccia, and D. R. Kelemen (2007). 'The Study of Critical Junctures: Theory, Narrative, and Counterfactuals in Historical Institutionalism', *World Politics* 59:3, 341–69.

D. Caramani (2010). 'Of Differences and Similarities: Is the Explanation of Variation a Limitation to (or of) Comparative Analysis?' *European Political Science* 9:1, 34–48.

N. Cartwright (2004). 'From Causation to Explanation and Back', in B. Leiter (ed.), *The Future for Philosophy*. Oxford: Oxford University Press, 230–45.

J. L. Casti (1989). *Paradigms Lost: Images of Man in the Mirror of Science*. New York: Morrow.

A. Chatterjee (2009). 'Ontology, Epistemology, and Multiple Methods', APSA 2009 Toronto Meeting Paper, http://ssrn.com/paper=1451632.

J. T. Checkel (2008). 'Process Tracing', in A. Klotz and D. Prakash (eds), *Qualitative Methods in International Relations: A Pluralist Guide*. Houndmills: Palgrave Macmillan, 114–27.

M. R. Cohen and E. Nagel (1934). *An Introduction to Logic and Scientific Method*. New York: Harcourt.

J. S. Coleman (1998). *Foundations of Social Theory*. Cambridge, Mass.: Harvard University Press.

D. Collier (1993). 'The Comparative Method', in A. Finifter (ed.), *Political Science: The State of the Discipline II*. Washington D.C.: American Political Science Association, 105–19.

—— (2010). 'Process Tracing: Introduction and Exercises', http://polisci.berkeley.edu/people/faculty/CollierD/Proc%20Trac%20-%20Text%20and%20Story%20-%20Sept%2024.pdf (accessed July 2, 2012).

D. Collier (2011). 'Understanding Process Tracing', *Political Science & Politics* 44:4, 823–30.

D. Collier, H. E. Brady, and J. Seawright (2004a). 'Critiques, Responses, and Trade-Offs: Drawing Together the Debate', in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Md.: Rowman & Littlefield, 195–227.

—— (2010). 'Outdated Views of Qualitative Methods: Time to Move On', *Political Analysis* 18:4, 506–13.

—— (2004b). 'Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology', in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham, Md.: Rowman & Littlefield, 229–66.

D. Collier and J. Mahoney (1996). 'Insights and Pitfalls: Selection Bias in Qualitative Research', *World Politics* 49:1, 56–91.

D. Collier, J. Seawright, and G. L. Munck (2004c). 'The Quest for Standards', in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham. Md.: Rowman & Littlefield, 21–50.

R. B. Collier and D. Collier (1991). *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America*. Princeton, N.J.: Princeton University Press.

J. Constantelos (2010). 'Playing the Field: Federalism and the Politics of Venue Shopping in the United States and Canada', *Publius: The Journal of Federalism* 40:3, 460–83.

J. A. C. Conybeare (1987). *Trade Wars: The Theory and Practice of International Commercial Rivalry*. New York: Columbia University Press.

T. D. Cook and V. C. Wong (2008). 'Better Quasi-experimental Practice', in P. Alasuutari, L. Bickman, and J. Brannen (eds), *The SAGE Handbook of Social Research Methods*. Thousand Oaks, Calif.: Sage, 134–65.

Craver, C. (2006). 'When Mechanismic Models Explain', *Synthese* 153:3, 355–76.

L. Cronqvist and D. Berg-Schlosser (2008). 'Multi-value QCA (mvQCA)', in B. Rihoux and C. Ragin (eds), *Configurational Comparative Methods*. Thousand Oaks, Calif.: Sage, 69–86.

E. G. DeFelice (1986). 'Causal Inference and Comparative Methods', *Comparative Political Studies* 19:3, 415–37.

—— (1980). 'Comparison Misconceived – Common Nonsense in Comparative Politics', *Comparative Politics* 13:1, 119–26.

C. Demetriou (2009). 'The Realist Approach to Explanatory Mechanisms in Social Science: More than a Heuristic?' *Philosophy of the Social Sciences* 39:3, 440–62.

G. De Meur, and D. Berg-Schlosser (1994). 'Comparing Political Systems: Establishing Similarities and Dissimilarities', *European Journal of Political Research* 26, 193–219.

—— (1996). 'Conditions of Authoritarianism, Fascism, and Democracy in Interwar Europe: Systematic Matching and Contrasting of Cases for "Small N" Analysis', *Comparative Political Studies* 29:4, 423–68.

D. Dessler (1991). 'Beyond Correlations: Toward a Causal Theory of War', *International Studies Quarterly* 35:3, 337–55.

D. de Vaus (2008). 'Comparative and Cross-National Designs', in P. Alasuutari, L. Bickman, and J. Brannen (eds), *The SAGE Handbook of Social Research Methods*. Los Angeles: SAGE, 249–64.

D. Dion (1998). 'Evidence and Inference in the Comparative Case Study', *Comparative Politics* 30:2, 127–45.

P. Downward, J. H. Finch, and J. Ramsay (2002). 'Critical Realism, Empirical Methods and Inference: A Critical Discussion', *Cambridge Journal of Economics* 26:4, 481–500.

D. W. Drezner (2000). 'Bargaining, Enforcement, and Multilateral Sanctions: When Is Cooperation Counterproductive?' *International Organization* 54:1, 73–102.

T. Dunning (2008). 'Improving Causal Inference: Strengths and Limitations of Natural Experiments', *Political Research Quarterly* 61:2, 282–93.

——(2007) 'The Role of Iteration in Multi-method Research', *APSA Qualitative Methods Newsletter* 5:1, 22–4.

A. Dür (2007a). 'Discriminating among Rival Explanations: Some Tools for Small-N Researchers', in T. Gschwend and F. Schimmelfennig (eds.), *Research Design in Political Science: How to Practice What They Preach*. Houndmills: Palgrave Macmillan, 180–200.

—— (2007b). 'Foreign Discrimination, Protection for Exporters, and U.S. Trade Liberalization', *International Studies Quarterly* 51:2, 457–80.

—— (2010). *Protection for Exporters: Power and Discrimination in Transatlantic Trade Relations, 1930–2010*. Ithaca, N.Y.: Cornell University Press.

B. Ebbinghaus (2005). 'When Less Is More: Selection Problems in Large-N and Small-N Cross-National Comparisons', *International Sociology* 20:2, 133–52.

S. Eckert (2010). 'Between Commitment and Control: Varieties of Delegation in the European Postal Sector', *Journal of European Public Policy* 17:8, 1231–52.

H. Eckstein (1975). 'Case Study and Theory in Political Science', in F. I. Greenstein and N. W. Polsby (eds), *Strategies of Inquiry. Handbook of Political Science*, vol. 7. Reading, Mass.: Addison-Wesley, 79–137.

K. M. Eisenhardt (1991). 'Better Stories and Better Constructs: The Case for Rigor and Comparative Logic', *The Academy of Management Review* 16:3, 620–7.

—— (1989). 'Building Theories from Case Study Research', *The Academy of Management Review* 14:4, 532–50.

C. Elman (2005). 'Explanatory Typologies in Qualitative Studies of International Politics', *International Organization* 59:2, 293–326.

C. Elman and M. F. Elman (2001). *Bridges and Boundaries: Historians, Political Scientists, and the Study of International Relations*. Cambridge, Mass.: MIT Press.

T. Falleti and J. Lynch (2008). 'From Process to Mechanism: Varieties of Disaggregation', *Qualitative Sociology* 31:4, 333–9.

T. G. Falleti, and J. F. Lynch (2009). 'Context and Causal Mechanisms in Political Analysis', *Comparative Political Studies* 42:9, 1143–66.

J. D. Fearon and D. Laitin (2008). 'Integrating Qualitative and Quantitative Methods', in J. M. Box-Steffensmeier, H. Brady, and D. Collier (eds) *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, 756–76.

J. A. Ferejohn (2004). 'External and Internal Explanation', in I. Shapiro, R. M. Smith, and T. E. Masoud (eds), *Problems and Methods in the Study of Politics*. Cambridge: Cambridge University Press, 144–64.

B. Flyvbjerg (2006). 'Five Misunderstandings about Case-Study Research', *Qualitative Inquiry* 12:2, 219–45.

—— (2001). *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again*. Cambridge: Cambridge University Press.

R. J. Franzese (2008). 'Context Matters: The Challenge of Multicausality, Context-Conditionality, and Endogeneity for Empirical Evaluation of Positive Theory in Comparative Politics', in C. Boix and D. E. Stokes (eds), *Oxford Handbook of Comparative Politics*, Oxford: Oxford University Press, 28–72.

D. A. Freedman (1999). 'From Association to Causation: Some Remarks on the History of Statistics', *Statistical Science* 14:3, 243–58.

—— (1991). 'Statistical Models and Shoe Leather', *Sociological Methodology* 21, 291–313.

M. Freitag and R. Schlicht (2009). 'Educational Federalism in Germany: Foundations of Social Inequality in Education', *Governance* 22:1, 47–72.

J. Frendreis (1983). 'Explanation of Variation and Detection of Covariation: The Purpose and Logic of Comparative Analysis', *Comparative Political Studies* 16, 255–72.

J. A. Frieden (1999). 'Actors and Preferences in International Relations', in D. A. Lake and R. Powell (eds), *Strategic Choice and International Relations*, Princeton, N.J.: Princeton University Press, 39–76.

J. Friedrichs and F. Kratochwil (2009). 'On Acting and Knowing: How Pragmatism Can Advance International Relations Research and Methodology', *International Organization* 63:04, 701–31.

W. A. Gamson (1961). 'A Theory of Coalition Formation', *American Political Science Review* 26, 373–82.

B. Geddes (1990). 'How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics', *Political Analysis* 2:1, 131–50.
—— (2003). *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. Ann Arbor: University of Michigan Press.
A. L. George (1979). 'Case Studies and Theory Development: The Method of Structured, Focused Comparison', in P. G. Lauren (ed.), *Diplomacy. New Approaches in History, Theory, and Policy*. New York: Free Press; London: Collier Macmillan, 43–68.
A. L. George and A. Bennett (2005). *Case Studies and Theory Development in the Social Sciences*. Cambridge, Mass.: MIT Press.
J. Gerring (2007a). *The Case Study Method: Principles and Practices*. Cambridge: Cambridge University Press.
—— (2010). 'Causal Mechanisms: Yes, but…', *Comparative Political Studies* 43:11, 1499–526.
—— (2005). 'Causation: A Unified Framework for the Social Sciences', *Journal of Theoretical Politics* 17:2, 163–98.
—— (2007b). 'Is There a (Viable) Crucial-Case Method?' *Comparative Political Studies* 40:3, 231–53.
—— (2008). 'The Mechanismic Worldview: Thinking Inside the Box', *British Journal of Political Science* 38, 161–79.
—— (2001). *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.
—— (2004). 'What Is a Case Study and What Is It Good For?' *American Political Science Review* 98:2, 341–54.
J. Gerring and R. McDermott (2007). 'An Experimental Template for Case Study Research', *American Journal of Political Science* 51:3, 688–701.
G. Goertz (2003). 'Cause, Correlation, and Necessary Conditions', in G. Goertz and H. Starr (eds), *Necessary Conditions. Theory, Methodology, and Applications*. Lanham, Md.: Rowman & Littlefield, 47–65.
—— (2006). *Social Science Concepts: A User's Guide*. Princeton, N.J.: Princeton University Press.
G. Goertz and J. S. Levy (2007). 'Causal Explanation, Necessary Conditions, and Case Studies', in J. S. Levy and G. Goertz (eds), *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*. London: Routledge, 9–45.
G. Goertz and J. Mahoney (2012). *A Tale of Two Cultures: Contrasting Qualitative and Quantitative Paradigms*. Princeton, N.J.: Princeton University Press.
G. Goertz and H. Starr (2003a). 'Introduction: Necessary Condition Logics, Research Design, and Theory', in G. Goertz and H. Starr (eds), *Necessary Conditions. Theory, Methodology, and Applications*. Lanham, Md.: Rowman & Littlefield, 1–24.
—— (2003b). *Necessary Conditions. Theory, Methodology, and Applications*. Lanham, Md.: Rowman & Littlefield.
J. A. Goldstone (2003). 'Comparative-Historical Analysis and Knowledge Accumulation in the Study of Revolutions', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press.
—— (1998). 'Initial Conditions, General Laws, Path Dependence and Explanation in Historical Sociology', *American Journal of Sociology* 104:3, 829–45.
—— (1997). 'Methodological Issues in Comparative Macrosociology', *Comparative Social Research* 16, 107–20.
J. H. Goldthorpe (1997a). 'Current Issues in Comparative Macrosociology: A Debate on Methodological Issues', *Comparative Social Research* 16, 1–26.

—— (2000). *On Sociology. Numbers, Narratives, and the Integration of Research and Theory*. Oxford: Oxford University Press.

—— (1997b). 'A Response to the Commentaries', *Comparative Social Research* 16, 121–32.

J. Gowa (2011). 'The Democratic Peace after the Cold War', *Economics & Politics* 23:2, 153–71.

J. M. Grieco (1990). *Cooperation among Nations: Europe, America, and Non-tariff Barriers to Trade*. Ithaca, N.Y.: Cornell University Press.

B. Grofman (2001). *Political Science as Puzzle Solving*. Ann Arbor: University of Michigan Press.

B. Grofman and C. Q. Schneider (2009). 'An Introduction to Crisp-Set QCA, with a Comparison to Binary Logistic Regression', *Political Research Quarterly* 62:4, 662–72.

A. Grzymala-Busse (2011). 'Time Will Tell? Temporality and the Analysis of Causal Mechanisms and Processes', *Comparative Political Studies* 44:9, 1267–97.

S. Haggard (1988). 'The Institutional Foundations of Hegemony: Explaining the Reciprocal Trade Agreements Act of 1934', *International Organization* 42:1, 91–119.

S. Haggard and R. R. Kaufman (2008). *Development, Democracy, and Welfare States: Latin America, East Asia, and Eastern Europe*. Princeton, N.J.: Princeton University Press.

P. A. Hall (2003). 'Aligning Ontology and Methodology in Comparative Research', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 373–404.

—— (2008). 'Systematic Process Analysis: When and How to Use It', *European Political Science* 7:3, 304–17.

G. Hawthorn (1991). *Plausible Worlds: Possibility and Understanding in History and the Social Sciences*. Cambridge: Cambridge University Press.

C. Hay (2006). 'Political Ontology', in R. E. Goodin and C. Tilly (eds), *The Oxford Handbook of Contextual Political Analysis*. Oxford: Oxford University Press, 78–96.

P. Hedström and R. Swedberg (1996). 'Social Mechanisms', *Acta Sociologica* 39:3, 281–308.

—— (1998). 'Social Mechanisms: An Introductory Essay', in P. Hedström and R. Swedberg (eds), *Social Mechanisms. An Analytical Approch to Social Theory*. Cambridge: Cambridge University Press, 1–31.

P. Hedström and P. Ylikoski (2010). 'Causal Mechanisms in the Social Sciences', *Annual Review of Sociology* 36:1, 49–67.

C. G. Hempel (1965). *Aspects of Scientific Explanation*. New York: Free Press.

F. Hendriks and A. Michels (2011). 'Democracy Transformed? Reforms in Britain and the Netherlands (1990–2010)', *International Journal of Public Administration* 34:5, 307–17.

C. R. Hitchcock (1995). 'The Mishap at Reichenbach Fall: Singular vs. General Causation', *Philosophical Studies* 78:3, 257–91.

J. Hogan and D. Doyle (2009). 'A Comparative Framework: How Broadly Applicable Is a "Rigorous" Critical Junctures Framework?' *Acta Politica* 44:2, 211–40.

K. R. Howe (2011). 'Mixed Methods, Mixed Causes?' *Qualitative Inquiry* 17:2, 166–71.

C. Howson and P. Urbach (2005). *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court.

D. Hume (2003 [1740]). *A Treatise of Human Nature*. Mineola, N.Y.: Dover.

C. Ichniowski and K. Shaw (2011). 'Insider Econometrics: A Roadmap for Estimating Empirical Models of Organizational Design and Performance', in R. Gibbons and J. Roberts (eds), *Handbook of Organizational Economics*.

K. Imai, L. Keele, D. Tingley, and T. Yamamoto (2011). 'Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies', *American Political Science Review* 105:4, 765–89.

E. M. Immergut and K. M. Anderson (2008). 'Historical Institutionalism and West European Politics', *West European Politics* 31:1, 345–69.

P. T. Jackson (2010). *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*. London: Taylor & Francis.

A. M. Jacobs (2008). 'The Politics of When: Redistribution, Investment and Policy Making for the Long Term', *British Journal of Political Science* 38:3, 193–220.

M. L. F. Jakobsen (2010). 'Untangling the Impact of Europeanization and Globalization on National Utility Liberalization: A Systematic Process Analysis of Two Danish Reforms', *Journal of European Public Policy*, 17:6, 891–908.

H. Jeffreys (1961). *Theory of Probability*. Oxford: Clarendon Press.

J. Johnson (2006). 'Consequences of Positivism: A Pragmatist Assessment', *Comparative Political Studies* 39:2, 224–52.

J. B. Johnson, H. T. Reynolds, and J. D. Mycoff (2007). *Political Science Research Methods* (6th edn). Washington, D.C.: CQ Press.

M. Kaeding (2008). 'Lost in Translation or Full Steam Ahead – The Transposition of EU Transport Directives across Member States', *European Union Politics* 9:1, 115–43.

A. Kaiser (1997). 'Types of Democracy: From Classical to New Institutionalism', *Journal of Theoretical Politics* 9:4, 419–44.

I. Katznelson (2003). 'Periodization and Preferences: Reflections on Purposive Action in Comparative Historical Social Science', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 270–301.

S. Kemp and J. Holmwood (2003). 'Realism, Regularity and Social Explanation', *Journal for the Theory of Social Behaviour* 33:2, 165–87.

D. Kerwer and M. Teutsch (2001). 'Elusive Europeanization: Liberalizing Road Haulage in the European Union', *Journal of European Public Policy* 8:1, 124–43.

G. King, R. O. Keohane, and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, N.J.: Princeton University Press.

G. King, and E. N. Powell (2008). 'How Not to Lie without Statistics'. Paper prepared for delivery at the 2008 Annual Meeting of the American Political Science Association.

E. Kiser and M. Hechter (1991). 'The Role of General Theory in Comparative-Historical Sociology', *American Journal of Sociology* 97, 1–30.

B. Kittel (2005). 'The American Political Methodology Debate: Where Is the Battlefield?' *Newsletter of the APSA Organized Section on Qualitative Methods* 3:1, 12–9.

—— (2006). 'A Crazy Methodology? On the Limits of Macro-Quantitative Social Science Research', *International Sociology* 21:5, 647–77.

A. Klotz (2008). 'Case Selection', in A. Klotz and D. Prakash (eds), *Qualitative Methods in International Relations: A Pluralist Guide*. Houndmills: Palgrave Macmillan, 43–58.

D. Y. Kono (2006). 'Optimal Obfuscation: Democracy and Trade Policy Transparency', *American Political Science Review* 100:3, 369–84.

D. Kühn and I. Rohlfing (2010). 'Causal Explanation and Multi-method Research in the Social Sciences'. IPSA Committee on Concepts and Methods, Working Paper Series Political Methodology, no. 26.

M. Kurki (2007). *Causation in International Relations: Reclaiming Causal Analysis*. Cambridge: Cambridge University Press.

R. Lane (1996). 'Positivism, Scientific Realism and Political Science: Recent Developments in the Philosophy of Science', *Journal of Theoretical Politics* 8:3, 361–82.

M. Lange (2009). *Lineages of Despotism and Development: British Colonialism and State Power*. Chicago: University of Chicago Press.

C. A. Lave and J. G. March (1975). *An Introduction to Models in the Social Sciences*. Lanham, Md.: University Press of America.

P. Lazarsfeld (1937). 'Some Remarks on the Typological Procedures in Social Research', *Zeitschrift für Sozialforschung* 6, 119–39.

K. Leavitt, T. R. Mitchell, and J. Peterson (2010). 'Theory Pruning: Strategies to Reduce Our Dense Theoretical Landscape', *Organizational Research Methods* 13:4, 644–67.

R. N. Lebow (2010). *Forbidden Fruit: Counterfactuals and International Relations*. Princeton, N.J.: Princeton University Press.

—— (2000). 'What's So Different about a Counterfactual?', *World Politics* 52, 350–85.

J. S. Levy (2008) 'Case Studies: Types, Designs, and Logics of Inference', *Conflict Management and Peace Science* 25:1, 1–18.

—— (2007) 'Qualitative Methods and Cross-method Dialogue in Political Science', *Comparative Political Studies* 40:2, 196–214.

E. S. Lieberman (2009) *Boundaries of Contagion: How Ethnic Politics Have Shaped Government Response to AIDS*. Princeton, N.J.: Princeton University Press.

—— (2001) 'Causal Inference in Historical Institutional Analysis: A Specification of Periodization Strategies', *Comparative Political Studies* 34:9, 1011–35.

—— (2005) 'Nested Analysis as a Mixed-Method Strategy for Comparative Research', *American Political Science Review* 99:3, 435–52.

S. Lieberson (1997). 'The Big Broad Issues in Society and Social History: Application of a Probabilistic Perspective', in V. R. McKim and S. P. Turner (eds), *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, Ind.: University of Notre Dame Press, 359–85.

—— (1998). 'Causal Analysis and Comparative Research: What Can We Learn from Studies Based on a Small Number of Cases?' in H.-P. Blossfeld and G. Prein (eds), *Rational Choice Theory and Large-Scale Data Analysis*. Boulder,Colo.: Westview, 129–45.

—— (1994). 'More on the Uneasy Case for Using Mill-Type Methods in Small-N Comparative Studies', *Social Forces* 72:4, 1225–37.

—— (1991). 'Small Ns and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases', *Social Forces* 70:2, 307–20.

S. Lieberson and J. Horwich (2008). 'Implication Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences', *Sociological Methodology* 38:1, 1–50.

A. Lijphart (1975). 'Comparable-Cases Strategy in Comparative Research', *Comparative Political Studies* 8:2, 158–77.

—— (1971). 'Comparative Politics and the Comparative Method', *American Political Science Review* 65:3, 682–93.

—— (1999). *Patterns of Democracy: Government Forms and Performance in Thirty-six Countries*. New Haven, Conn.: Yale University Press.

P. Lipton (1991). *Inference to the Best Explanation*. London: Routledge.

D. Little (2010). 'Causal Mechanisms', in *New Contributions to the Philosophy of History*. Dordrecht: Springer, 97–120.

—— (1991). 'Causal Analysis' (chap. 2), 'Rational Choice Theory' (chap. 3), 'Toward Methodological Pluralism' (chap. 11), in D. Little (ed.), *Varieties of Social Explanation: An Introduction to the Philosophy of Social Science*. Boulder, Colo.: Westview Press.

—— (1998). *Microfoundations, Method, and Causation: On the Philosophy of Social Scienes*. New Brunswick, N.J.: Transaction.

I. S. Lustick (1996). 'History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias', *American Political Science Review* 90:3, 605–18.

P. Machamer, L. Darden, and C. F. Craver (2000). 'Thinking about Mechanisms', *Philosophy of Science* 67:1, 1–25.

J. L. Mackie (1965). 'Causes and Conditions', *American Philosophical Quarterly* 2, 245–64.

J. Mahoney (2010). 'After KKV: The New Methodology of Qualitative Research', *World Politics* 62:1, 120–47.

—— (2001). 'Beyond Correlational Analysis: Recent Innovations in Theory and Method', *Sociological Forum* 16:3, 575–93.

—— (2005). 'Clarifying Comparative-Historical Methodology', *APSA Qualitative Methods Newsletter* 3:1, 19–22.

—— (2004). 'Comparative-Historical Methodology', *Annual Review of Sociology* 30, 81–101.

—— (2007a). 'Debating the State of Comparative Politics: Views from Qualitative Research', *Comparative Political Studies* 40:1, 32–8.

—— (1999) 'Nominal, Ordinal, and Narrative Appraisal in Macrocausal Analysis', *American Journal of Sociology* 104:4, 1154–96.

—— (2007b). 'Qualitative Methodology and Comparative Politics', *Comparative Political Studies* 40:2, 122–44.

—— (2003a). 'Strategies of Causal Assessment in Comparative Historical Analysis', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 337–72.

—— (2000). 'Strategies of Causal Inference in Small-N Analysis', *Sociological Methods & Research* 28:4, 387–424.

—— (2003b) 'Tentative Answers to Questions about Causal Mechanisms'. Paper presented at the 2003 annual meeting of the American Political Science Association.

J. Mahoney and G. Goertz (2004). 'The Possibility Principle: Choosing Negative Cases in Comparative Research', *American Political Science Review* 98:4, 653–69.

—— (2006). 'A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research', *Political Analysis* 14:3, 227–49.

J. Mahoney, E. Kimball, and K. L. Koivu (2009). 'The Logic of Historical Explanation in the Social Sciences', *Comparative Political Studies* 42:1, 114–46.

J. Mahoney, and D. Rueschemeyer (2003a). 'Comparative Historical Analysis. Achievement and Agendas', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 3–38.

—— (2003b). *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 3–38.

J. Mahoney, and P. L. Terrie (2008). 'Comparative-Historical Analysis in Contemporary Political Science', in J. M. Box-Steffensmeier, H. E. Brady, and D. Collier (eds), *The Oxford Handbook of Political Methodology*, Oxford and New York: Oxford University Press, 737–55.

J. Mahoney, and K. A. Thelen (2009). 'A Theory of Gradual Institutional Change', in J. Mahoney and K. A. Thelen (eds), *Explaining Institutional Change: Ambiguity, Agency, and Power*. Cambridge: Cambridge University Press, 1–37.

J. Mahoney, and C. M. Villegas (2007). 'Historical Enquiry and Comparative Politics', in C. Boix and S. C. Stokes (eds), *The Oxford Handbook of Comparative Politics*. Oxford: Oxford University Press, 73–89.

P. T. Manicas (2006). *A Realist Philosophy of Social Science: Explanation and Understanding*. Cambridge and New York: Cambridge University Press.

E. D. Mansfield, and J. L. Snyder (2005). *Electing to Fight: Why Emerging Democracies Go to War*. Cambridge, Mass.: MIT Press.

J. G. March and J. P. Olsenin (2006). 'The Logic of Appropriateness', in M. Moran, M. Rein, and R. E. Goodin (eds), *The Oxford Handbook of Public Policy*. Oxford: Oxford University Press, 689–708.

J. A. Maxwell (2010). 'Using Numbers in Qualitative Research', *Qualitative Inquiry* 16:6, 475–82.

R. Mayntz (2004). 'Mechanisms in the Analysis of Social Macro-phenomena', *Philosophy of the Social Sciences* 34:2, 237–59.

D. McAdam, S. Tarrow, and C. Tilly (2008). 'Methods for Measuring Mechanisms of Contention', *Qualitative Sociology* 31:4, 307–31.

T. J. McKeown (1999). 'Case Studies and the Statistical World View', *International Organization* 53:1, 161–90.

J. S. Mill (1874). *A System of Logic, Ratiocinative and Inductive*. New York: Harper & Brothers.

S. L. Morgan and C. Winship (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

H. J. Morgenthau and K. W. Thompson (1985). *Politics among Nations: The Struggle for Power and Peace*. New York: Knopf.

R. B. Morton and K. C. Williams (2010). *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge: Cambridge University Press.

B. A. Most and H. Starr (1989). *Inquiry, Logic, and International Politics*. Columbia: University of South Carolina Press.

G. L. Munck (2005). 'Ten Fallacies about Qualitative Research', *Qualitative Methods – Newsletter of the APSA Organized Section on Qualitative Methods* 3:1, 2–5.

G. L. Munck and J. Verkuilen (2002). 'Conceptualizing and Measuring Democracy – Evaluating Alternative Indices', *Comparative Political Studies* 35:1, 5–34.

K.-H. Nassmacher (2010). 'The Dilemma of Depth versus Breadth in Comparing Political Systems Empirically…and How to Overcome It', *European Political Science* 7, 113–25.

E. Nichols (1986). 'Skocpol on Revolution: Comparative Analysis vs. Historical Conjuncture', *Comparative Social Research* 9, 163–86.

R. Northcott (2010). 'Natural-born Determinists: A New Defense of Causation as Probability-raising', *Philosophical Studies* 150:1, 1–20.

H. Obinger, S. Leibfried, C. Bogedan, E. Gindulus, J. Moser, and P. Starke (2010). *Transformations of the Welfare State: Small States, Big Lessons*. Oxford: Oxford University Press.

J. S. Odell (2004). 'Case Study Methods in International Political Economy', in D. F. Sprinz and Y. Wolinsky-Nahmias (eds), *Models, Numbers, and Cases: Methods for Studying International Relation*. Ann Arbor: University of Michigan Press, 56–80.

R. J. Oliver (2011). 'Powerful Remnants? The Politics of Egalitarian Bargaining Institutions in Italy and Sweden', *Socio-Economic Review* 9:3, 533–66.

A. J. Onwuegbuzie and N. L. Leech (2005). 'Taking the "Q" Out of Research: Teaching Research Methodology Courses without the Divide between Quantitative and Qualitative Paradigms', *Quality & Quantity* 39, 267–96.

J. M. Owen (1994). 'How Liberalism Produces Democratic Peace', *International Security* 19:2, 87–125.

R. Pahre (2005). 'Formal Theory and Case-Study Methods in EU Studies', *European Union Politics* 6:1, 113–45.

—— (2001). 'Most-Favored-Nation Clauses and Clustered Negotiations', *International Organization* 55:4, 859–90.

—— (2008). *Politics and Trade Cooperation in the Nineteenth Century: The 'Agreeable Customs' of 1815–1914*. Cambridge: Cambridge University Press.

—— (1998). 'Reactions and Reciprocity: Tariffs and Trade Liberalization from 1815 to 1914', *Journal of Conflict Resolution* 42:4, 467–92.

J. M. Paige (1999). 'Conjuncture, Comparison, and Conditional Theory in Macrosocial Inquiry', *American Journal of Sociology* 105, 781–800.

J. H. Park and N. Jensen (2007). 'Electoral Competition and Agricultural Support in OECD Countries', *American Journal of Political Science* 51:2, 314–29.

S. Park (2009). 'Philosophical Responses to Underdetermination in Science', *Journal for General Philosophy of Science* 40:1, 115–24.

P. Pierson (2000). 'The Limits of Design: Explaining Institutional Origins and Change', *Governance* 13:4, 475–99.

—— (2004). *Politics in Time*. Princeton, N.J.: Princeton University Press.

B. M. Pollins (2007). 'Beyond Logical Positivism: Reframing King, Keohane, and Verba', in R. N. Lebow and M. I. Lichbach (eds), *Theory and Evidence in Comparative Politics and InternationalRrelations*. Houndmills: Palgrave, 87–106.

D. Prakash and A. Klotz (2007). 'Should We Discard the "Qualitative" versus "Quantitative" Distinction?', *International Studies Review* 9:4, 753–70.

A. Prontera (2010). 'Europeanization, Institutionalization and Policy Change in French and Italian Electricity Policy', *Journal of Comparative Policy Analysis: Research and Practice* 12:5, 491–507.

A. Przeworski (2007). 'Is the Science of Comparative Politics Possible?', in C. Boix and S. C. Stokes (eds), *The Oxford Handbook of Comparative Politics*. Oxford: Oxford University Press, 147–71.

A. Przeworski and H. Teune (1970). *The Logic of Comparative Social Inquiry*. New York: Wiley-Interscience.

J. Quadagno and S. J. Knapp (1992). 'Have Historical Sociologists Forsaken Theory?: Thoughts on the History/Theory Relationship', *Sociological Methods and Research* 20:4, 481–507.

C. C. Ragin (1981). 'Comparative Sociology and the Comparative Method', *International Journal of Comparative Sociology* 22:1–2, 102–20.

—— (1992). '"Casing" and the Process of Social Inquiry', in C. C. Ragin and H. S. Becker (eds), *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge: Cambridge University Press, 217–26.

—— (1987). *The Comparative Method: Moving beyond Quantitative and Qualitative Strategies*. Berkeley, Calif.: University of Berkeley Press.

—— (2000). *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

—— (2006a). 'How to Lure Analytic Social Science out of the Doldrums: Some Lessons from Comparative Research', *International Sociology* 21:5, 633–46.

—— (2008). *Redesigning Social Inquiry*. Chicago: Chicago University Press.

—— (2006b). 'Set Relations in Social Research: Evaluating Their Consistency and Coverage', *Political Analysis* 14:3, 291–310.

—— (1997). 'Turning the Tables: How Case-Oriented Research Challenges Variable-Oriented Research', *Comparative Social Research* 16, 27–42.

C. C. Ragin and D. Zaret (1983). 'Theory and Method in Comparative Research – Two Strategies', *Social Forces* 61:3, 731–54.

B. Rihoux and G. De Meur (2008). 'Crisp-Set Qualitative Comparative Analysis (CSQCA)', in B. Rihoux and C. Ragin (eds), *Configurational Comparative Methods*. Thousand Oaks, Calif.: Sage, 33–68.

R. Rogowski (1995). 'The Role of Theory and Anomaly in Social-Scientific Inference', *American Political Science Review* 89:2, 467–70.

I. Rohlfing (2008). 'What You See and What You Get: Pitfalls and Principles of Nested Analysis in Comparative Research', *Comparative Political Studies* 41:11, 1492–514.

I. Rohlfing (2012). 'Causal inference and case selection in comparative hypothesis testing via process tracing', typescript.

S. Rosato (2003). 'The Flawed Logic of Democratic Peace Theory', *American Political Science Review* 97:4, 585–602.

D. Rueschemeyer (2003). 'Can One or a Few Cases Yield Theoretical Gains?', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 305–32.

—— (2009) *Usable Theory: Analytic Tools for Social and Political Research*. Princeton, N.J.: Princeton University Press.

J. Runde and M. de Rond (2010). 'Evaluating Causal Explanations of Specific Events', *Organization Studies* 31:4, 431–50.

F. Russo and J. Williamson (2007). 'Interpreting Causality in the Health Sciences', *International Studies in the Philosophy of Science* 21:2, 157–70.

W. C. Salmon (1998). *Causality and Explanation*. New York: Oxford University Press.

G. Sartori (1991). 'Comparing and Miscomparing', *Journal of Theoretical Politics* 3:3, 243–57.

—— (1970). 'Concept Misformation in Comparative Politics', *American Political Science Review* 64:4, 1033–53.

J. Savolainen (1994). 'The Rationality of Drawing Big Conclusions Based on Small Samples: In Defense of Mill's Methods', *Social Forces* 72:4, 1217–24.

R. A. Sayer (2010). *Method in Social Science: A Realist Approach*. London and New York: Routledge.

J. Schaffer (2003). 'Overdetermining Causes', *Philosophical Studies* 114:1–2, 23–46.

C. Q. Schneider and C. Wagemann (2012). *Set-Theoretic Methods for the Social Sciences. A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

—— (2010) 'Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy Sets', *Comparative Sociology* 9:3, 397–418.

G. Schneider and N. P. Gleditsch (2010). 'The Capitalist Peace: The Origins and Prospects of a Liberal Idea', *International Interactions* 36:2, 107–14.

P. A. Schrodt (2010). 'Seven Deadly Sins of Contemporary Quantitative Political Analysis'. Working paper.

K. A. Schultz (2001). *Democracy and Coercive Diplomacy*. Cambridge Cambridge University Press.

J. Seawright and D. Collier (2004). 'Glossary', in H. E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Lanham, Md.: Rowman & Littlefield, 273–313.

J. Seawright and J. Gerring (2008). 'Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options', *Political Research Quarterly* 61:2, 294–308.

R. Sil (2000). 'The Foundations of Eclecticism: The Epistemological Status of Agency, Culture, and Structure in Social Theory', *Journal of Theoretical Politics* 12:3, 353–87.

R. Sil and P. J. Katzenstein (2010a) 'Analytic Eclecticism in the Study of World Politics: Reconfiguring Problems and Mechanisms across Research Traditions', *Perspectives on Politics*, 8:2, 411–31.

—— (2010b) *Beyond Paradigms: Analytic Eclecticism in the Study of World Politics, Political Analysis Series* (Houndmills: Palgrave Macmillan).

T. Skocpol (1986). 'Analyzing Causal Configurations in History: A Rejoinder to Nichols', *Comparative Social Research* 9, 187–94.

—— (2003). 'Doubly Engaged Social Science', in J. Mahoney and D. Rueschemeyer (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, 407–28.

T. Skocpol and M. Somers (1980) 'The Uses of Comparative History in Macrosocial Inquiry', *Comparative Studies in Society and History* 22:2, 174–97.

D. Slater (2009). 'Revolutions, Crackdowns, and Quiescence: Communal Elites and Democratic Mobilization in Southeast Asia', *American Journal of Sociology* 115:1, 203–54.

D. Slater and E. Simmons (2010). 'Informative Regress: Critical Antecedents in Comparative Politics', *Comparative Political Studies* 43:7, 886–917.

N. Smelser (1973). 'The Methodology of Comparative Analysis', in D. Warwick and S. Osherson (eds), *Comparative Research Methods*. Englewood Cliffs, N.J.: Prentice-Hall, 42–86.

E. Sober (2009). 'Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads', *Philosophical Studies* 143:1, 63–90.

H. D. Soifer (2010). 'The Causal Logic of Critical Junctures'. Committee on Concepts and Methods Working Paper Series, no. 24.

P. Starke (2008). *Radical Welfare State Retrenchment in Comparative Perspective*. New York: Palgrave Macmillan.

G. Steinmetz (2004). 'Odious Comparisons: Incommensurability, the Case Study, and "Small N's" in Sociology', *Sociological Theory* 22:3, 371–400.

B. Steunenberg and M. Kaeding (2009). '"As Time Goes By": Explaining the Transposition of Maritime Directives', *European Journal of Political Research* 48:3, 432–54.

A. L. Stinchcombe (1991). 'The Conditions of Fruitfulness of Theorizing about Mechanisms in Social-Science', *Philosophy of the Social Sciences* 21:3, 367–88.

—— (1968). *Constructing Social Theories*. New York: Harcourt Brace & World.

R. Stryker (1996). 'Beyond History versus Theory: Strategic Narrative and Sociological Explanation', *Sociological Methods & Research* 24:3, 304–52.

D. T. Studlar (2010). 'What Explains the Paradox of Tobacco Control Policy under Federalism in the U.S. and Canada? Comparative Federalism Theory versus Multi-level Governance', *Publius* 40:3, 389–411.

S. Tarrow (2010). 'The Strategy of Paired Comparison: Toward a Theory of Practice', *Comparative Political Studies* 43:2, 230–59.

K. Thelen (1999). 'Historical Institutionalism in Comparative Politics', *Annual Review of Political Science* 2, 369–404.

K. Thelen and S. Steinmo (1992). 'Historical Institutionalism in Comparative Politics', in S. Steinmo, K. Thelen, and F. Longstreth (eds), *Structuring Politics: Historical Institutionalism in Comparative Analysis*. Cambridge: Cambridge University Press, 1–32.

C. G. Thies (2002). 'A Pragmatic Guide to Qualitative Historical Analysis in the Study of International Relations', *International Studies Perspectives* 3:4, 351–72.

G. Thomas (2010). 'Doing Case Study: Abduction Not Induction, Phronesis Not Theory', *Qualitative Inquiry* 16:7, 575–82.

—— (2011). 'A Typology for the Case Study in Social Science Following a Review of Definition, Discourse, and Structure', *Qualitative Inquiry* 17:6, 511–21.

M. Tight (2010). 'The Curious Case of Case Study: A Viewpoint', *International Journal of Social Research Methodology* 13:4, 329–39.

C. Tilly (1997). 'Means and Ends of Comparison in Macrosociology', *Comparative Social Research* 16, 43–53.

—— (2001). 'Mechanisms in Political Processes', *Annual Review of Political Science* 4, 21–41.

—— (2004). 'Social Boundary Mechanisms', *Philosophy of the Social Sciences* 34:2, 211–36.

C. Trampusch (2010). 'Employers, the State and the Politics of Institutional Change: Vocational Education and Training in Austria, Germany and Switzerland', *European Journal of Political Research* 49:4, 545–73.

G. Tsebelis (2002). *Veto Players: How Political Institutions Work*. Princeton, N.J.: Princeton University Press.

S. Van Evera (1997). *Guide to Methods for Students of Political Science*. Ithaca. N.Y.: Cornell University Press.

B. Vis (2009). 'Governments and Unpopular Social Policy Reform: Biting the Bullet or Steering Clear?' *European Journal of Political Research* 48:1, 31–57.

E. J. Wagenmakers, R. Wetzels, D. Borsboom, and H. van der Maas (2011). 'Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi', *Journal of Personality and Social Psychology* 100:3, 426–32.

D. Waldner (2012 [forthcoming]). 'Process Tracing and Causal Mechanisms', in H. Kincaid (ed.), *The Oxford Handbook of the Philosophy of Social Science*. Oxford: Oxford University Press.

—— (2007). 'Transforming Inference into Explanation: Lessons from the Study of Mass Extinctions', in R. N. Lebow and M. I. Lichbach (eds), *Theory and Evidence in Comparative Politics and International Relations*. Houndmills: Palgrave, 145–75.

H. A. Walker and B. P. Cohen (1985). 'Scope Statements – Imperatives for Evaluating Theory', *American Sociological Review* 50:3, 288–301.

S. Walter (2010). 'Globalization and the Welfare State: Testing the Microfoundations of the Compensation Hypothesis', *International Studies Quarterly* 54:2, 403–26.

—— (2008). 'A New Approach for Determining Exchange-Rate Level Preferences', *International Organization* 62:3, 405–38.

P. V. Warwick and J. N. Druckman (2006). 'The Portfolio Allocation Paradox: An Investigation into the Nature of a Very Strong but Puzzling Relationship', *European Journal of Political Research* 45:4, 635–65.

J. Waskan (2011). 'Mechanistic Explanation at the Limit', *Synthese* 183:3, 389–408.

J. A. Wasserman, J. M. Clair, and K. L. Wilson (2009). 'Problematics of Grounded Theory: Innovations for Developing an Increasingly Rigorous Qualitative Method', *Qualitative Research* 9:3, 355–81.

E. Weber (2007). 'Social Mechanisms, Causal Inference, and the Policy Relevance of Social Science', *Philosophy of the Social Sciences* 37:3, 348–59.

B. Western (2001). 'Bayesian Thinking about Macrosociology', *American Journal of Sociology* 107:2, 353–78.

J. Woodward (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
—— (2011). 'Mechanisms Revisited', *Synthese* 183:3, 409–27.
K. Yesilkagit and J. G. Christensen (2010). 'Institutional Design and Formal Autonomy: Political versus Historical and Cultural Explanations', *Journal of Public Administration Research and Theory* 20:1, 53–74.
R. K. Yin (2008). *Case Study Research: Design and Methods*. Thousand Oaks, Calif.: Sage.
B. Zangl (2008). 'Judicialization Matters! A Comparison of Dispute Settlement under GATT and the WTO', *International Studies Quarterly* 52:4, 825–54.
M. Zelditch (1971). 'Intelligible Comparisons', in I. Vallier (ed.), *Comparative Methods in Sociology: Essays on Trends and Applications*. Berkeley: University of California Press, 267–307.
D. Ziblatt (2009). 'Shaping Democratic Practice and the Causes of Electoral Fraud: The Case of Nineteenth-Century Germany', *American Political Science Review* 103:1, 1–21.
A. S. Zuckerman (1997). 'Reformulating Explanatory Standards and Advancing Theory in Comparative Politics', in M. I. Lichbach and A. S. Zuckerman (eds), *Comparative Politics: Rationality, Culture and Structure*. New York: Cambridge University Press, 277–310.

Index