USING CASE STUDIES IN THE SOCIAL SCIENCES:

METHODS, INFERENCES, PURPOSES

Attilia Ruzzene

Ruzzene, A.

Using case studies in the social sciences. Methods, inferences, purposes.

Printing: Ridderprint BV, www. ridderprint.nl

Cover: Ridderprint BV, www. ridderprint.nl

Illustration by Minjung Kim (Gwangju 1962, --).

Front cover: *Fullness in void* (2006)

Back cover: *Sigh and me* (1998)

USING CASE STUDIES IN THE SOCIAL SCIENCES:

METHODS, INFERENCES, PURPOSES

Gebruik van case studies in de sociale wetenschappen:
methoden, gevolgtrekkingen en doeleinden

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus
Prof.dr. H.A.P. Pols
and in accordance with the decision of the Doctorate Board

The public defence shall be held on
Thursday, 20 November 2014 at 15:30 hrs

by

Attilia Ruzzene
born in Pordenone, Italy

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral committee

Promotors:

Prof.dr. J.M. Reiss

Prof.dr. J.J. Vromen

Other members:

Prof.dr. W. Hout

Prof.dr. M.S. Morgan

Prof.dr. F.A. Muller

For Santiago, and for Franck

Acknowledgments

I started working on this project several years ago; since then, many people contributed to its development in different ways. I feel a sense of deep gratitude towards them all.

The methodological interest that informs this thesis first arose in the department of economics Cognetti de Martiis, in Torino. I thank David Lane for awakening in me the sense of philosophical wonder; Roberto Marchionatti for pushing me to follow this drive; and Franco Donzelli for his generous and genuine interest in my intellectual path.

Eipe has been in these years my second family. It has changed me in ways that were unimaginable before: it changed my way of seeing the world, of experiencing it, and of making sense of it. Here, I felt deeply challenged and enormously engaged. This was a gift of incommensurable value.

Caterina Marchionni has been a great teacher and wonderful companion who helped me to move the first steps in this inquiry.

My friends made this experience worthwhile. I love them all, and felt very much loved by them: Anil, Mary, Luis, Clemens, Tom, Tyler, Daniel, Josh, Pedro, Sine, Ana, and Melissa. They have been my friends, my fellows, my brothers, and sisters. Their smiles and words made my days shine. These are the jewels I carry within.

Over the years Eipe has attracted many other bright students who enriched my life in unexpected ways. François Claveau has been a wonderful fellow in this experience. Many others joined Eipe afterwards. I feel deep gratitude towards them all.

I would have never accomplished this project without the patient help of Ticia Herold, Lizzy Patilaya-Ternatus, Mark Dambrink, Manon Geluk, and Ceciel Meiborg.

Many other people crossed my path in these years and gave a precious contribution to this experience: Piet Steenbakkers, James McAllister, Federica Russo, Phyllis McKay Illari, Petri Ylikoski, Maria Jimenez Buedo, David Teira Serrano, Stephen Turner, Paul Roth, Mary Morgan, Marcel Boumans, Fred Muller, Ingrid Robeyns, Conrad Heilmann, and Constanze Binder.

The hardest words to find are those most deserved. I owe the accomplishment of this thesis to my supervisors, Jack Vromen and Julian Reiss. They gave me the chance to join Eipe, to explore and develop myself in this environment, and to fulfill a project that was professional as much as personal. I thank them for the trust they had in me, for the respect they showed for my work, and for understanding the twists and turns of my path. Jack has been an exemplar of intellectual integrity and humbleness. In his supervision I found support, encouragement, and profound care. Julian triggered the passion for what I do, and still helps me to keep it alive. His teachings and insights have been a continuous source of inspiration, and shaped this thesis in uncountable ways. The intellectual debt I owe to him is enormous, and so is the gratitude for the most valuable gift a teacher could ever offer to his students, that is, becoming an independent thinker.

In these years I had often the feeling of reaching boundaries: the limitations of my intellectual capacity, and the more bitter limitations of me as a human being. I tried to face these limitations, and to push the boundaries a little bit further. It has been a painful exploration which made my life worthwhile. If I had the strength to delve in it, I owe it to my family, for it is a living example of truthfulness, courage, and compassion.

I gratefully acknowledge the financial support provided by the Erasmus Institute for Philosophy and Economics.

Chapter 2 is based on “Process-tracing as an Effective Epistemic Complement” published in *Topoi*, 33 (2): 61-72, 2014.

Chapter 3 is based on “Drawing Lessons from Case Studies by Enhancing Comparability” published in *Philosophy of the Social Sciences*, 42 (1): 99-120, 2012.

Chapter 4 is based on “Policy making in developing countries: from prediction to planning” forthcoming in *Journal of Economic Methodology*.

USING CASE STUDIES IN THE SOCIAL SCIENCES: METHODS, INFERENCES, PURPOSES

Table of Contents

1. WORKING WITH CASE STUDIES IN THE SOCIAL SCIENCES: THE ISSUES AHEAD

1.1 INTRODUCTION: CASES AND CASE STUDIES

Despite fads and fashions in the academic culture, case-based reasoning has proved to be a persistent form of analysis in the social sciences, in the humanities, and even in moral thinking. Broadly understood, case-based reasoning locates the ultimate source of our epistemic and moral intuitions in the concreteness and idiosyncrasy of particulars. Even though they can be traced back to a common root, different traditions of reasoning with cases and of using case studies coexist in the academic landscape. To some extent, the divide between traditions maps onto the distinction between practice-oriented and theory-oriented disciplines, the former being concerned with the solution of concrete problems and the latter mostly aiming at the formulation of abstract conceptualizations and general principles.

One can discern three main *contexts of use*: the case study as pedagogical tool, as method of treatment or intervention, and as an epistemic strategy. Even though, as a matter of fact, contexts of use to some extent match scientific fields, they do so loosely, in the sense that a given field can be hospitable to multiple uses of case studies.

In practice-oriented fields, which to a large extent aim at producing and transmitting practical knowledge that is *knowledge of how* to practice a profession, the study of cases became an essential *pedagogical* tool. Ever since case studies were first introduced in the Harvard Law School in 1870 by C.C. Langdell, and soon after were adopted by the medical and business schools in Harvard,[1] the case-based form of reasoning is being used to familiarize the students with the principles of the discipline and the types of problems they will confront eventually as professional practitioners (Forrester 1996, Creager et al. 2007). The case, in fact, instantiates the basic principles, illustrates both recurrent and atypical problems, and offers practical solutions to those problems. Knowledge of cases is

[1] For a discussion of the use of case studies for teaching purpose in the business school see McNair (1954).

expected to guide the future practitioner through the specificity of the situations she will eventually confront.

In fields such as psychology and psychiatry the case-study is the clinical study. The treatment of the patient creates case-based knowledge in the first place, which is in turn used on subsequent occasions as raw evidence to diagnose, explain, and actually treat other patients. In these domains, case-based reasoning and knowledge, though used to some extent as a pedagogical tool, informs primarily the therapeutic practice.[2] One might look at the art of casuistry as a similar context of use. Casuistry was directed at adjudicating moral issues and based its moral judgments on detailed acquaintance with a huge repertoire of cases, rather than on the principled knowledge of moral theories. It was the method of settling moral disputes and dilemmas used by the priest and the savior, regarded as fully legitimate until Blaise Pascal's *Provincial Letters* contributed to its demise (Jonsen and Toulmin 1988). Pascal attacked fiercely the Jesuits of Paris for using case-based moral reasoning to placate wealthy Church donors while punishing poor penitents. It is still unclear what Pascal's intentions were: whether "he was writing *within* the casuistic tradition, as a *rigorist* objecting to the *laxism* of his contemporaries", or whether he was rejecting the whole tradition as deceptive or compromising (15). Be that as it may, as a matter of fact since Pascal's attack the casuist's method fell into disrepute for a very long time.

In the social sciences, case studies are regarded primarily as an epistemic strategy.[3] It is plausible to distinguish between weaker and stronger epistemic usage of case studies. At the weaker end of the spectrum, case studies occupy an ancillary role in the scientific investigation. That is, they are subordinate to other research strategies for illustrative purposes: in this context the case study is the illustration of a concept, a claim, or even a theory, as instantiated in a concrete case. Alternatively, case studies are the final output of a long and complex research procedure of inquiry, case study research (CSR), aimed at

[2] For a distinction between clinical study and case study see Eckstein (2000 [1975]). For a criticism of the use of case studies in psychiatry see Grϋnbaum (1988); for a defense see Runyan (1982) and Al Rubaie (2002).

[3] In the present thesis, (research) strategy and method are used as synonyms.

acquiring familiarity with, interpreting, characterizing, explaining, and theorizing phenomena of scientific interest; and, in turn, at using these descriptions, explanations, theorizations to predict and control social outcomes. When used as epistemic strategies in this stronger sense, case studies posit hypotheses about phenomena. These hypotheses are subject to methodological scrutiny for what concerns their credibility and their capacity to fulfill the research goals.

Despite being widespread, the use of case studies to establish hypotheses in the social sciences has typically been regarded with suspicion, or with lenience, by some part of the scientific community (Gerring 2007a, Flyvbjerg 2006). Some scholars simply dismissed case studies as merely providing *anecdotal* knowledge and likened them to mere story-telling (Campbell 1975, Eysenck 1976)[4]. Other scholars adopted a more charitable view and tried to rehabilitate what they call the case study method in the arena of scientific methods (Lijphart 1971, Eckstein 2000 [1975]). They looked at it as a "special case" of the experimental and statistical methods. Similar to the latter in purpose and logic, it would mainly differ from them as to the strength of its findings. Typically, it would be understood as a research strategy of limited reliability, to which one resorts when the other designs, regarded as stronger in these respects, cannot be employed.

More recently, scholars started looking at case studies with renewed attention and it is reasonable to talk of a new trend in the field. Testimony of this turn is the increasing number of methodologically informed publications in the social sciences and in philosophy.[5] Several circumstances have contributed to this methodological "awakening": the growing dissatisfaction with the old view that saw merely the method of case studies as a last resort, to be employed in the absence of a better alternative; the growing belief that skepticism about case studies was probably due to a lack of understanding the method; finally, the realization that case studies have a potential to offer that is still

[4] Both Campbell and Eysenck have been initially fierce critics of CSR who have later softened and modified their original views (Flyvberg 2006).

[5] Important works in this areas are Becker and Ragin (1992), Bennett and Elman (2006), Brady and Collier (2004), George and Bennett (2005), Gerring (2007a), Ragin (2000).

unexplored. This new perspective emphasizes the specificity of case studies and aims at assessing how they can best fulfill their potential.

This new perspective informs my thesis. Its main focus is the use of case studies in the social sciences as an epistemic strategy to formulate, establish, and generalize causal hypotheses. A secondary focus is an investigation into the use of causal findings generated in case studies to inform policy making in the social realm. The case study, as I understand it, is the thick analysis of social phenomena as they occurred in circumscribed contexts. It is thus a *situated* analysis, rich in detail and directed to represent faithfully the complexity of the object of interest. As such, it constitutes a reservoir of local knowledge that becomes evidentially useful when understanding and intervening in the social world in order to fix it, change it, and improve it. One way in which case studies can perform this evidential role is by formulating causal hypotheses in a rigorous and conscientious way.

In this chapter, I characterize what can be regarded as two alternative views of case studies and the understanding of science in which they are embedded.

The first approach flourished in the 70s and looked at case studies as a special, and typically weaker, form of the experimental, statistical, or comparative methods. Since this approach tends to evaluate case studies by criteria belonging to other methodological traditions, it can be said to present a *heteronomous* paradigm. This view prevented the full grasp of the specificity and potential of case studies. The second, alternative view, which developed during the last decades, is taking shape gradually and is still far from being fully articulated. This approach strives for an understanding of case studies liberated from the narrow mindset that caricatures case studies as the method of last resort. In particular, it sees case studies as an *autonomous epistemic genre* (Morgan 2012). It thus carries the promise to come to terms with the idiosyncrasy that renders this research strategy so peculiar and challenging.

The two approaches are rarely found in their ideal typical form and there is no neat dividing line between authors and historical periods. Recent contributors to the debate are not always fully consistent in their adherence to the new paradigm and are still often conditioned by the old one. It is nevertheless useful to keep the two views distinct in order to state more clearly the reasons why the former approach should be abandoned and the latter is where we need to move instead.

1.2 THE HETERONOMOUS PARADIGM

The traditional way to understand Case Study Research (CSR) is to look at it as a "special case" of the other research strategies, be it the statistical method, the comparative, or yet the experimental one (Morgan 2012). This view emphasizes the affinity in logic and purposes between CSR and the other research strategies and reduces their differences to matters of degree. CSR would, in fact, differ from the other methods essentially because it focuses on a *single case* rather than many (Lijphart 1971). From the homogeneity of purpose among the methods, it follows that CSR is characterized as stronger, or weaker, than the other strategies in some relevant respects, such as its capacity to discover phenomena, formulate new hypotheses, and suggest new concepts (heuristic power), or its capacity to provide rational grounds for phenomena, theories, and concepts already in use (testing power).

Within the traditional view, positions differ as to the comparative strengths and weakness of CSR. Some scholars, for instance, see it as stronger than the other methods as to heuristic power while others recommend the use of statistical methods for formulating new theories and concepts (Eckstein 2000 [1975]). On the other hand, scholars disagree on whether the testing power constitutes a comparative advantage or disadvantage of CSR. Eckstein, for example, defends the admittedly unconventional view that regards CSR as comparatively powerful as a testing tool, while the majority in this tradition defends the superiority of the alternative methods for theory justification (Flyvbjerg 2006).

Despite disagreement as to the comparative advantages or disadvantages of CSR, the traditional view emphasizes the homogeneity of purpose and rationale between CSR and the other research methods. In particular, adherents to this view typically maintain the following tenets:[6]

a) Science centers on the formulation of very broad, or universal, generalizations;

b) It primarily aims to provide *explanations* of scientific phenomena;

c) Scientific explanations employ general propositions (see a above) that relate two or more variables, or *regularities*.[7]

d) The scientific practice is structured around:

1. The context of discovery: the process of finding new scientific hypotheses.
2. The context of justification: the process by means of which scientific hypotheses are judged.[8]

According to this view the comparative strength of a method depends on the extent to which it contributes either to the context of discovery or to the context of justification. Scientific hypotheses are seen in this context as general propositions that describe empirical regularities. When assessed in this framework, the heuristic and testing power of CSR depends on how CSR, regarded as "the study of a phenomenon for which we report

[6] These four tenets are by no means exhaustive of the traditional view discussed here. The adherents to this view of CSR also share to a large extent a similar explanatory model, hold to a sharp distinction between *facts* and *theory*, and so on. My purpose here is not to fully articulate the underlying conception of science; rather, it is presenting the underlying tenets that are more directly relevant to their assessment of CSR.

[7] I shall discuss below that *regularities* are but a form of scientific generalizations. Mechanisms are generalizations of a different form, at least according to some scholars (see section 1.3).

[8] Paul Hoyningen-Huene notices how the distinction between context of discovery and justification (the DJ distinction), as used in the 1960s and 1970s, is not just one distinction but a set of intermingled distinctions (2006: 119). He thus distinguishes five versions of the DJ distinction (six including the one he proposes): two of them concern the object-level that is discovery and justification as things in the world; the other four distinguish between discovery and justification at the meta-level by appealing to the methodological and disciplinary differences in the *analysis* of discovery and justification. In my understanding, adherents to the traditional view of CSR *most probably conflate* the five versions and *certainly endorse* the first version which sees discovery and justification as *distinct temporal processes*. This version is regarded as untenable by Hoyningen-Huene because it is often not possible in the history of science to identify processes of discovery as opposed to processes of justification (2006: 121; see also section 1.3 below).

only a single measure on any pertinent variable" (Eckstein 2000 [1975]: 124), contributes either to the formulation of new regularities or to their corroboration.

This view as applied to CSR is epitomized in some influential contributions such as Arend Lijphart (1971), Harry Eckstein (2000 [1975]), Joseph Campbell (1975), and Stanley Lieberson (1991). However, it is not exhausted there. Even though certainly dominant among social scientists back in the 1970s, traces of it still survive nowadays in some major contributions to the literature on CSR (Vennesson 2008).

1.2.1 DEGREES OF FREEDOM AND GENERALIZATIONS

In *Comparative Politics and the Comparative Method* (1971) Lijphart articulates an approach to CSR that can be regarded as paradigmatic of the traditional view:[9] he characterizes science as an explanatory endeavor to which the various research strategies contribute by establishing empirical regularities among variables of interest. Establishing regularities in a reliable way demands, however, controlling the other relevant variables, which the various research strategies attain to a different extent (1971: 683). Lijphart thus articulates a *hierarchy* of methods with (1) the ideal controlled experiment on top, and, in descending order, (2) the randomized experiment, (3) the statistical method, (4) the comparative method, and (5) CSR respectively. The methods in the hierarchy share the same logic but pursue control by different means and to a different extent. These means, in fact, are not all equally effective to the end of control. In particular, the lower a given strategy is in the hierarchy, the more imperfect is the control it achieves on the relevant variables.

The ideal controlled experiment creates two identical situations in which perfect control of the relevant variables is achieved and the relationship between the variables of interest is therefore established with certainty. In most actual scientific practice this ideal situation can only be approximated and the confidence in the conclusion weakened accordingly.

[9] Even though Lijphart's view is certainly mainstream within the traditional view not everybody shares his hierarchical view on methods that sees the ideal experiment as the gold standard (see below).

(2) When the ideal controlled experiment is not available, randomized experiments can (sometimes) be used that aim to create two equivalent groups in which causal factors are identically distributed. To this end subjects are assigned to either group by way of a randomizing procedure. One of the groups is exposed to a treatment, the experimental group, and the other is given a control, the control group. A positive experimental result is evidence for the causal claim that the difference in the outcome observed in the two groups is due to the stimulus provided. Insofar as the randomization is performed successfully, the assumption that the causal factors are identically distributed in the two groups is satisfied and the ensuing conclusion is safely established.[10]

(3) At the next level below, the statistical method relies on the same logic of control but uses observational data that cannot be manipulated experimentally. The manipulation of the data is in this case conceptual and performed by way of partial correlation. The data sample is divided into different sub-groups within which the relevant variables are held fixed. Each sub-group is examined to see whether the regularity of interest holds when the other relevant variables are controlled for. Unlike the experimental design where *all* relevant variables are taken care of by way of a successful randomization,[11] the statistical method attains control only for those variables that are *known* to be relevant. For this reason Lijphart regards it as an approximation to the controlled experiment.

[10] This formulation is somewhat imprecise in two respects. First, other assumptions besides the causal factors being identically distributed in the two groups need to be satisfied for "the ensuing conclusion to be safely established". They include assumptions about the relationship between probability and causation and the inference of probabilities from frequencies (see also Chapter 4). Second, more than just successful randomization is required for causes to be identically distributed. In fact, randomization when properly implemented avoids *systematic* differences between the treatment and control group that might be due to the assignment procedure (selection bias). However, it only avoids "chance differences" between the two groups in the long run or for high numbers that are not frequently available in the social sciences. To use the methodologist's language, groups are equated "on *expectations* at pretest" that is, the *means* of relevant variables can differ between groups in the short run due to sampling error (Shadish, Cook and Campbell 2002). Furthermore, differences that are neither due to sampling error or selection bias might arise between the treatment and control group. They might be caused, for example, by failure of blinding, failure of compliance, and so on.

[11] See fn. 9 above.

(4) The comparative method is, regarded as an approximation of the ideal experiment, weaker than the statistical method which "it resembles in all respects except one: the number of cases is too small to permit systematic control by means of partial correlation (1971: 684)". The number of cases is crucial for the degree of control that one can obtain: one, in fact, has to be able to observe whether the regularity in question persists when the other relevant variables are held fixed. For this, one needs to examine a large number of cases. One should thus resort to the comparative method only when the number of available cases is too small for finding reliably partial correlations. The degree of control one achieves by means of this method is therefore very limited.

There are ways, however, to alleviate the problem that the comparative method encounters of too many variables to control for and too few cases to examine. Lijphart proposes a few strategies to this end. They are either directed at increasing the number of cases subject to analysis, for instance by using variables that are more widely applicable and thus comparable across a larger number of cases in such a way that the sample is enlarged accordingly. Alternatively, they reduce the number of variables that need to be controlled for, for example by a general commitment to theoretical parsimony and by judiciously restricting the analysis to the really key variables while omitting those of only marginal importance (686-690).

(5) The lack of control becomes a much more troubling concern in CSR where the number of cases reduces to one. If one examines a single case, in fact, one only obtains a single observation for each variable of interest.[12] Due to lack of variation in the putative cause and lack of control on the other relevant variables, it seems thus impossible to establish a reliable conclusion by means of case studies. Multiple explanations are in fact compatible with the available evidence and the observations one has are insufficient to rule them out. The lack of control in CSR has been famously reframed as the problem of degrees of freedom (Campbell 1975, George and Bennett 2005).

[12] This is the definition of case study provided by the traditional view (see Eckstein's definition in sec. 1.2).

This notion, widely used in statistics,[13] was then carried over to the case-study context. In general, the number of degrees of freedom is always equal to the number of observations minus the number of necessary relations among those observations (Walker 1940). If applied to a statistical sample, it is the difference between the sample size (the number of independent observations) and the parameters estimated on those observations, such as sample means and variance. In statistics it is a crucial concept because the sampling distribution depends from it, and if the number of degrees of freedom are miscalculated the level of significance of the test is also misinterpreted (1940: 260). In general, a high number of degrees of freedom is needed in order to have a stronger design. Intuitively, the higher the number of observations, the better is the evidence against which to test alternative hypotheses. It follows that the chance that one hypothesis fits the whole data is lower and that alternative hypotheses are ruled out is higher. Hence, if a hypothesis fits the data, the confidence in it is also higher.

Transferred to CSR, what one would observe in this case is a total lack of degrees of freedom; actually, *negative* degrees of freedom (George and Bennett 2005: 28). If one treats the case study as the study of a *single case* in the statistical sense, in fact, this would be understood as a situation in which a single independent observation is obtained (N=1) and too many necessary relations are imposed on that one observation (the relevant variables). One thus faces here a situation of under-determination of theory by the data where multiple explanations that fit the same evidence equally well are available, and one does not have the means to rule all but one out. We would therefore have a very weak design where the chance of each alternative hypothesis of fitting the data is very high, and the confidence in it is correspondingly lower.

It is important to stress that this conclusion depends on treating the case study as the study of one case where a single observation is collected on each relevant variable. This view was already challenged by Campbell (1975: 179) and since then by many others (among which Gerring 2004, 2007a, George and Bennett 2005). Even though the type of

[13] The notion was introduced by physician James Clerk Maxwell in the 19th century.

answers provided by each author slightly differs, the general response has been that the number of observations (and, thus, the evidence against which the hypotheses are tested) dramatically increases if one considers that observations are collected not just for the single variables of interest, but also for the various, observable, implications generated by the theory. In some sense, says Campbell, the researcher has tested the theory with degrees of freedom coming from the multiple implications of any one theory; and she does not retain the one theory unless most of these are also confirmed (1975: 182).

There is a distinct but related problem arising from the limited number of cases. As mentioned above, the view of CSR that is described here depends on a very specific view of science broadly understood. Science is regarded by these scholars, first and foremost, as an activity orientated to the production of general knowledge (Eckstein 2000 [1975]) that takes the form of empirical regularities. When the empirical material is constituted by a single case, the evidential basis seems too narrow to extract general knowledge that is valid in circumstances different than the ones actually examined. From this *prima facie* plausible observation, these scholars concluded that, if the central goal of the scientific practice consists in establishing generalizations, then it is hard to understand what place CSR can occupy in such an endeavor. If learning about the general is rendered difficult in the analysis of a single case, the worthiness of the whole enterprise becomes disputable. In this respect Lijphart comments (691):

> The scientific status of the case method is somewhat ambiguous because science is a generalizing activity. A single case can constitute neither the basis for a valid generalization nor the ground for disproving an established generalization.

This conclusion, however, highlights a tension in the view described by Lijphart. Even if one agrees with Lijphart that generalizing is problematic in case studies and that science primarily aims at broad generalizations, he would still have to explain *how* experiments solve this problem successfully so as to deserve to be ranked first in his hierarchy of methods while case studies are confined to the bottom of the ranking. In fact, experiments and case studies seem to face a very similar challenge. Well renowned experimentalists say in this respect:

> Most experiments are highly localized and particularistic. They are almost always conducted in a restricted range of settings, often just one, with a particular version of one type of treatment rather than, say, a sample of all possible versions [...] A conflict seems to exist between the *localized* nature of the causal knowledge that individual experiments provide and the more generalized causal goals that research aspires to attain (Shadish, Cook and Campbell 2002: 18-19).

If one pays heed to what these authors say, one is led to think that also experiments establish very narrow causal conclusions. The tension in Lijphart's view might be then explained by the (implicit) assumption that results obtained via experimental methods *automatically* generalize. This is arguably a strong assumption for which no justification is provided. Lijphart seems to entertain the idea that the high level of control achieved via experiments *entails* generalizability. This would be, however, an invalid argument. Control and generalizability are two distinct issues and, whatever their relation, certainly the former does *not entail* the latter. I will discuss these issues below under the label of internal and external validity.

Given the lack of control that affects CSR and its limited capacity for producing general knowledge, Lijphart places the CSR at the very bottom of its hierarchy of methods. If the CSR deserves a place among the scientific methods *at all*, however, it remains to be specified what its positive contribution to scientific progress is. Lijphart thinks that, despite the serious problems that affect them, case studies can still contribute to theory development even though in a less direct way (1971: 691).

1.2.2 CASE STUDIES: WHAT ARE THEY GOOD FOR, THEN?

Aware of the challenges that CSR confronts due to its lack of control and generalizing power, Lijphart outlines a taxonomy of case studies where the possible alternative uses of the method are articulated. The way in which case studies can be useful depends on the distinctive relation that each type has to theoretical generalizations. Lijphart distinguishes the following types (1971: 691):

1. A-theoretical case studies;
2. Interpretive case studies;
3. Hypothesis-generating case studies;
4. Hypothesis-confirming case studies;
5. Hypothesis-infirming case studies;
6. Deviant case studies.

Similar taxonomies are outlined by scholars such as Eckstein (2000 [1975]), Flyvbjerg (2006), Vennesson (2008), where the classification, and the principle behind it, remains essentially the same even though some change is introduced in the characterization of each type and the terminology in use. In what follows I will provide a brief description of the types in Lijphart's taxonomy, and refine it when useful with details provided by the other contributors. The general purpose of this description is not to establish which taxonomy is the most accurate or useful, but rather to illustrate the logic they share.

1. <u>Atheoretical case studies</u>

Atheoretical case studies move in a theoretical vacuum (1971: 691). They are neither guided by theory nor aim at the formulation of theoretical generalizations: they are characterized by Lijphart as purely descriptive endeavors.[14] As such their usefulness resides in providing empirical material for further enquiry; in other words, they are data-gathering tools that contribute to theory formation at best indirectly. Eckstein calls this type of case-studies "configurative-idiographic" (2000 [1975]: 132). They aim to present exhaustive depictions of the overall configuration of the individual so as to capture its uniqueness (the configurative element). Furthermore, they bring about the significance of the facts collected by largely intuitive interpretation claiming validity on the ground that

[14] Lijphart is categorizing ideal-types. He clarifies that in the practice case studies seldom fit a specific category neatly. He specifies: "An actual instance of an atheoretical case study probably does not exist, because almost any analysis of a single case is guided at least by some vague theoretical notions and some anecdotal knowledge of some other cases, and usually results in some vague hypotheses or conclusions that have a wider applicability" (1971: 691).

intensive study and empathetic feel for cases provide authoritative insights into them (the idiographic element).

Even though Eckstein agrees with Lijphart that these studies essentially amount to the collection of facts without theoretical import, he also emphasizes that they are often based on philosophical assumptions less innocent than Lijphart seems to suggest. In the end, Eckstein's position towards this type of studies is much more critical than Lijphart's as he sees them as irremediably affected by theoretical poverty. Even though Eckstein does not deny that these studies can be persuasive and subtle in the intuitions they offer, their incapacity (actually, the explicit refusal) to collect facts in a systematic way renders them useless for theory building.

2. Interpretive case studies

Similar to atheoretical case studies, interpretive case studies are selected because of an interest in the case rather than in the formulation of a general theory (Lijphart 1971: 692). They have a stronger link to theory than atheoretical case studies but the link is still weak. In fact, they do not contribute to theory development in any meaningful way. Rather, they rely on previously established generalizations that are found applicable to the context of interest and in this way help shed light on the case at hand.

Eckstein calls this a "discipline-configurative" type of studies (2000 [1975]: 134). The case is explained by subsuming it under well-established propositions: the outcome observed is in fact inferred by deduction from the extant theory and a set of specified antecedent conditions. According to Eckstein, the explanation of the case (in his words, the interpretation) is successful if it is logically *compelled* by the theory: one should be able to demonstrate that, given the regularity and the characteristics of the case, the outcome must have occurred or had a high probability of doing so (136).

3. Hypothesis-generating case studies

Lijphart asserts that hypotheses-generating case studies are selected for purpose of theory-building. They rely to a very limited extent on previously developed but vague generalizations. The purpose of these studies is to contribute to the formulation of new generalizations in areas where no theory, or very unsatisfactory theory, exists. The generalizations thus formulated ought then to be tested subsequently by further empirical inquiry.

In Eckstein's taxonomy, hypothesis-generating case studies are called "heuristic" (2000 [1975]: 137). He remarks that there is a good track of records of case studies used for heuristic purposes which typically consist in discerning important general problems and possible theoretical solutions. One reason why this might be the case is that case studies consist in the intensive analysis of a subject, do not restrict the inquiry to a limited set of variables, and in so doing increase the probability that new variables and critical relationships among them be discovered (138).

4. Theory-confirming case studies

Theory-confirming case studies are developed within the framework provided by well-developed theory. The general purpose of this type of studies is testing theoretical generalizations by providing evidence in support of the extant theory. Lijphart maintains that the theoretical value of these studies is limited if the theory has been confirmed already by a large number of cases. In other words, if the theory obtained already substantial evidence in its support, the confidence in the hypothesis is only marginally increased by one additional positive instance. The theoretical value of these studies is however enhanced when the studied cases turn out to be *crucial* (1971: 692).

Crucial cases are characterized by "extreme values on one of the variables" and thus constitute a "crucial test of the propositions" (ibid.). Lijphart is not very specific in his definition and does not explain why crucial cases are so important for testing. In general, crucial cases are such that one would expect from them either a strong refutation or a strong confirmation of the theory. The former, what Eckstein and Flyvbjerg call "least-likely" cases, are especially useful in theory confirmation. A least-likely case, in fact, is such that, given the characteristics of the case, one would expect a strong rejection of the theory. If the study of a least-likely case eventually *confirms* the theory, the confidence in the theory is remarkably increased, or at least it increases more than it would if the case was not crucial.

5. Theory-infirming case studies

Similar considerations as above extend to theory-infirming case studies, which are also developed within the framework of a well-developed theory. They test theoretical generalizations and provide negative evidence for it that is, evidence that refutes the theory. Also in this case, if the theoretical proposition is solidly based on a large number of cases, the theoretical value of one additional infirming study is limited as it decreases the confidence in the theory only marginally.

Furthermore, the infirming power of the case study is higher if it uses a *crucial* case. A crucial case is useful for a theory-infirming study if one would expect from it a strong confirmation of the theory. This type of crucial cases is called by Eckstein and Flyvbjerg "most-likely". Most-likely cases are such that, given the characteristics of the case, one would expect the theory to be strongly supported in the circumstances. If the study of a "most-likely" case eventually *refutes* the theory, the corresponding decline of confidence in the theory is particularly severe or, at least, more severe than the decline one would have if the case was not crucial.

6. Deviant case studies

Finally, deviant case studies select cases that are known to deviate from well-established generalizations; they are known as outliers (Flyvbjerg 2006). They are used to develop theory further either including new variables or refining the variables in use by disclosing why the case at hand is deviant. They thus weaken the original proposition, but replace it with a modified version of it, to be tested against new cases.

Even though the scholars mentioned above share the same organizing principle for their taxonomy, they see the comparative strength of the CSR in different *types* of case studies. Lijphart seems to emphasize their heuristic power while acknowledging that they are weak in terms of testing power when contrasted with the other research strategies in the hierarchy. In this position, he is not alone. It used to be conventional wisdom among social scientists that case studies are most valuable in the context of discovery rather than justification, and this view is still endorsed in more recent contributions (Gerring 2004, 2007a).

Eckstein holds instead a rather unconventional view. He defends the idea that the comparative strength of the CSR resides in the testing power of crucial cases. In his view testing a theory is an *effort to falsify*, and such an effort is most profitable when most-likely and least-likely cases are identified (Eckstein 2000 [1975]: 146).[15] Despite the differences in emphasis, these scholars can be understood as adopting a similar view on CSR, and on science more in general: science is seen as a generalizing endeavor aimed at explaining phenomena of interest by formulating empirical regularities. In this perspective, they share and try to solve the same fundamental puzzle about CSR: how possibly do "N = 1" types of study contribute to science so understood in any meaningful way?

1.3 THE NEW PROJECT: CSR AS AN AUTONOMOUS EPISTEMIC GENRE

An alternative view of CSR is, however, on offer. It started taking shape recently in the methodological debate on qualitative research and its relationship with the quantitative tradition that followed the publication of *Designing Social Inquiry* by King, Keohane, and Verba in 1994 (Brady and Collier 2004, George and Bennett 2005). It is also traceable in isolated contributions that either preceded or lie at the margin of that debate (Creager et al. 2007, Flyvbjerg 2006, Forrester 1996, Geertz 1973, Morgan 2012, Vennesson 2008). These authors do not belong to the same philosophical tradition. Nevertheless, they contributed maybe unintentionally, and to different extent, to what can be seen as the same project, that is, liberating CSR from the statistical, comparative, and experimental mindset. They tend to emphasize the distinctiveness of CSR and the difference in logic and purpose, rather than the similarity, with respect to the other research strategies. Thus, at its core this view can be regarded as different from the approach sketched above in the recognition of CSR as an autonomous *style of reasoning* which, as such, needs to be evaluated on its own terms.

[15] See Flyvbjerg (2006) for a contemporary criticism of the conventional wisdom on the lack of testing power of case studies.

The notion of "style of reasoning" is used by Ian Hacking to refer to six methods which historically became a permanent component of science (1992). Hacking adopts the definition by historian of science Alistair Crombie who explains ostensibly the notion by pointing to six styles. In the formulation by Hacking they are:

> a) The simple method of postulation exemplified by the Greek mathematical sciences.
>
> b) The deployment of experiment both to control postulation and to explore by observation and measurement.
>
> c) Hypothetical construction of analogical models.
>
> d) Ordering of variety by comparison and taxonomy.
>
> e) Statistical analysis of regularities of populations, and the calculus of probabilities.
>
> f) The historical derivation of genetic development (1992: 4).

Hacking is not interested in the specific way in which every style came into being. Every style, in fact, became independent of its own history to be what we now regard as a "rather timeless canon of objectivity, a standard or model of what it is to be reasonable about this or that type of subject matter" (1992: 10). Rather, he is concerned with what makes of a certain method a style of reasoning. The necessary condition for being a style of reasoning is the introduction of novelties in the form of objects, evidence, sentences, modalities, and possibilities.

Thus, in Hacking's account each style should introduce novelties of most or all of the listed types, and should do so in an open-textured, ongoing and creative way. He gives the example of mathematicians who "do not just introduce a few sorts of abstract object, numbers and shapes, and then just stop. The type 'abstract object' is open-ended once we begin reasoning in a certain way" (1992: 12). In addition to the ostensive explanation and necessary condition, Hacking suggests that what *constitutes* a style of reasoning is a set of techniques of *self-stabilization*: it is by way of these techniques that each style persists in its peculiar and individual way (1992: 16).

John Forrester proposed to extend the notion so as to include "reasoning in cases" as a seventh style on top of the six listed by Hacking (Forrester 1996: 2). Forrester motivates his proposal with the purpose of characterizing the dominant style of reasoning in psychoanalysis and the related sciences since the early Twentieth century. His project is, however, more ambitious than that. Reasoning in cases is not only typical of fields such as psychiatry, psychology, and criminology; rather, there is a methodological continuum from the psychoanalytic case, to the case used as a pedagogical tools in certain academic traditions, to the style of reasoning that is typical of the man of practice.

Thus, Forrester launches a unifying philosophical project that he leaves, however, unfulfilled. To some extent his line of thought is followed up more recently by Mary Morgan. Morgan goes back to Hacking's conception of styles of reasoning, which she re-labels *epistemic genres*. She writes in defense of case studies as an epistemic genre, and legitimizes her project by citing Hacking's view as follows: "It was a matter of history that each epistemic genre developed its own generic way of finding and validating knowledge, so that work within that genre came to be judged within that epistemic genre and by its community of practitioners, not according to the rules, or in the terms, of any other genre" (2012: 671). It then follows that if CSR constitutes indeed an epistemic genre similar considerations should be extended to it as well.

This view of CSR sits comfortably in a conception of science rather different from the traditional view presented above. This alternative approach is not as coherent and uniform as the one to which it reacts is; it hosts in fact a range of positions which, however various, challenge the old paradigm in similar respects. Generally speaking, the new approach sees science as a much less homogeneous practice than the old view would have it. In particular, it challenges to varying extents the four tenets presented above (see sec. 1.2). The first tenet to be disputed is:

a. The (indiscriminate) characterization of science as a *generalizing* epistemic practice.

Objections to the first tenet come in different forms and are not all equally radical. Some scholars understand this statement as referring to universal generalizations and point out that some part of science is “content” with generalizations of a more limited scope. This is the milder objection to (a). Other scholars understand the first tenet as having normative import to the effect that general theoretical knowledge is the only (scientific) knowledge there is; or, less radically, is more valuable than concrete context-dependent knowledge. These scholars reject this normative stance, argue that concrete context-dependent knowledge is as valuable, if not more, than theoretical knowledge, and that its value is proved by the crucial role this type of knowledge plays in human learning (Flyvbjerg 2006). Still other authors entertain a radically different view of science, or at least some fields in it. They see some part of science as a case-based epistemic activity where *specific instances* of scientific practice in the form of *case studies, models, exemplars,* occupy a central role in learning and research strategies (Creager et al. 2007). Related to this is the challenge to the second tenet:

b. The primary aim of science is the explanation of scientific phenomena.

The new view tends to de-emphasize the prominence of explanation as the unique, and/or most important, goal in science. On the one hand, it rehabilitates description as an important goal on its own decoupled from explanation and not just preliminary to it, as the old paradigm would have it. On the other hand, it introduces the idea of “thick description” that is providing accounts of situated phenomena which are rich in detail and faithful to the complexity of the studied subject. The ultimate goal is not to explain the specific by subsuming it under a general law; it is rather providing understanding of the specific case by revealing its complexity. Description so understood is pursued for its own sake and is not seen any longer as mere “data gathering” functional to subsequent explanation. Furthermore other purposes of science are more explicitly contemplated. Some authors see in policy making one of the main goals of scholarly research. The idea behind this view is that scientific inquiry should lead to the production of knowledge that is useful ultimately for the policy maker.

c. The only generalizations relevant to science are empirical regularities.

Several authors object to tenet (c) by pointing out that general knowledge of scientific interest does not only come in the form of empirical regularities but of causal processes and mechanisms (Coleman 1986, Stinchcombe 1991). The debate on the concept and use of causal mechanisms in the social sciences is too vast to do justice to it in a few lines. This debate has grown massively in the last years, with contributions from many social scientists and philosophers of science.[16] The point of interest here is that these scholars defend the importance of mechanisms vis-à-vis regularities in the philosophical and methodological discourse about science. They defend this claim on the following grounds:[17] mechanisms are distinct from regularities; the epistemic strategies conducive to the discovery of mechanisms and regularities are different; mechanisms play an equally, if not more, important role than regularities in the explanation of social phenomena.

d. Discovery and justification are distinct temporal processes.

Even though, as noticed above, strong objections have been raised against the version of the DJ distinction that sees justification and discovery as distinct temporal processes (see footnote 5), traces of it still survive nowadays in several contributions. The general idea behind the rejection is that in the practice one cannot separate the moment and process by means of which hypotheses are constructed from the moment and process by means of which they are tested. If this is indeed the case, CSR cannot be said to contribute either to the context of discovery or to the context of justification (Vennesson 2008).

1.3.1 CASE, CASE STUDIES, AND CASE STUDY RESEARCH

The new perspective leads to re-conceptualizing the notions of case, case study, and case study research. Definitions that were typical of the traditional view are dropped. In key contributions to the old view, case, case study, and case study research are defined as follows:

[16] Prominent contributors to the debate on the role of mechanisms in the social sciences are Bunge 1997, 2004, Elster 1989, 1998, Hedström and Swedberg 1998, Hedström and Ylikoski 2010, Little 1991, 1998, Mayntz 2004.

[17] This is a gross generalization. The philosophical positions in this debate are almost as numerous as the contributors. My purpose here is just conveying the sense of what are the issues at stake.

> CASE: A phenomenon for which we report and interpret only a single measure on any pertinent variable (Eckstein 2000 [1975]: 123-4).
>
> CASE STUDY: the study of a single case.
>
> CASE STUDY METHOD:[18] is the method applied to the study of a single case, rather than many as in the statistical method, or just a few as in the comparative method (Lijphart 1971: 691).

These definitions seem not only strongly influenced by a statistical viewpoint but also hardly informative, and inaccurate. Certainly, there is much more to a case study than just *reporting* and *interpreting*, whatever that means in this context, a single measurement on any pertinent variables; furthermore, it is also disputable that the case study only reports a *single* measurement on these variables, rather than many (George and Bennett 2005). Even though the old definitions are rejected and replaced by new ones, there is not agreement yet on a definition of case, case study, and case study method within the individual disciplines, let alone across them. The following definition by political scientist Pascal Vennesson are probably quite representative of the average feeling. Not all features they include, however, would be regarded as *defining* case study by the scholars in this new trend.[19] Disagreement on basic concepts shows that the new trend has not developed yet into a fully-fledged view.

> CASE*: A phenomenon, or an event, chosen, conceptualized and analyzed empirically as a manifestation of a broader class of events.
>
> CASE STUDY*: is a research strategy based on the in-depth investigation of one, or a small number, of phenomena in order to: (i) explore the configuration of each case, and (ii) elucidate features of a larger class of (similar) phenomena, (iii) by developing and evaluating theoretical explanations.

[18] Here and throughout the thesis I treat "case study research" and "case study method" as synonyms. I preferred to use "case study method" in this context to be faithful to the original discussion.

[19] I will discuss these features below.

One can find similar definitions in George and Bennett (2005), Gerring (2004, 2007a) and Ragin (2000). These definitions deserve some comments, and some refinements.

In the first place, *case* is a theoretical construct: it is not the specific phenomenon or event, but the conceptualization of it. Something such as a phenomenon or event in political science, an individual in psychology and psychiatry, a cultural practice in anthropology is *turned into* a case by the study of it. Forrester, citing Foucault talking about the development of the clinical sciences, says that the examination, surrounded by all its documentary techniques, makes each individual a 'case' (1996: 12).

The above definition is, however, too restrictive when it defines the case as the manifestation of a broader *class* of events. The theoretical work behind the selection and analysis of cases is often much more fluid. The class of reference is often not established at the beginning of the research, can change along the process, or emerge at its conclusion. Sometimes it might even remain indeterminate.

Furthermore, the purpose of the study does not need to shed light on a larger class of elements, but might simply be the exploration of the single case. This does not imply that generalizations from a case study are not possible or even undesirable. Rather, the other cases to which the results can be generalized are often identified *once* the case study has been accomplished. In fact, results from previous case studies may turn out to be applicable to new cases in an unexpected or unplanned way. In a similar vein, case studies do not always have such a strong and explicit theoretical motivation. Thus, the purpose of the study does not need to be the development and evaluation of a theoretical *explanation*. It might just be the characterization, or description, of the case at hand. In this sense, features (ii) and (iii) above should not be regarded as a defining feature of case studies but as additional epistemic benefits of this research strategy.[20]

Finally, case study and case study research (or case study method) should be treated as distinct notions. The definition of case study given above refers in fact to the research

[20] In particular, that feature (ii) is indeed an epistemic benefit of CSR will be argued in the next section and in Chapter 3.

strategy, hence to case study research. The case study is the (written) *output* of this research strategy when employed to study one or a few cases.

Morgan provides a characterization of case studies that draws on the practitioners' literature but at the same time emphasizes some of the aspects that to some extent fail to be present in their definitions. She identifies the following as characteristic features of this research strategy (Morgan 2012: 668):

I. Boundedness: a case study investigates a *bounded whole* object of analysis;

II. Open-endedness: the boundary between subject of analysis and context is not clear at the start of research and may remain fluid during the study;

III. Depth: the case study creates a considerable depth of engagement with the subject and dense evidential materials across a range of aspects of the topic;

IV. Multi-method: many potential research methods may be used within the case study.

V. Complexity: the outcome is a complex, often narrated, account that typically contains some of the raw evidence as well as its analysis.

The pressing issues that CSR confronts can be understood in terms of *validity* and *relevance*.[21] I define and discuss validity and relevance in the sections that follow. First I distinguish between *internal* and *external* validity. Rather than assessing CSR in terms of its heuristic or testing power, and try to establish whether CSR contributes to the context of discovery or justification, what matters here are the conditions under which CSR can obtain internal validity, and what criteria are adequate to assess whether a given case study is in fact *internally valid*. Furthermore, rather than assessing whether and to what extent case studies contribute to the formulation of empirical regularities, the broader question of interest here is how one *generalizes* from a given case study. As I shall argue below, there are different strategies to generalize in CSR, one such a strategy is by producing results that are *externally valid*. Finally, once one acknowledges that the social sciences can aim at

[21] Whereas the notion of validity has been used quite extensively in the literate on CSR (see, for instance, Gerring 2007a, Morgan 2012), to the best of my knowledge the notion of relevance was never explicitly adopted and discussed.

a variety of purposes all equally worthy, the pressing issue is under what conditions a case study provides results that are *relevant* vis-à-vis the purposes at stake.

If CSR is to be treated as an epistemic genre, as the experimental and statistical methods are, then it has its own way of finding, validating, and, I would add, generalizing knowledge. The challenge that lies ahead for those scholars that see in it an autonomous style of reasoning is to understand which way this is.

1.3.2 INTERNAL VALIDITY AND EXTERNAL VALIDITY

The concept of validity refers to the correctness of scientific claims. The methodological literature typically draws a distinction between *internal* and *external* validity. Scientific results are *internally* valid if, and only if, the inference from the evidence to hypothesis within a given study is correct (Reiss 2008). The internal validity of a study, thus, is compromised when throughout the inferential process from evidence to hypothesis mistakes are made. Mistakes might be due to fallacious reasoning, to the violation of the assumptions, or to an incorrect measurement procedure. Internal validity is a scientific ideal since, as a matter of fact, it is very hard to control for all sources of error in the inferential procedure, and have adequate knowledge that the relevant errors have been in fact controlled. Within a given study, one *approaches*, rather than attains, internal validity. The extent to which internal validity can be achieved varies from method to method. Certain methods have higher chance to deliver conclusions that are internally valid because of the type of inference they rely on (deductive rather than inductive) or because they have strategies to ensure that the assumptions are satisfied.

When discussing internal validity in relation to CSR one can tackle the issue from two points of view. On the one hand, one can ask *how specifically* CSR produces results that are internally valid. This strategy amounts to establishing what conclusions can be reached vis-à-vis the evidence at hand, how, and with what degree of confidence. It thus examines the *techniques* by means of which conclusions are inferred and aims at identifying under which conditions these techniques are *correctly employed.* These are *conditions of internal validity* and pertain to the inferential procedure from evidence to hypotheses. On the other hand, one can formulate an assessment of the case study based on criteria other than inferential correctness. Rather than examining the techniques by means of which the evidence is analyzed and the conclusion derived, one would focus on features of the narrative itself. This way typically abstains from judgments of *correctness*; rather, it imposes criteria, or standards, on the narrative as various as coherence, credibility, persuasiveness, or significance, and use those as indicative of the validity of the study. The two approaches coexist in the literature on CSR. Proponents of either approach regard their strategy as conducive to ascertain the internal validity of CSR. Both, however, are subject to specific challenges.

In the former case, the difficulty arises from the fact that CSR relies on the use of several techniques that are often employed jointly. Upon examining the assumptions, inferential procedure, and type of evidence proper of each technique, one thus needs to formulate the conditions that are conducive to correct conclusions, and the degree of confidence one can have in conclusions thus reached. Furthermore, one needs to specify how the techniques are used correctly when employed jointly, and how the joint use of these techniques affects the confidence in the final conclusion. The techniques most commonly used for analyzing evidence in CSR are process-tracing and the method of comparison.

The method of comparison shares the same "logic" that underlies the comparative method discussed in section 1.2.1.[22] The method of comparison refers to a *technique* of data analysis that is employed in case study research for within-case and across-case analysis (CSR can in fact focus on the study of one or a few cases). In CSR the results obtained by the method of comparison are always integrated with other types of evidence to develop a complex narrative which the method of comparison could not possibly develop on its own. The comparative method is a research strategy that consists in the comparison of a (typically higher) number of cases to establish causal relations between relevant variables. We can thus regard CSR and the comparative method as distinct study designs which can

[22] This logic was first articulated by J.S. Mill in his description of the method of difference and method of agreement (1858).

both be directed to causal analysis. The former differs from the latter in that it draws on a variety of causal insights and techniques, among which the method of comparison, the results of which are eventually integrated into the narrative.

Whereas the method of comparison is well mastered and understood, process-tracing is a technique whose inner workings, despite broad usage, are not fully understood yet. Even though its conditions of validity are not fully articulated, it is characterized as directed to the identification of causal processes and the circumstances in which these processes unfold. The literature typically refers to the former as *mechanisms*, and to the latter as the *systems* in which the mechanisms are embedded. When process-tracing is used jointly with the method of comparison, the description of the mechanism, and of the conditions in which it is triggered, complements, enriches, and refines the causal relation that the method of comparison identifies. In Chapter 2, I will try to clarify what conditions of validity do apply to process-tracing and how, and to what extent, the method of comparison and process-tracing when employed jointly increase the plausibility of the general conclusion.

An alternative, and perhaps more controversial, way to assessing validity in CSR is by imposing criteria on the final output (Morgan 2012). Typically, the final output is a narrated account which has a very complex form and "contains some of the raw evidence as well as its analysis and ties together the many different bits of evidence" (2012: 668). Standards can therefore be established on how the different bits of evidence should tie together, or on specific features of the resulting narrative. Inspired by Neil MacCormick's work on legal cases, Morgan suggests that valid case study accounts exhibit the following criteria: consistency with all the evidence found, coherence within the account (the bits of evidence fit together), and credibility of the explanation in social scientific terms (2012: 674). This approach to internal validity does not find full consensus in the scientific community. While standards are acknowledged as leading to valuable case studies, they are also criticized for being either too loose or simply inadequate to ensure rigor (Geertz 1973, Morse et al. 2002). Whether they are in fact conducive to internal validity is therefore still open to discussion.

Scholars belonging to the old tradition worried about the scientific usefulness of case studies because of their lack of generalizing power. Since this tradition sees science primarily as aiming at finding regularities that explain phenomena of interest, it is hard to appreciate how the study of a single instance can help formulate broad generalizations of this sort. This is, however, a very specific way to characterize what is a much more complex issue: *generalizability* pertains to scientific findings whose scope is more encompassing than the subject of the study. Addressing generalizability requires one to specify (i) *what* should be generalized from a given study, (ii) *to what* such a result should be generalized, (iii) and *how*. Different answers can be given to these questions. I distinguish three approaches that differ in the answers given to each of the three questions above. In the first two cases, generalizability can be understood in terms of *external validity* and what is actually generalized is an empirical finding; in the third case, the generalization is theoretical.

Scientific conclusions are *externally valid* if they are correct in the studied context and in other contexts yet unstudied. Thus, whereas internal validity is obtained by way of correct inference from evidence to hypothesis within a given study, external validity is obtained by way of correct inference from the results obtained in a context studied directly (the original context) to a hypothesis about a new context yet unstudied (the target context). As such, external validity is not only achieved if one formulates general conclusions that are universally, or almost universally, valid. It is obtained any time scientific conclusions about a target context are inferred correctly from results that were established in the original context; that is, they are *extrapolated* correctly.

Under the broader umbrella of external validity, the scholars of the old school provided a possible answer, very specific indeed to the three questions posed above. Any study should establish an empirical regularity to be generalized, if not universally, then to a broader class of cases of which the studied case constitutes a representative instance. This view on generalizability by no means belongs only to the traditional view. It is still very present in the contemporary debate on CSR, even though in a slightly revised form as one can easily tell from some very common definitions of case study actually in use.[23]

However, that the conclusions be generalizable to a broader class of elements is not the only modality for CSR to achieve external validity, or any other design for that matter. Successful extrapolation from case to case is another modality that is, according to some scholars, more amenable to CSR (Forrester 1996, Creager et al. 2007). This alternative approach gives different answers to each of the three questions posed above. As to the object of the generalization, it is not the empirical regularity as such, but the causal mechanism responsible for the phenomenon of interest and the causal conclusion it elicits. The target of the generalization is another or a few other cases where similar phenomena are observed. Finally, the inference is made by analogical reasoning. The studied case is not regarded as an instance that is representative of a broader class of elements but is, more generally, a situated phenomenon that is *similar* to others in some relevant respects.[24]

In the third approach, the case study is seen as the occasion for formulating conceptual and theoretical generalizations rather than empirical results to be directly applied to other cases, be they classes of elements or concrete instances. Robert Yin describes this alternative as a form of *analytic generalization* and contrasts it with the statistical generalization typical of other research strategies. In CSR "the mode of generalization is analytic generalization, in which a previously developed theory is used as a template with which to compare the empirical results of the case study" (Yin 2003: 33). In a similar spirit, Clifford Geertz talks of clinical inference, or inference *within cases*. Rather than beginning with a set of observations and attempting to subsume them under a governing law, such inference "begins with a set of signifiers, or symptoms, and attempt to place them within an intelligible frame" (1973: 165). The task is constructing an analytic system in whose terms what is generic to the studied case will stand out against the other determinants of human behavior. The generality the case study contrives to achieve, Geertz suggests,

[23] See section 1.3.1 above for my discussion of some recent definitions of case study; see Chapter 3 for a thorough discussion and criticism of what I call the traditional view on external validity.

[24] It can be argued that the former approach discussed above is interpretable to some extent as a special case of the latter, more general, approach.

"grows out of the delicacy of its distinctions, not the weep of its abstractions" (Geertz 1973).

In Chapter 3, I shall explore and defend the second approach to external validity. This line of reasoning about external validity seems in fact the most promising. On the one hand, it rescues CSR from the statistical viewpoint that informs the first approach and opens the way to a better understanding of CSR as an autonomous epistemic genre. Furthermore, unlike the third approach, it focuses on the causal conclusions that can be elicited from the case studies and carried over to other contexts. What is under scrutiny in this thesis is the capacity of case studies to generate valid causal knowledge under the assumption that this is the evidence we primarily need for policy making. In line with the general spirit of the present work, in the chapters that follow I only focus on the formulation and extrapolation of causal claims in CSR and omit analytic generalizations instead.

1.3.3 RELEVANCE

Case studies can be assessed in a further respect besides their internal and external validity, namely with respect to the *relevance* of their results. Valid hypotheses are those hypotheses that are regarded as *correct* by the extant background knowledge; relevant hypotheses are those conclusions that are *adequate* with respect to purposes. Validity claims thus differ from relevance claims first and foremost because they address distinct properties of scientific hypotheses: the former express judgments about the correctness of the inferential procedure by means of which conclusions are derived within the study, or outside of it; the latter express judgments on the adequacy of those conclusions to further purposes, fulfill tasks, or solve problems. As they target different properties of scientific conclusions, judgments of validity and judgments of relevance are based on different sorts of considerations.

Validity claims are directed to evaluate whether a hypothesis is justified by the evidence: matter for this the inferential logic of the method employed and the assumptions on which it works. Relevance claims are directed to evaluate whether the results delivered by a given method are epistemically adequate to try to solve and achieve the problem and purpose in question. This type of analysis presupposes:

1) The definition of the purpose/problem at stake;
2) The characterization of the *kind* of conclusions required;
3) The assessment of what kind of results a *given* method provides.

Relevance claims are justified if 1), 2), and 3) are also justified. This implies that they can be contested by impugning either 1), 2), or 3). One can impugn 3) by objecting that the conclusion licensed by the method in question is not exactly the type of information that is required to fulfill the purpose of interest as specified in 2). For instance, if control is the goal of interest and one regards counterfactual claims as necessary to this end, one might object to the relevance of case study evidence because it corroborates singular causal hypotheses that cannot support counterfactuals. Or, one can impugn 2) by objecting that the specified kind of scientific result, even if actually provided by the method in question, is not adequate for the purpose or problem as defined in 1). For instance, one might object that counterfactuals are not the type of claims one need for control purposes, after all, and that one needs evidence of causal processes instead.

Alternatively, one can impugn 1) by arguing that the problems and purposes as actually defined in 1) are either illegitimate or incorrectly posed. Purposes and problems can be defined at a rather abstract level, for instance by referring to the fourfold distinction between description, explanation, prediction, and control, or at a higher level of detail. When considering control, one might define it as the formulation of policies to bring about intended effect in the social realm.[25] And one might object to this aim as an illegitimate goal of science on the ground that science should stay aloof from all that concerns politics, policies and power. Alternatively, one might specify in greater detail the type of problem, policy, or yet effect of interest, for instance, by suggesting that the aim of economics is erasing poverty in developing countries and leveling out inequalities. The objection to purposes defined at this level of specification is eminently an empirical matter, and draws

[25] See section 1.3.3.1 for a similarly general, though more precise, characterization of policy making.

to a large extent on factual considerations about the context in which the purpose at hand is to be reached or the problem solved.

Even though validity and relevance are distinct concerns, relevance claims do relate in fact to validity claims. In particular it seems reasonable to argue that when assessing the kind of results that a given method provides (point 3 above) one should only pay heed to those results that are reached by way of a valid inference. In this sense, relevance *presupposes* validity. Validity is, however, far from sufficient for relevance. In particular, relevance claims draw on a much broader set of considerations than the validity of those scientific conclusions whose relevance is ultimately at stake. This set of considerations encompasses among others formulating the aims which can legitimately be pursued by way of a scientific inquiry, and assessing which methods actually help further those aims.

1.3.3.1 CASE STUDY RELEVANCE FOR POLICY MAKING

Besides the traditional aims of scientific research, such as description and explanation, some contributors to the recent debate on CSR have started considering more seriously control. Case studies thus seem to be methodologically interesting not only for their descriptive and explanatory virtues, but also as evidence for policy makers. As a matter of fact, case studies *are used* for intervention purposes in several fields. In medicine, psychology or psychiatry the prominent goal of the clinical study is the treatment of the patient both directly and indirectly by building clinical knowledge to be used in future treatments; similarly, social and economic policies implemented by regional and international institutions are also largely built on case-based knowledge.

There are specific features of case studies that seem to render them relevant to policy making on an intuitive ground. The case study can be in fact characterized as providing *contextual*, *concrete*, and *processual* evidence.

In the first place, it is an investigation into phenomena that are circumscribed in time and space, where the boundary between context and subject of investigation remains fluid during the research (Morgan 2012).[26] The output of this procedure is thus a characterization of the phenomenon that is left permeable for contextual information. Furthermore, case studies are thick descriptions that use dense evidential material. This type of analysis is extremely rich in detail and, in this sense, formulated at a relatively high level of concreteness. By focusing in fact on localized phenomena, the case study can afford to retain information that ought instead to be sacrificed in the process of abstraction that more general analyses demand. Finally, in certain circumstances case studies explain by describing the causal process, or mechanism, by means of which certain effects are brought about.[27] In particular, this form of explanation amounts to identifying the conditions, the factors, the events, and their configuration that were responsible for the phenomenon of interest.

Thus seen, case studies are relevant to policy making understood as a practice aimed at producing intended effect in the social world by manipulating some feature of it.

First, policies operate in contexts: they produce the intended effect in conjunction with other causal factors. Using a more technical language, policies are INUS conditions, that is, they are an insufficient but non-redundant part of a condition which is itself unnecessary but sufficient for the occurrence of the outcome of interest.[28] Consider an example used by Nancy Cartwright and Jeremy Hardie to illustrate this point (Cartwright and Hardie 2012). California implemented a project to improve the academic achievement of its school pupils by reducing class-size in the mid90s. Class-size reduction is an INUS condition in Mackie's account. On its own it does not contribute to the effect of interest, as shown by California's experience, unless supported by other factors. For instance, availability of space and quality of teachers are other non-redundant but insufficient *parts* which, in conjunction with reduced class size, increase students' proficiency. In this sense, thus, policy making is to a large extent a *situated* practice: when involved in actual decision making policy makers target contexts in their specificity. From this follows the relevance

[26] See point II in section 1.3.1.

[27] They can do so by employing process-tracing as a technique of causal analysis (see section 1.3.2).

[28] The term 'INUS conditions' is due to Mackie (1988).

of evidence such as case studies that is *contextual* and thus provides clues to what other factors besides the policy of interest need to be present for the outcome to occur.

Furthermore, policy makers try to sort the intended effect by modifying certain features of the target contexts, by erasing existing practices, or by creating new ones. In other words they *manipulate* the context. Manipulation requires detailed information about what is exactly to be modified and how it can be modified: policy making needs thus to be informed by descriptions fine-grained enough to enable the decision maker to have an answer to this type of questions; this is why evidence that is *concrete* plays an important role in it. Finally, the idea of manipulating certain aspects of reality *in order to produce* the intended effect presupposes causal justification. Policy making so understood is rational only if supported by causal beliefs that is, by the confidence that there is some relation between the object to be manipulated and the outcome of interest such that the former can be exploited to modify the latter. Processual knowledge is causal knowledge of a specific type which, as such, is prima facie relevant to this end.[29]

Even though the idea of case studies as relevant to policy making seems to have some appeal, the debate among philosophers and social scientists is still poorly developed. Several research avenues have, however, been opened and might be worth exploring. As early as the mid-70s, leading political scientist Alexander George worried about the gap between policy makers' needs and the orientation dominant in scholarly works. He is concerned about the conditions present in the field of international relations at that time and fears a bias in current research towards *grand theories* at the expense of research with higher practical relevance (George 1976, 1994, 1997).

George takes a modest though challenging stance. He admits that the gap between policy making and scholarly research cannot be eliminated. Political, financial, ethical, and practical considerations, in fact, inform decision making besides the factual knowledge that science provides. Nevertheless, George urges, the gap could be *bridged* by scholarly knowledge that fulfills the standards for relevance. In sum, scholars face a choice between

[29] Relevance as defined in section 1.3.3 presupposes validity to some extent. In Chapter 2 I will characterize the conditions in which case studies are likely to deliver *valid* causal knowledge.

taking into consideration the need of policy makers and ignoring it while being driven by purely scholarly interest.

The plea for more relevant research by field specialists is voiced again and articulated further in the later work by George and Bennett (2005). They remark that the *grand theories* of scholarly research in international relations are generalizations broad in scope and probabilistic in form; as such, this form of knowledge is insufficient to guide the policy maker in the choice of the strategy to implement in target contexts. It needs to be integrated by *usable knowledge,*[30] which consists of propositions that, though narrower in scope, specify the conditions under which these limited generalizations hold. In this way *usable knowledge* can be guide to action because it allows the policy maker to better explain the opponent' actions and make a diagnosis of whether the extant conditions are adequate for the adoption of a given strategy. George and Bennett see case studies as a fruitful instrument for the production of *usable knowledge.*

The fruitfulness derives from the case study, which has to identify the causal mechanism by means of which explanatory variables bring about the outcome of interest. The focus on mechanisms, as opposed to broad generalizations, has two distinctive benefits. First, the conditions under which the more modest generalization holds will be clarified. Secondly, focusing on mechanisms yields a description at a lower level of abstraction; in this way, they would identify variables that are likely to be more easily manipulated, and hence of utmost interest to the policy makers.

The proposal by George and Bennett is certainly insightful and praiseworthy. They make an effort to spell out what the positive contribution of case studies can be to policy making once the policy maker's needs and the specificity of CSR are taken into consideration. It leaves, however, much unspecified. Policy making is a complex practice organized in various stages each of which relies on a variety of evidential inputs. The role of case studies can be clarified further by specifying at which stages they are relevant, why, and how they complement the other sources of evidence.

[30] George and Bennett borrow the term *usable knowledge* from the work by Lindblom and Cohen *Usable Knowledge, Social Science and Social Problem Solving* (1979).

In this effort, social scientists and philosophers working on case studies would be assisted by more developed strands of literature on policy evaluation and evidence-based policy. As to the former,[31] several contributors emphasized the specific and crucial role of case studies with respects to other forms of evaluation (Eck 2006, Scriven 2008). In the discussion on evidence-based policy, CSR still occupies a minor role if compared to other more popular strategies of research.[32] This does not come as a surprise when one considers the fact that case studies lie at the bottom in the hierarchies of evidence formulated by most institutional evaluators. The scholars who see the potential of case studies in policy making should thus feel the urge to engage in this debate as well. This would force them to clarify further why case study evidence is distinctive, how it differs from evidence generated otherwise, and how exactly it complements it.

1.4 CONCLUSION

Understanding CSR as an autonomous epistemic genre, rather than a (weaker) form of the statistical or experimental methods, adds complexity to the analysis of case studies while making room at the same time for new research avenues. This perspective draws attention to the striking features of CSR and leads to a re-conceptualization of the old definitions and problems. What was simply understood as a problem of control is thus re-framed as an issue of internal validity, of which control constitutes just one specification. Similarly, the generalizability potential of CSR is not limited to the formulation of regularities that apply to the class of elements of which the case constitutes a representative instance. The forms of possible generalizations from, and within, case studies are various in content, target, and inferential modality. Finally, the discussion of the relevance of the CSR is also liberated from the old distinction between the heuristic and justificatory moment of scientific inquiry. In the old view the usefulness of the case study was in fact determined by the degree of support it gives to the theory at hand. Possible uses of case studies were either the formulation of new theories or the testing of extant hypotheses. In

[31] Reference point in this literature is Pawson (2006), and Pawson and Tilley (1997).

[32] As an example of very recent and important contribution to the field see Cartwright and Hardie (2012).

a broadened perspective other uses of case studies become apparent and their relationship with theory, now more broadly understood, is accordingly more complex.

The chapters that follow are motivated by these considerations. They are informed by a view on CSR as an autonomous epistemic genre and address the methodological problems case studies confront by exploring some of the avenues outlined above.

In Chapter 2, I address internal validity in historical narratives. Historical narratives are case studies that aim to formulate and substantiate causal hypotheses by articulating descriptions of the sequences of events leading to the outcome of interest. They typically make use of process-tracing to draw causal inference, and often rely on the additional use of the methods of comparison. Despite the important role of historical narratives in the social sciences, how process-tracing operates in the narratives is still poorly understood. The debate on process-tracing in fact, even though it is growing thanks to a number of recent contributions, is still muddy and under-developed. In particular, there are no shared criteria to assess its epistemic contribution; moreover, the conditions proposed so far tend to tie the validity of the findings to the use of specific kinds of evidence and are thus unhelpful when this specific evidence is not available.

I argue that the proposed conditions are unduly restrictive and fail to acknowledge the actual contributions which process-tracing can offer to valid causal inference. I formulate new conditions to assess process-tracing performance in cases in which the favorable evidential circumstances do not occur and existing criteria fail to apply. By discussing process-tracing, I touch upon what is currently a hot topic in philosophy of science; that is the role of causal mechanisms and mechanistic knowledge in the sciences. The majority of scholars who are interested in process-tracing also agree that it somehow helps the detection of causal mechanisms. The agreement, however, stops there: scholars entertain many different notions of mechanisms and disagree on the role mechanisms actually play in explanation and causal inference. From my discussion, it will become apparent how the complexity of this debate contributes to render our understanding of process-tracing even more problematic.

In Chapter 3, I address the problem of generalizability. I provide an outline of what I define as the traditional view on external validity. This approach is conditioned by a statistical viewpoint on CSR and reduces external validity to issues of mere representativeness.[33] In so doing it leads the debate on the generalizability of case-study results to a dead end as it quickly dismisses external validity as the downside of CSR. At the same time, it suggests that CSR is comparatively stronger in providing results internally valid. On this ground this approach recommends the use of case studies when internal validity is the main research goal of interest, while turning to other methods when one pursues generalizations instead. This outcome is unfortunate because, as a matter of fact, case studies are often performed with the explicit or implicit purpose of drawing lessons from the studied case to be carried over to new contexts yet unstudied.

I attempt to release this tension by examining the assumptions behind the traditional view on the external validity of CSR. Some of these assumptions have already been addressed, and actually disputed, in the current debate. In Chapter 3, I focus instead on those assumptions that, to the best of my knowledge, have not been addressed yet and seem to be responsible for the dead end in which the discussion among social scientists seems to be trapped now. In particular, I suggest that the debate should focus on how make case studies comparable rather than how select the typical case. Typicality and comparability are concepts closely related but distinct. The traditional view conflates the two and thus runs into confusion about what external validity is really about and how it can be addressed in a fruitful manner. I surmise that by enhancing the comparability of studies unnoticed room for improvement is made for formulating more reliable assessment of the external validity of results obtained in case studies.

In Chapter 4, I discuss issues of relevance when policy making purposes are at stake. In particular, I focus on the debate on the use and usefulness of randomized controlled trials (RCTs) to find the key to economic and social development. The participants to this debate agree that RCTs are affected by limited external validity, and that this impinges on

[33] This is the first of the three approaches to external validity I briefly characterize in section 1.3.2.

their usefulness for policy making. They diverge, however, on the strategies to overcome this problem. I analyze three alternatives that are found in the economic literature: replication of RCTs, which has been proposed by the promoters of RCTs; cross-country regressions, which have been typically endorsed by RCT-skeptics; and the causal models proposed by James Heckman. I argue that these strategies succeed in their attempt to a different, and limited, extent.

Proponents of the first two strategies fail to take into adequate consideration the distinction between external validity and relevance, and treat the latter as a spill-over of the former. Their strategies, in fact, aim to improve the external validity of causal effects on the assumption that relevance will automatically follow. I argue that this is not the case, because external validity and relevance are distinct concerns and should thus be addressed separately. The proposal by Heckman succeeds in delivering causal effects that are, as a matter of fact, more relevant to policy makers' purposes. I argue, however, that his model cannot adequately address the type of problems policy makers are likely to face in developing contexts. Whereas Heckman's model is equipped to face problems of *prediction*, in developing contexts policy makers face problems of *planning*. Planning is a complex procedure that depends on various pieces of evidence and raises several concerns. Causal effects are but one epistemic input in this procedure; case-study evidence is also relevant to the crucial phases of planning.

2 PROCESS TRACING AS AN EFFECTIVE EPISTEMIC COMPLEMENT

2.1 INTRODUCTION: FINDING MECHANISMS IN THE SOCIAL SCIENCES

During the last decades, philosophers of science and social scientists have promoted the view that knowledge of mechanisms can help causal inference in the social sciences (Elster 1989, Coleman 1986, Little 1991, 1998, Bennett and George 1997, George and Bennett 2005, Steel 2004, 2008). They have regarded covariation analysis based on statistical methods as subject to severe limitations, and proposed the identification of mechanisms as a promising strategy to overcome its problems. Their views differ, however, in the epistemic status given to mechanisms.

Some scholars consider them necessary for causal inference (Little 1991, 1998). Others regard them as insufficient on their own for this purpose, and recommend the joint use of mechanistic and statistical evidence (Bennett and George 1997, George and Bennett 2005). Still others maintain that there is no general demand for mechanistic knowledge; whether we need it depends on contextual factors and on substantive background knowledge (Kincaid 1996, Steel 2008). Setting this dispute aside, mechanisms can only play a significant role in causal inference if social scientists have means to identify them correctly. Some scholars suggest that the method of process-tracing might be helpful in this respect (Bennett and George 1997, George and Bennett 2005, Steel 2004, 2008).

Process-tracing is a widely used strategy of causal inquiry in the social sciences. Despite the ample usage, a formalized template of process-tracing is not available yet (Gerring 2007a), and there is no shared understanding of how it works, and what its strengths and weaknesses are. Scholars tend to agree that it is a method of within-case causal analysis that helps causal inference by detecting mechanisms, and in these respects it differs essentially from the other empirical strategies in the social sciences. This superficial agreement hides deeper differences, however. Views differ in their definition of social mechanisms,[34] the circumstances in which they should be studied, and what type of evidence is required to identify them correctly. As a consequence, there is neither a shared understanding of when process-tracing is a fruitful strategy to employ, nor are there agreed-upon criteria to tell apart successful applications of process-tracing from unsuccessful ones.

Process-tracing is sometimes said to be helpful for identifying causal mechanisms when it also makes use of statistical and experimental evidence, or when *matched* by a theory of social phenomena fully specified beforehand[35] (Steel 2008, Gerring 2007a, Bennett and George 1997, George and Bennett 2005). In much social scientific practice, however, these evidential sources are either unavailable or only offer a partial support.

Historical narratives are often a case in point. When process-tracing is used in case study research as a method of causal analysis, it helps develop historical narratives, which are case studies that provide causal explanations of phenomena of interest in the form of descriptions of the sequence of events leading to the outcome. Process-tracing plays a central role in this type of studies because the narrative rests on the causal mechanism that process-tracing identifies. In this context mechanisms often are either infrequent or unique socio-historical patterns that cannot be studied in a controlled setting. Statistical or experimental data are, thus, hardly available, or only provide partial support, which is insufficient on its own to fully outline the mechanism. Furthermore, the mechanisms are typically identified upon the examination of heterogeneous piecemeal evidence and not derived from a theory fully specified beforehand.[36]

[34] See Hedström and Ylikoski (2010) and Gerring (2007b) for an overview of the various definitions of mechanism in the philosophical and social scientific literature.

[35] In what follows I will sometimes refer to these theories as "fully-fledged" for short.

[36] Background theories play often a role in historical narratives but, as I shall discuss below, this is not the role assigned to them by current accounts of process-tracing. This consideration also extends to analytic narratives which constitute a subset of historical narratives that make use of formal models (i.e. rational choice and game theoretic models) to derive hypotheses about the relevant mechanisms. Even in these studies the exact relation between the formal model and the mechanistic hypothesis is far from obvious. For an example of analytic narratives see Bates et al. (1998). For a critical discussion of the role of models in analytic narratives see Alexandrova (2009).

In these cases it is unclear what counts as a successful application of process-tracing, and it is left open what the epistemic status of its findings is. For want of practicable criteria for assessing process-tracing contribution, it seems that it is to be acknowledged a "mere" heuristic role. This conclusion, however, obscures the fact that in the narrative process-tracing often plays an important *complementary* role to other techniques of causal analysis in achieving valid causal inference, and that whether it succeeds in doing so does not only depend on the kind of evidence that backs it up. It thus remains to be seen under what circumstances process-tracing can be an effective complement for valid causal inference in situations where reliable evidence of other kind such as statistical or experimental data is not present.

In this chapter, I shall propose two criteria to determine whether process-tracing helps to achieve valid causal inference when used as an adjunct to other strategies of causal analysis. The fulfillment of these criteria ultimately depends on the capacity of process-tracing to provide a *complete* characterization of the mechanism responsible for the causal relationships of interest. If process-tracing identifies *complete* mechanisms, it can infer causal sequences of events which are in turn evidence for the causal relationship in question.

The criteria I propose are valuable in two respects. First, even though they do not constitute a condition of internal validity on their own, they can be shown to be effective in distinguishing successful applications of process-tracing from unsuccessful ones. And since their fulfillment is not tied to the use of any specific kind of evidence, they are likely to be applicable "across the board", that is, also in some of the cases in which the favorable evidential circumstances do indeed obtain. Furthermore, upholding these criteria allows a more nuanced assessment of process-tracing inference beyond the sharp distinction between a "mere" heuristic role and full-blown validity.

The chapter is organized as follows. In section 2.2, I sketch some views on process-tracing as an alternative strategy of causal inquiry, including Daniel Steel's framework, which will be considered in detail. In section 2.3, I argue that the conditions of validity Steel proposes are tailored for specific applications of process-tracing, and fail to apply to historical narratives. In section 2.4, I propose two criteria to single out successful applications of process-tracing when historical narratives are not developed against the background of fully-fledged social theories. In section 2.5, I explain how process-tracing can help causal inference effectively if it meets the proposed criteria. Section 2.6 concludes this chapter.

2.2 PROCESS TRACING AS AN ALTERNATIVE STRATEGY OF CAUSAL INQUIRY

2.2.1 CURRENT ACCOUNTS OF PROCESS TRACING

The debate on process-tracing hosts a variety of positions. Scholars have different views on what social mechanisms are, how exactly process-tracing can lead to their discovery, and whether it can do so reliably. Arguably, the characterization of process-tracing they offer is contingent on the specific definition of mechanism each of them endorses, and so does their ultimate view on the capacity of process-tracing to draw valid causal inferences. In this section and the next, I provide an overview of the most influential positions in the literature, focusing on their definition of social mechanism and the related understanding of process-tracing. Two main accounts of process-tracing will emerge. The former characterizes process-tracing as reconstructing sequences of events responsible for outcomes of interest, and as using mainly historical or qualitative evidence to this end. The latter characterizes process-tracing as identifying mechanisms understood as a set of entities and activities responsible for general causal relationships, and does not put any constraints on the kind of evidence it relies on.

2.2.1.1 DANIEL LITTLE: PROCESS TRACING AS HISTORICAL INVESTIGATION

Daniel Little (1991, 1998) conceives of causal relationships between social phenomena as constituted by causal mechanisms. To him asserting that C causes E is to assert that C in the context of typical causal fields brings about E through a specific mechanism. In particular, "social causal relations are constituted by the causal powers of various social events, conditions, structures, and the like, and the singular causal mechanisms that lead from antecedent conditions to outcomes." (Little 1998: 161). This metaphysical thesis is accompanied by an epistemic requirement. Little in fact maintains that to establish credibly a causal relationship between C and E, one needs to identify the mechanism connecting the former to the latter. Only if the mechanism leading from C to E has been de discovered, one can assert confidently that C causes E. Knowledge of mechanisms is in this sense necessary for causal inference. Little sees mechanisms as a series of events (C_1, C_2, C_3...) that links C to E, where the transition from an antecedent to its consequent is governed by a causal law (Little 1991: 15). It follows that the claim that C causes E rests on the identification of the series of causally connected events that lead from the former to the latter.

Process-tracing helps with the analysis of causal mechanisms, because it inquiries into the history of the event of interest and formulates causal hypotheses about its course. This research strategy is typically used to answer singular causal questions such as: Why did the Nicaraguan revolution occur? Why did European growth slow down after September 11th 2001? Little characterizes process-tracing as collecting and analyzing historical evidence, broadly understood as comprising all knowledge of social, economic, and political circumstances, and facts that are somehow relevant to the event of interest. Process-tracing then aims to reconstruct the chain of events that actually led to the outcome of interest. In Little's account the chain so conjectured is valid evidence for the hypothesis in question only if it identifies a causal mechanism. That is, the events that make up the chain ought to be causally connected. Process-tracing can contribute successfully to causal inference insofar as it responds adequately to this challenge.

2.2.1.2 ANDREW BENNETT AND ALEXANDER GEORGE: PROCESS TRACING AS COLLIGATION

Andrew Bennett and Alexander George (1997, 2005) define mechanisms as the causal processes and intervening variables through which explanatory variables produce their effects. The notion applies to processes, including intentions, expectations, information, small group and bureaucratic decision-making dynamics, coalition dynamics, strategic interaction and so on (1997: 1). Bennett and George do not regard the identification of mechanisms as sufficient for causal inference; nevertheless they see it as a necessary epistemic complement to evidence of correlation. They remark that analyses only based on comparative methods or statistical techniques are plagued by the so-called problem of confounders: an observed correlation between variables can in fact be explained either by one variable causing the other, or by unobserved third factors causally related to both. Because of their inability to achieve full control of causally relevant variables, the statistical and comparative methods cannot draw valid causal conclusions. According to Bennett and George, identifying the mechanism between the variables of interest provides decisive evidence that the variables are indeed causally related; in so doing, it would allegedly provide a solution to the problem of confounders.

In their view, process-tracing is the method that generates and analyzes data on the mechanisms linking the putative cause to the observed effects. They relate it to the practice of colligation used in historical explanation, which consists in tracing the minute chains of events that brought more complex phenomena about (1997: 5). Unlike purely historical explanation, however, process-tracing further couches the empirical observations in the analytical terms the research-design first identifies. In their account process-tracing is, in fact, a research strategy deeply loaded with theory. Its contribution to causal inquiry depends ultimately on how "tight" its relationship with theories of social processes is. George and Bennett distinguish two approaches: process verification and process induction (1997: 10). In the former case, the application of process-tracing is guided by a fully-fledged theory which already obtained some empirical support. In this case, process-tracing collects empirical observations to be matched to the predictions entailed by the theory. In the latter case, when theories of the former type are inapplicable, the investigator sets out for "the inductive purpose of finding one or more potential causal paths which can then be rendered as more general hypotheses for testing against other cases (1997: 12)".

2.2.1.3 JOHN GERRING: PROCESS TRACING AS DETECTIVE WORK

John Gerring presents a rather sketchy account of process-tracing, even though he fully acknowledges the important role it occupies in case-study research (2007a). Two strategies of causal inference are available to the researcher who performs case studies. She either employs an experimental template or uses process-tracing. In the former case, the study is organized in such a way that variation is observed in the phenomenon of interest and the putative cause only, while keeping the other relevant factors fixed. Since we are not in the laboratory, however, this high level of control is hardly attainable in case study research, and the experimental template is seldom applicable. Process-tracing reconstructs the causal chain relating the relata in question. How these chains are actually identified is not discussed by Gerring. He rather emphasizes that process-tracing uses peculiar evidence to this end. In particular, he says that *multiple types* of evidence are employed for the verification of a single inference (Gerring, 2007a: 173). Process-tracing, he suggests, collects strands of heterogeneous evidence more akin to the sparse clues in detective work than to the observations on homogeneous samples typical of other empirical methods.

The aforementioned scholars regard process-tracing as a strategy of causal inquiry *alternative to* other methods of causal inference in the social sciences. The type of contrast they draw changes, however, from scholar to scholar. Daniel Little, Andrew Bennett and Alexander George regard process-tracing as an alternative to the comparative and statistical methods (Little 1991, Bennett and George 1997, George and Bennett 2005). Gerring treats it as an alternative to the experimental template *in* case study research (Gerring 2007a). All these accounts suggest that what is distinctive about process-tracing is the *kind* of evidence it uses. It fails, however, to be acknowledged that, as a matter of fact, process-tracing often uses a mixed evidential basis that includes both quantitative and qualitative evidence. More importantly, the circumstances in which process-tracing is a valuable strategy to use are not made clear by these authors. Since its evidence is characterized as non-experimental, non-statistical, or vaguely qualitative, process-tracing is portrayed (if perhaps to some extent unintendedly) as the method of “last resort”, to which scholars turn for want of a better alternative.

Daniel Steel also regards process-tracing as an alternative to other strategies of causal inference. What is distinctive of it, though, is not whether it uses evidence of qualitative, historical, or quantitative nature; rather, it is the level of analysis on which process-tracing focuses: whereas the alternative strategies of causal inference examine evidence directly pertaining to the causal relationship of interest, process-tracing uses evidence that is indirectly related to it. It is thus mostly valuable in those epistemic circumstances where only indirect evidence is available. Unlike the other scholars, Steel proposes a framework which, while taking into account the fact that process-tracing typically relies on a mixed evidential basis, also clarifies *when* process-tracing is the most fruitful strategy to use (Steel 2008).

2.2.2 DIRECT VERSUS INDIRECT CAUSAL INFERENCE

Steel distinguishes between methods of direct and indirect causal inference (2008: 175). These two methods use different types of evidence for *analogous* inferential purposes, such as learning about the causal relationships among macro-features of a complex system.

Consider the set of variables V, the causal relationships among which one intends to investigate. They might be the macro-features of a system S, such as inflation and unemployment if the system is an economy; or, to take Steel’s example, exposure to Aflatoxin B_1 and development of liver cancer if the system is an organism (2008: 187-8). The researcher faces a choice between two approaches: she can proceed to draw causal conclusions among variables in V using evidence about those same variables. She would thus employ *direct causal inference* (ibid.). Alternatively, she might try to draw the same conclusions by using evidence about a distinct and yet related set of variables M and thus employ *indirect causal inference* instead. M contains the properties of the components of system S, which she is ultimately investigating. If S is an economy and the relationship of interest is between inflation and unemployment, M might include the households and their consumption behavior, the firms and their employment policy. Thus, rather than focusing on the systemic features directly, the researcher would study them indirectly by first examining the structure responsible for them.

The system components and their relationships when associated in such a way as to give rise to macro-level regularities constitute what Steel defines as a *causal mechanism*.[37] In the social realm the mechanism's components are agents grouped into categories associated with characteristic modes of behavior. Thus, social mechanisms involve reference to some categorization of agents into relevantly similar groups defined by a salient position their members occupy vis-à-vis others in the society. In the description of the mechanism, the relevant behavior of the agent is often assumed to be a function of the group to which he or she belongs (2008: 48). The characteristic behavior of agents constitutes what one calls a practice, or custom. When applied successfully, process-tracing identifies the set of practices that link together to form a social mechanism (2008: 190). As an illustration of the type of mechanisms process-tracing detects, we consider one of the cases Steel uses to describe its operation.

This case concerns the Trobriand society, studied by leading anthropologist Bronislaw Malinowski (1935). Malinowski uses indirect evidence to support his claim of the existence of a causal relationship between (C) the possession of many wives and (E) wealth and influence among Trobriand chiefs. In absence of statistical evidence about the relationship between C and E, this causal claim is supported by evidence about the relevant social processes in Trobriand society. The social practices for which Malinowski collects evidence are the following: (1) the custom whereby brothers are required to contribute substantial gifts of yams to the households of their married sisters – gifts that are larger than usual when the sister is married to a chief. Furthermore, (2) yams are the primary means used by chiefs to finance their political endeavors and public projects (Steel 2004: 67). In virtue of (1), the possession of more wives causes the possession of more yams; in

[37] The definition of mechanism I endorse is rather close to Steel's but generalizes it in important ways: components in social mechanisms are not necessarily associated with *characteristic* modes of behavior; more generally, they engage in activities/actions in virtues of their properties and their organization within a given system. Furthermore, mechanisms do not give rise necessarily to regularities but also to singular causal relations. I'll come back to these points in section 2.4 below.

virtue of (2), the possession of more yams leads to greater influence among Trobriand chiefs. Insofar as (1) and (2) hold, we can conclude that C causes E. Process-tracing serves to convince us of the existence of (1) and (2).

In what follows I shall discuss how Steel presents his proposal as a solution to the problem of underdetermination that typically plagues forms of direct causal inference in the social sciences (2008: 174).

Most forms of direct causal inference are in fact based on the use of comparative and statistical methods. Steel remarks that conditional probabilities and statistical associations are often insufficient evidence to identify causal relationships. Common causes (i.e. confounders) might be responsible for the observed correlation between variables and remain undetected by the scientist. Furthermore, it can still be the case that the causal structure cannot be identified by means of probabilities only and more sophisticated strategies such as the use of instrumental variables,[38] even though they might help circumvent the problem, are often inapplicable. Process-tracing might offer a solution to the problem of underdetermination faced by the strategies of direct causal inference *as it confronts a distinct underdetermination problem*. This consists in the identification of underlying mechanisms by using evidence about the system components rather than focusing directly on the systemic causal relation itself. Since the two problems are distinct it seems plausible to Steel that there can be cases in the social sciences in which one can be solved successfully while the other cannot.

Steel illustrates the point in relation to the case of Trobriand society. Imagine that V = $\{W, S, N\}$, where W, S, and N are variables indicating Wealth, Social status, and Number of wives, respectively (2008: 193). It is plausible, says Steel, that there is a causal connection between each pair of these variables that is unmediated by the third: status is likely causally linked to wealth as cause or effect independently of number of wives; it is likely that status and number of wives are linked by a path that is not mediated by wealth;

[38] Instrumental variables are used for causal inference in econometric regressions when the relationship of interest is likely to be confounded. For a more extensive discussion of instrumental variables along Reiss' lines see section 4.2.2.

and it is also likely that wealth and number of wives are linked by a path unmediated by status. It follows that alternative causal structures are equally plausible and compatible with the same statistical evidence. Consider the following diagrams each representing plausible alternative causal structures (193):

Fig. 2.1

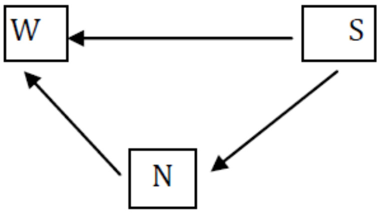
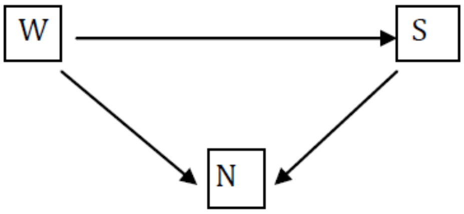


Fig. 2.2



Under standard assumptions about the relationship between probability and causality,[39] neither the case in which *S* causes *N* and *W*, and *N* causes *W* (fig. 2.1) nor the case in which *W* causes *S* and *N*, and *S* causes *N* (fig. 2.2) generate probabilistic independencies among the variables: each pair of variables is probabilistically dependent both marginally and conditional on the third variable. Given to the absence of probabilistic independencies between *W*, *S*, and *N* the two graphs are indistinguishable and the statistical evidence cannot help establish whether *N* in fact causes *W*. If adequate instrumental variables are not available, strategies of direct inference cannot solve the underdetermination problem they face.

Process-tracing uses evidence about a distinct but related set of variables. In this case it investigates the practices that together sustain the causal connection between *N* and *W*. The epistemic problem it confronts is the correct interpretation of these practices. How can one establish whether yams are used as means to finance public project and political

[39] Steel assumes the faithfulness condition (FC), which asserts that the only probabilistic independencies in acyclic causal structures are those entailed by the Causal Markov Condition (CMC). The CMC asserts that, conditional on its direct causes, a variable is probabilistically independent of any set of other variables that does not include its effects. FC and CMC are discussed and defended by Judea Pearl, Peter Spirtes, Clark Glymour, and Richard Scheines, among others. See Hitchcock (2010) for a discussion of these points.

endeavors rather than as savings which the wife's family offers for the upbringing of offspring? And that this transaction strengthens the reputation of the man who receives rather than contributes yams? Even in this case, as in the case of direct inference above, multiple explanations are plausible. The problem of underdetermination at stake is, however, distinct from the one above as it regards *other variables*: flows of yams, kinship relations, public projects, and the practices of which they are part. Furthermore, these practices are typically identified by different means, such as interviews, prolonged observation, and active engagement with the community. Steel's point is that there are cases in the social sciences in which the epistemic circumstances are more favorable to the identification of the underlying social practices rather than the systemic causal relation itself.

If, as Steel suggests, process-tracing is a strategy to exploit epistemic circumstances distinct from the ones on which methods of direct inference rely, it might then be helpful in two ways:

1. as a *surrogate* of the alternative methods;
2. as a *complement* of the alternative methods.

1. As a surrogate it would be used *in place* of the alternative methods in cases in which the latter are not applicable in a fruitful manner. In this case the direct benefit of employing process-tracing would consist in *extending* the range of cases where causal inference is possible in the social sciences.

2. Alternatively, as a complement it might be used *in tandem* with the other methods in circumstances that are somehow favorable to both, and still have the benefit of *increasing* the overall confidence in the causal conclusion. The two ways of process-tracing (1, 2) are not mutually exclusive. Process-tracing might be employed as a surrogate in certain cases and as a complement in others. It is, however, plausible that the strength of the causal inference varies in the two cases. For in way 1 the burden of drawing valid causal inference lies entirely on process-tracing, whereas in way 2 it is shared by two or more strategies employed in a mutually supportive way.

If process-tracing is employed as a surrogate method (1), it establishes at best qualitative causal claims. As Steel points out in relation to Malinowski's study of Trobriand society, all one can conclude from successful identification of the mechanism is that there is *at least* one path through which the number of wives exerts a positive influence upon wealth among Trobriand chiefs (2008: 192). That *some* mechanism is identified correctly, does not exclude that other mechanisms might be present which the social scientist fails to notice. And if the other mechanisms happen to exert causal influence in the opposite direction, the overall effect of their joint operation is nil.

If, by contrast, one uses process-tracing as an adjunct method to the strategies of direct inference this problem can be more successfully circumvented. Let us assume that statistical dependence had been established between the relevant macro-variables in the Trobriand case. This evidence supports the hypothesis that the number of wives exerts positive causal influence on wealth, helps assess the strength of this influence, and indirectly rules out (to some extent) the hypothesis that undetected mechanisms might be exerting a neutralizing effect. Be it employed as a surrogate method or in a complementary fashion, it remains to be established whether and how process-tracing can solve the underdetermination problem it faces.

2.3 VALIDITY CONDITIONS OF PROCESS TRACING

If process-tracing is to be a strategy for formulating causal hypotheses *and* providing evidence for them, as Steel indeed suggests, it ought to solve its underdetermination problem, which amounts to identifying the relevant mechanism. According to Steel, this is facilitated considerably in specific evidential circumstances.

If available, statistical and experimental evidence is particularly helpful to establish generalizations about the system components (2008: 188-9). As an example Steel refers to the study by Donohue and Levitt (2001), who propose a novel explanation of the sharp decrease in the crime rate in the United States since the Nineties onwards. Donohue and Levitt hypothesize that "the Supreme Court's 1973 decision in Roe v. Wade legalizing abortion nationwide potentially fits the criteria for explaining a large, abrupt, and continuing decrease in crime" (2001: 380). This hypothesis is partly supported by *direct* evidence about the correlation between the legalization of abortion and drop in the crime rate. Furthermore, a mechanism consisting of the following links is said to be operating: (1) unwanted children are more likely to be born in adverse socio-economic conditions; (2) children born in adverse socio-economic conditions are more likely to engage in criminal behavior. Links (1) and (2) are supported by evidence from experimental and statistical studies. If this evidence is regarded as valid, it follows from links (1) and (2) that the abortion has a disproportionate effect on the births of those who are most at risk of engaging in criminal behavior; thus, were abortion legalized, the crime rate would eventually go down.

The problem process-tracing faces, however, more often consists in interpreting social practices in the absence of statistical and experimental evidential support. In this case, Steel suggests, in line with Todd Jones (1999), that cognitive psychology might be of help by offering evidence about general tendencies at the psychological and cognitive level (2008: 194). In this field controlled experiments are more a practical possibility than in the rest of the social sciences. What cognitive psychology can provide, more precisely, is the background knowledge required for the interpretation of the practice. This is however insufficient to identify fully the various norms, customs, and social practices in place in the various contexts of interest. Steel thus suggests that process-tracing is likely to solve its specific underdetermination problem if the relevant practices present the following characteristics:

> 1. The practice in question is exhibited in publicly accessible settings.
> 2. There is no prohibition, taboo, or other obstacle to open discussion of the practice.
> 3. The practice is transparent to participants, in the sense that participants have a reasonably clear understanding of its functioning (2008: 195).

Steel specifies that conditions (2) and (3) facilitate learning what participants regard as the rules and practices, while (1) allows for comparison with actual behavior (ibid.)

These conditions seem to be tailored for specific applications and settings in which process-tracing can rely on the use of ethnographic techniques such as the method of participant-observation (DeWalt & DeWalt 2002). Participant-observation is a data-collection method that aims to gain intimate and close familiarity with a given group of individuals (religious, occupational, and sub-cultural groups, or a particular community) and their practices through an intensive involvement with people in their cultural environment, usually over an extended period of time. To compare what the agents say with their actual behavior, and thus satisfy the criteria above the researcher ought to be granted access to the social setting where the practice enacts and be able to interact with the agents involved in the practice. More in general, the criteria above presuppose that the practices in question are *directly* observable.

This is hardly ever the case in social science. This might be due to a variety of factors: trivially, the researcher might be denied direct access to the setting; more interestingly, the setting in which the practice is displayed might be too broad or the practice too complex for it to be observable in any obvious way; or, the evidence might be available only from indirect and historical sources. Many fields in the social sciences, such as political science, social history, comparative sociology, and some subfields in economics, hence cannot rely on the use of ethnographic techniques; yet they draw on a form of analysis that is largely based on process-tracing. In particular, process-tracing is used in these cases to develop causal accounts in the form of historical narratives. Since probabilistic and experimental evidence plays here a marginal role and ethnographic techniques are often inapplicable, the use of process-tracing in these cases cannot be assessed by the criteria listed above.

Scholars such as Little, George, and Bennett pay closer attention to the use of process-tracing in historical narratives and have shown awareness of the difficulty this use entails. Historical narratives are case studies[40] aimed to articulate and substantiate a causal hypothesis. Typically, they do so by reconstructing and describing the sequence of events that lead from the alleged cause to the purported effect. The sequence of events is valid

[40] See section 1.3.1 for a definition of case study.

evidence for the hypothesis at hand only if the events that make up the chain[41] are causally connected. Daniel Little believes that process-tracing confers just temporal order to the events considered and only *formulates* hypotheses about the causal connections between them. In a similar vein, George and Bennett claim that process-tracing on its own only tracks temporally ordered chains of micro-events. In other words, in both accounts process-tracing cannot distinguish between accidental and causal sequences of events. Evidence that testifies that the sequence identified is *causal* needs to come from elsewhere, and in particular, from well specified theories of social phenomena. In absence of a theory, process-tracing is regarded merely as a helpful heuristic strategy.

This conclusion can be explained by the causal assumptions these authors hold. Whereas George and Bennett do not render explicit their view, Little asserts that mechanisms are fundamental to causation: there is a causal relation between variables if and only if there is a causal mechanism connecting them (1991: 25). However, if one considers the definition of mechanism Little endorses, one concludes that he holds ultimately a Humean theory of causation: causation is a temporally asymmetric, stable association between events. Little, in fact, defines mechanisms as series of events leading from cause to effect where the transition from an event to the other is governed by lawlike *regularities*. Social mechanisms, in particular, *derive* from lawlike regularities that govern the behavior of individuals (1991: 18). It thus seems that Little ultimately sees regularities as prior to mechanisms:[42] to identify mechanisms one needs knowledge of regularities; knowledge which cannot be offered by process-tracing as it typically deals with unique series of events.

Claims of causal connectedness between events in the chain thus require support from theoretical evidence in the form of deductions from established theory. Little cites the example of the Communist revolution in China in the late 1930s and asserts that it is

[41] In what follows I use "sequence" and "chain" as synonyms.

[42] Little's view can be summarized by the slogan "no regularities, no mechanisms". He further specifies that "regularities stem from the properties or powers of a range of entities" (1991: 18). He does not tell us, however, what kind of things in the world *powers* are, nor does he specify how one would proceed to identify them. Even though he tries to give more substance to his theory of powers in later contributions (1995), his theory remains underspecified from an epistemic point of view.

plausible to regard the worldwide Great Depression as causally responsible for the Chinese revolution that followed, because the former significantly affected the Chinese rural economy. This claim ought to be supported by an account of the particular causal sequence of events that led from the putative cause to the effect. To identify this sequence, however, one should substantiate the claim that the events in the chain are causally connected by appealing to theory. Little clarifies:

> We note that singular event **a** is followed by event **b**, and we argue that this was to be expected on theoretical grounds. Suppose for example that it is held that falling prices for cotton in the international market in the 1930s caused Chinese peasant activism. This causal judgment may be supported by a theoretical analysis of peasant political motivation focusing on the connection between peasant economic security and political behavior. (1991: 30)

One needs a theory of political behavior that specifies what effect is to be expected when certain regularities about human behavior hold. In the example above, it might assert (among other things) that in worsening economic circumstances peasant communities join or promote radical political movements, or both. Process-tracing provides historical evidence that in the 1930s the economic conditions in the countryside were indeed severe *and* that peasants actually joined in radical movements. The causal claim about the actual mechanisms is validated if, and only if, it is entailed by the established theory and confirmed by the series of occurring events. The events in the series are causally related on behalf of the regularity posited by the theory, which we need to ground the claim that they constitute a mechanism. Process-tracing amounts to matching the empirical evidence about token events with the predictions entailed by the theories in this case. If there were no theory to guide its application, process-tracing would not have the capacity to distinguish causal from accidental sequences, and its contribution would be heuristic.

By distinguishing sharply the justificatory and heuristic role of process-tracing, however, this view neglects that further meaningful distinctions can be drawn to assess more precisely the epistemic contribution of process-tracing. In what follows I shall argue for a distinction between those applications of process-tracing that fail and those that succeed in providing evidence for causal inference in the absence of a well specified theory from which hypotheses about the relevant mechanism can be derived. Even in cases in which process-tracing is not fully backed-up by a social theory it can still succeed *as an effective complement for causal inference*. Its success does not depend strictly speaking on the *kind* of evidence it uses, and is not jeopardized by the absence of an established theory. Rather, it depends on how well supported by the evidence the narrative that it articulates is. Only if this evidence is sufficient to outline a mechanism that is *complete* vis-à-vis the historical hypothesis in question, process-tracing can help causal inference effectively.

Before moving to this discussion, a further point needs to be made about the distinction between the justificatory and the heuristic power of process-tracing. Internal validity should be regarded as a scientific ideal, which the various methods of causal inquiry approach to varying degrees. Julian Reiss sees validity as having to do with the correctness of the inference from evidence to hypothesis (Reiss 2008); he remarks that the inference can be regarded as correct once *known* sources of errors are controlled for. Trivially, we cannot expect to be able to control sources of errors that are unknown at the time of the inquiry.[43] Furthermore, even if we have principled knowledge of *which* errors we can make, we might still be unable to control for them, or to have full certainty that control has been effectively achieved.

For these reasons, it might be helpful to regard Reissian validity as the end of a spectrum. At the other end, validity is at its minimum, and the method can be regarded as having only heuristic capacity.[44] That is, it can formulate new hypotheses but cannot provide evidence for them. The point I will try to make with the discussion that follows is that there might be room for further fruitful distinctions along this continuum.

[43] Unless we control for them accidentally that is, without being aware that we have.

[44] This presupposes that one rejects the sharp distinction between context of discovery and context of justification as in Steel (2008).

2.4 PROCESS TRACING AS AN EFFECTIVE EPISTEMIC COMPLEMENT

In historical narratives process-tracing is often employed in tandem with other strategies of direct causal inference, such as the method of comparison. In these cases, the narrative typically offers:

(**GH**) General Hypothesis, which encompasses the cases compared;

(**HH**) several Historical Hypotheses, one for each case considered.

We can regard GH and HH as type and token causal claims respectively. Type causal claims describe causal relations in which relata are generic (e.g. expansive educational policies improve economic development), whereas token causal claims describe causal relations between singulars (e.g. UK government decision to rescue the Royal Bank of Scotland in September 2010 jeopardized the support of British taxpayers).

Thus, GH characterizes the causal relationships among variables at the more general level: its scope typically covers all cases under consideration; HH describes the causal connections among events occurring in each particular case. The historical narrative develops at the interplay of the two (sets) of hypotheses. It is typically structured around the description of several causal chains each connecting the variables in GH as instantiated in the specific cases. The claim was made that HH *provides evidence* for GH, and that it does so only if the causal chains it describes are *uninterrupted* (George and Bennett 1997). In George and Bennett's account, the causal chain is uninterrupted if, and only if, each link conforms to the expectations previously formulated on the basis of the background theory. Whereas absence of disruption is a plausible requirement, it is possible to provide an account of it that is more accurate, and at the same time, less restrictive than the one proposed by George and Bennett.

In what follows I shall argue that when employed jointly with the comparative method, process-tracing contributes *effectively* to causal inference if, and only if, it identifies causal chains that are *continuous*. My argument develops in two steps. I first provide a criterion for the *continuity* of causal chains as posited by HH, and illustrate by way of an example how process-tracing can fail to satisfy it. The usefulness of this criterion will become apparent once I introduce and discuss a second criterion, related to the former, which establishes when process-tracing succeeds as an *effective complement* for causal inference. By reconstructing continuous causal chains process-tracing partakes in a broader framework for causal inference where valid conclusions are derived by gradually eliminating rival hypotheses. I shall discuss in detail how and why this is the case in section 2.5.

Here is a criterion for continuous causal chains:

> **Continuity Criterion (CC):** The causal chain in (HH) is continuous if, and only if, process-tracing articulates a *complete causal mechanism* vis-à-vis the historical hypothesis in question.

Consider first the distinction between causal *chains* and causal *mechanisms*. Causal chains are sequences of events that are causally connected. Causal mechanisms are sets of components, and the related properties, organized in a system in such a way as to give rise to causal chains. In drawing this distinction, my view differs from Little's who identifies the mechanism *with* the chain of events itself. It also differs from Steel's who defines social mechanisms as only giving rise to macro-level *regularities*. I do not regard the production of *regularities* as a defining feature of causal mechanisms.[45] Social mechanisms can also bring about unique series of events. If the mechanism is triggered repeatedly and is embedded in a stable system then it also gives rise to regularities. In historical narratives, the mechanism, or mechanisms, in question is responsible for the chains of events as it unfolded in each case *and*, if instantiated in more the one case, for the general causal relationship as well.

Completeness is achieved if, and only if, every link in the chain is backed up by the mechanism at work, that is, the mechanism is structured in such a way that the interactions among its components are responsible for the entire sequence of events in the causal chain. The description of the mechanism typically includes the characterization of the relevant components, of the system in which they are embedded, and the type of relationships among them. Relationships among components are typically, though not necessarily, governed by social practices, norms, customs, laws, and so on. The sequence

[45] For a similar view see Reiss (2008: 108-110).

of events is produced by interactions among the mechanism components that, once triggered, unfold in virtue of, and are shaped by, the components structured relationships. The causal chain has a gap whenever one of the links fails to be explained by reference to the set of underlying structured relationships, or some subset in it. In this case the requirement of completeness is not fulfilled and the causal chain is *discontinuous*.

Consider the following example. In *Regional Advantage*, regional planner Anna Lee Saxenian (1994) explains the diverging economic performance of Route 128 and Silicon Valley, economic regions located in the Boston area and Northern California, respectively. Silicon Valley and Route 128 had a period of explosive economic growth during the 1970s. The astounding success of the two regions was driven by highly-sophisticated technology production: while Silicon Valley was an international leader in the market for semiconductors, Route 128 had a major role in the production of mini-computers. Their position was severely threatened in the mid-80s by a global turmoil. Japan became a heavy competitor in the market of semiconductors, while Route 128 minicomputers lost market shares to start-ups (from inside and outside the United States), producing personal computers and mainframes. This unexpected change in the global context dramatically hit the two regions. They reacted to the downturn, however, in opposite ways. Silicon Valley swiftly recovered taking control again of the more sophisticated and highly customized segment within the market for semiconductors. Route 128 remained stuck in the minicomputer market, which became increasingly thinner due to the impressive demand expansion for mainframes and personal computers. What we observe at the end of the 80s is thus a thriving Silicon Valley and a steady decline of Route 128.

Comparing the two cases, Saxenian formulates the following general hypothesis:

(GH) Industrial system → Economic adaptation[46]

The industrial system is the structure in which relationships are organized in the economic region. The relations comprise:

- ties among socio-cultural institutions, political powers and firms;

[46] Here and below arrows stand for causal relations, unless specified otherwise.

- relations between firms and suppliers;
- power and work relations within the firm.

Saxenian claims industrial systems to be responsible for the degree of adaptation of the regional economy to the relevant environment. This relationship depends on the specific chain of events that unfolded in each case from the emergence of a given industrial system to the economic performance at the end of the 80s. Silicon Valley, says Saxenian, first developed, then partly abandoned but ultimately enforced a *network-based* industrial system (NBS); Route 128 created and remained stuck in what she defines as an independent-firm based system (IFS). Saxenian formulates two historical hypotheses that posit that the emergence, abandonment, and enforcement of a given system were responsible for the specific path of alternative growth and decline that was observed in each case. Fig 2.3 represents the causal chain that unfolded in Silicon Valley (HH_{SV}).[*]

Two types of mechanisms are posited to back-up the causal chain. They are characterized by similar systems and components but differ in the way in which components are arranged therein. Both mechanisms are embedded in a system – the global market for technology - characterized by fast and sudden changes, and comprise agents in the productive, social, and political sectors. In the first mechanism (NBS),[47] firms and suppliers are tied by horizontal links, companies are characterized by a decentralized power structure and dense relations exist between firms and the political and social agents in the region. The interactions among the components are informed by practices such as a balanced mixture of competition and cooperation, information disclosure, flexibility in the job market, social informality, and so on. The second mechanism (IFS)[48] is arranged in the following way: firms are organized in a vertically integrated manner keeping exchanges with suppliers to the minimum, ties to local institutions are almost non-extant, and power inside the firm is strongly centralized. In this case the interactions among the

[*] HH_{sv} refers to the historical hypothesis about Silicon Valley. Saxenian also formulated HH_{R128}, a historical hypothesis about Route 128. I do not consider the latter to avoid rendering the discussion unnecessarily complicated.

[47] NBS stands for Network Based System.

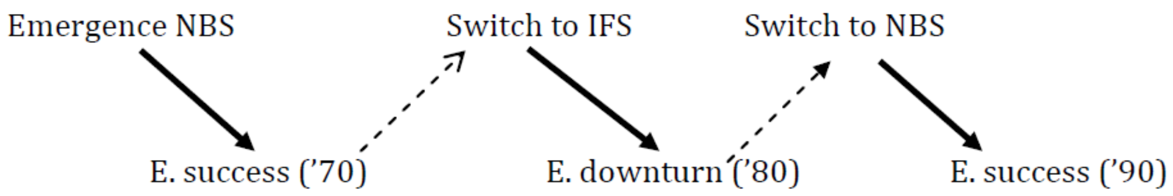
[48] IFS stands for Independent-Firm based System.

agents are governed by practices of fierce competition, information secrecy, job market rigidity, and social ties informed by strong formality.

Continuity is granted to the causal chain if each link is supported by (one of) the posited mechanisms. This seems, however, not to be the case. Whereas some links in the sequence are indeed backed-up by the mechanisms (continuous arrows, fig. 2.3), others are not (dotted arrows, fig. 2.3). Mechanism NBS explains how agents adapt to sudden changes in the relevant environment by relying on practices of information exchange, mixed cooperation and competition, and flexibility (thick arrows, fig. 2.3). On the other hand, mechanism IFS explains how practices that endorse rigidity, information secrecy, and fierce competition prevent adaptation when similar unexpected changes occur, and in this way lead to economic downturn (thin arrows, fig. 2.3).

Consider instead the dotted arrows. Saxenian claims that because of the astounding growth of the 70s entrepreneurs in Silicon Valley “abandoned the network”; and, as a consequence of the severe downturn in the early 80s, they went back to the network (1994: 88). She further claims that the decision of quitting the network system and adopting it again depends on the agents’ capacity (or lack thereof) to correctly understand the reasons of their failure and success. No account is given of why this would be the case and the mechanisms above do not offer such an explanation. The suggested causal link needs to be supported by reference to the (cognitive) mechanisms and social practices in virtue of which the agents formulate perceptions (and misperceptions) of the surrounding environment and restructure their relationships accordingly.

Fig 2.3 (HH$_{SV}$)*



That a causal chain is continuous in the sense described above does not entail that the mechanisms that back it up are correctly identified. The criterion of completeness is not tailored to assess whether process-tracing succeeds or fails in solving its underdetermination problem, to use Steel's terminology. Nevertheless, it can be useful to assess whether process-tracing can help to reach a valid causal inference when used jointly with strategies of direct inference. If the causal chain is not backed up by a complete mechanism, it is discontinuous and fails to connect the causal variables of interest. Hence, it cannot be used effectively to support the general hypothesis GH:

> **Effective Complement (EC).** Process-tracing is an effective complement for (GH) if, and only if, it posits continuous causal chains in (HH).

I shall argue in section 2.5 that if process-tracing satisfies criterion EC it helps reach valid causal inference by way of the gradual elimination of alternative hypotheses. It is, however, important to notice that discontinuity in the causal chain does not render the evidence in its support strictly speaking *invalid*. Consider the case illustrated above. It might well be the case that the described mechanisms are correctly identified; still, they do not back up each link in the causal chain. Hence, process-tracing here provides only partial evidence for the historical hypothesis in question. Nothing excludes that if this evidence were further developed into the description of complete mechanisms, process-tracing would also be an effective complement for causal inference.

The criterion of completeness, thus, cuts across the sharp distinction between heuristic and valid evidence I pointed out above.[49] In particular, that the posited mechanism is incomplete does not imply that the evidence is invalid; vice versa, that it eventually develops in a complete mechanism is not guarantee to full validity. Nonetheless, I shall argue below it can still help causal inference effectively.

2.5 PROCESS TRACING IN THE GENERAL ELIMINATION METHODOLOGY

[49] See last paragraph in section 2.3.

I claimed above that process-tracing is an effective complement for causal inference if, and only if, it posits continuous causal chains in the historical hypotheses HH (effective complement criterion, EC); I further claimed that it posits continuous causal chains if, and only if, it articulates mechanisms that are complete vis-à-vis HH (continuity criterion, CC). If process-tracing meets these criteria when used jointly with the method of comparison, it furthers valid causal inference by helping the elimination of alternative hypotheses. It partakes, in fact, in a more general procedure of eliminative induction that consists in the identification of possible causes of the outcome in question and gradually narrowing down the alternative hypotheses. In what follows I shall first characterize this eliminative procedure within a more general framework of eliminative induction, and then, by way of a case study, explain why continuity is a crucial condition for achieving eventually valid causal inference.

Process-tracing contributes to causal inference within the framework of the General Elimination Methodology (GEM) described by Michael Scriven (2008). Scriven illustrates GEM in its simplest instantiation as follows:

> Every child acquires a repertoire of possible causes for a large number of effects before reaching school age; for example, they know that the vase on the table by the window can be knocked over by the wind, the shades, the cat, a sibling, a playmate, or a grown-up. When they encounter the effect, they begin to sift the list and check for indicators, either immediately observable or quickly accessible, that will eliminate one or more candidates and eventually may identify the responsible cause. This is the basic case of hypothesis creation and verification and it is the essential element, even if subliminally and non-inferentially, in all careful causal explanations (Scriven 2007: 9).

According to Scriven, GEM is the basis for all causal claims:[50] it is used in scientific practice and in everyday situations; it is employed whether one uses background

[50] In my view, Scriven does not really provide sufficient evidence for the claim that GEM underlies *all* causal claims, nor that it underwrites all strategies listed below. However, I think that process-tracing when used jointly with the method of comparison *is* underpinned by GEM, as I shall argue in what follows.

knowledge of theoretical and empirical kind, practical knowledge, or formal strategies of causal inquiry. It operates by gradually narrowing down the range of possible causes of the phenomenon in question until a valid explanation is eventually obtained. In ordinary circumstances practical knowledge of a tacit and implicit kind is often sufficient, as in the case of the competent mechanic who is dealing with a break failure (2008: 21). When new and complex phenomena are examined, as typically happens in the social sciences, one relies on the application of research designs proper.

Scriven suggests the following ways to establish causation relying on GEM:

(i) Direct critical observation, such as visual, and tactile observation. Scriven claims that in certain instances causation can be directly and reliably observed, or more generally, experienced. One s*ees* causation by looking at the pen falling when Jack releases his right hand. Similarly, one *experiences* causation by pushing the speed pedal when driving the car.[51]

(ii) Reported and validated observation, e.g. case studies. Scriven refers here to those case studies that consist in *reporting* previous observations of causal relationships of the sort described in (i).

(iii) Direct or simple inductive inference from (i) or (ii). Upon having seen the pen falling when Jack released his right hand, one can explain the pencil on the floor by Jack's releasing his left hand. Scriven mentions as an instance of this strategy, the inference to the effects of meteorites on the far side of the moon's surface, prior to satellite launching.

(iv) Simple GEM inference, e.g., autopsy, engineering breakdown, and so on; that is, the elimination of alternative explanations which is based on

[51] Scriven defends his claim that causation is directly and reliably observable by referring to experimental findings about concepts development in early childhood and by appealing to the fact that causal claims based on eyewitness testimony (when they meet certain standards such as normal vision, propinquity, absence of motive to lie, etc.) are used for establishing a case in the court of law. For example, eyewitness testimony of a shooting is used to establish whether the victim was actually assaulted by the suspect. Scriven acknowledges that his position is controversial and highly disputed (2007: 5). This fact, however, does not impinge on the relevance of the methodological framework he proposes which accounts both for the cases in which causation is directly experienced *and* for the cases in which it is actually inferred.

background knowledge of possible causes *and* direct observation. If the car does not start, I'll first check the tank because I know that one possible cause is that I run out of fuel; if the tank is full, I conjecture that the battery must be out of charge. The same strategy underwrites the inference made by the coroner when performing an autopsy.

(v) Theoretical inference, based on the use of analogy or theory. This strategy concerns especially fields in which direct observation is not available, such as cosmogony, or geology. Examples are the theory of continental drift to explain mountain ranges, or the dust cloud produced by the meteorite fall to explain the extinction of dinosaurs (Waldner 2003).

(vi) Direct manipulation e.g. in the kitchen and lab. By way of tinkering and making things happen one learns about causation, for instance by lowering the flame one learns to prevent water from spilling; similarly, by raising the flame one learns to make fried egg.

(vii) Randomized Controlled Trials, (RCTs): experimental designs involving at least two groups of subjects, the control group and the experimental group, between which the subjects are distributed by way of a randomizing procedure.

(viii) Quasi-RCT. Experimental situations in which, unlike RCTs, treatment is not assigned to the units by way of a randomizing procedure, e.g. pharmacology.

(ix) Quasi-experimentation. Experimental situations that lack some of the features of the randomized controlled experiment. For instance, post-test only designs are quasi-experiment where the outcome in the two groups is measured and compared only after the treatment. Scriven refers here to fields such as pedagogy, addiction studies, and international aid.

(x) Natural experiments: observational studies where experimental conditions "naturally" occur that is, they are neither brought about by the

> experimenter's intervention nor are they subject to her control. In these situations the treatment is assigned "as *if* random."

I suggest that GEM also underpins the *joint use* of several research strategies together. In particular, when process tracing is employed as an adjunct method to strategies of direct causal inference, they together rely on GEM. In this case, where theory (v) often plays a substantive role, GEM finds application through a sort of *division of labor* among the strategies employed. Process-tracing thus partakes in this general inferential procedure with its own specific contribution. Before describing its role more precisely, let's consider the logical steps underlying GEM. GEM works under the following conditions:

> 1. The general premise is the deterministic principle: all macro events have a cause.
> 2. The first "premise from practice" is the list of possible causes (LOPC) of events of the type in which we are interested, e.g. learning gains, reduction of poverty, extension of life for AIDS patients…An LOPC usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualization; the context of investigation determines the appropriate distance parameters.
> 3. The second practical premise is the list of modus operandi for each of the possible causes (MOL)[52]. Each cause has a set of footprints, a short one if it is a proximate cause, a long one if it is a remote cause, but in general the MO is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective.
> 4. The fourth premise comprises "the facts of the case", and these are now assembled selectively by looking for the presence or absence of factors listed in the MO of each of the possible causes. Only those causes are (eventually) left standing whose MO is completely present. Ideally, there will be just one of these, but sometimes more than one, which are then co-causes (Scriven 2008: 21).

The causal inference proceeds through (a) the listing of possible causes of the effect in question, (b) the listing of their *modus operandi*, and (c) the collection of empirical data to establish the relevant cause by means of the evidence about which MO is actually at

[52] MOL stands for list of modus operandi. MO stands for modus operandi.

work. When process-tracing is employed in conjunction with other research strategies it typically serves at steps (b) and (c). Consider the study by Donohue and Levitt discussed above.[53] Background knowledge supplies the list of possible causes, that is, the explanations so far provided in the literature for the sharp decrease in the crime rate in the United States since 1991. They include the increased use of incarceration, improved policing strategies, the decline in the crack and cocaine trade, and so on (Donohue and Levitt 2001: 380). Unlike the case of the mechanic dealing with a break failure, or the detective searching the motive behind a murder, scientific studies as this one generally aim at proposing novel explanations of the effect of interest. They aim at *enlarging* the list of possible causes currently available in the academic literature. Strategies of direct inference are primarily orientated to identify the novel cause and provide some evidence that this is a prime suspect.

The list of MOs is either retrieved from background knowledge or the outcome of theoretical work. Whether empirical facts are required to prove the absence of the alternative MOs depends on specifics of the case. In the case above, empirical information *not pertaining* to the MOs is considered sufficient to exclude the competing causes. For example, the fact that "Many cities that have not improved their police forces have nonetheless seen enormous crime declines (2001: 380)" is sufficient to exclude the improvement in policing strategies as the main causal factor.[54] In other cases, further empirical information is required to exclude the competing causes. Process-tracing enters at this stage and, in particular, searches for clues that the MO of the proposed cause was actually in place. In this case process-tracing consists in recollecting the experimental and statistical evidence for the theoretical links constituting the mechanism between the legalization of abortion and the drop in the crime rate that followed twenty years later. As argued in the sections above, however, evidence of experimental and statistical character is not always central to process-tracing. Nevertheless, one observes a similar division of labor among research strategies in historical narratives. In this case the application of

[53] Cfr. Section 2.3.

[54] The authors do not exclude that it maybe had a role, but that it was a major one.

GEM is slightly more complicated, as the following case study, very popular among sociologists and political scientists, shows.

2.5.1 THE CASE STUDY: *STATES AND SOCIAL REVOLUTIONS*

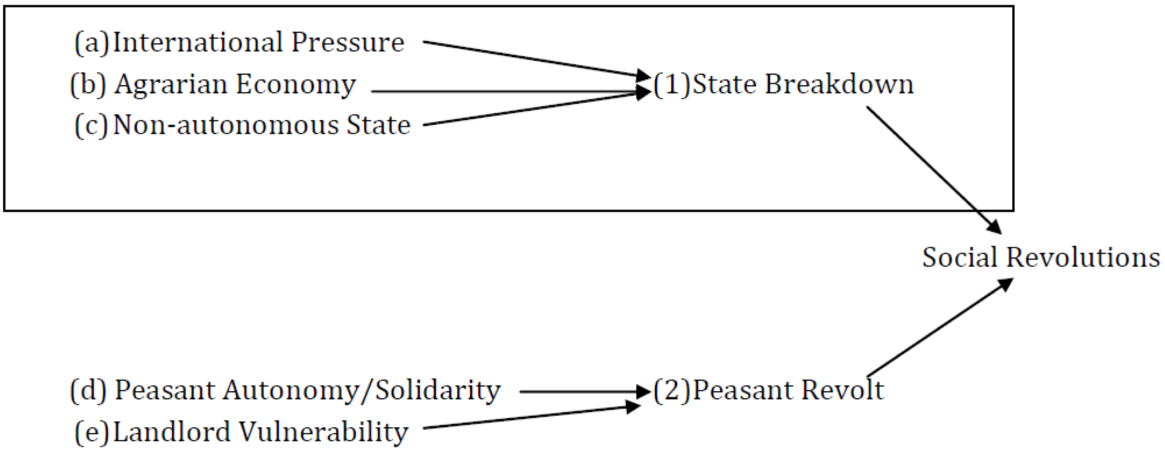
Theda Skocpol (1979) studies social revolutions in France (1788-9), China (1911-1916), and Russia (1917). *States and Social Revolutions* is a historical narrative that draws extensively on the method of comparison:[55] positive and negative outcomes of social revolutions are contrasted and, in accordance with the logic of Mill's method of agreement and difference, similarities and differences among the cases examined. The outcome of the investigation is rather complex. A set of circumstances proves in fact consistently present in the three positive cases (France, China, and Russia), and only partly so in the negative ones (Japan, Prussia, England, and Germany).[56] The method of comparison does not lead to a pattern of causal interaction, but only individuates the factors that have causal relevance in the cases at hand. Causal order needs in fact to be conferred on the set of relevant circumstances to establish causal priority and interactions among them.

By identifying underlying mechanisms, process-tracing reconstructs the causal chains that unfolded in each case and connects the relevant factors. The following pattern thus emerges upon the joint use of process-tracing and the method of comparison. Skocpol identifies two proximate causes of social revolutions: (1) state breakdown and (2) peasant revolt; and five ultimate causes: (a) international pressure, (b) agrarian economy, and (c) non autonomous state as causes of state breakdown; (d) peasant autonomy and solidarity and (e) landlord vulnerability as causes of peasant revolt.

[55] See Mahoney (1999) for a methodological discussion of this study.

[56] Prussia and Germany refer to Prussian Reform Movement (1807-1814) and German (failed) revolution (1848-50) respectively.

Fig. 2.4 General hypothesis in *States and Social Revolutions*

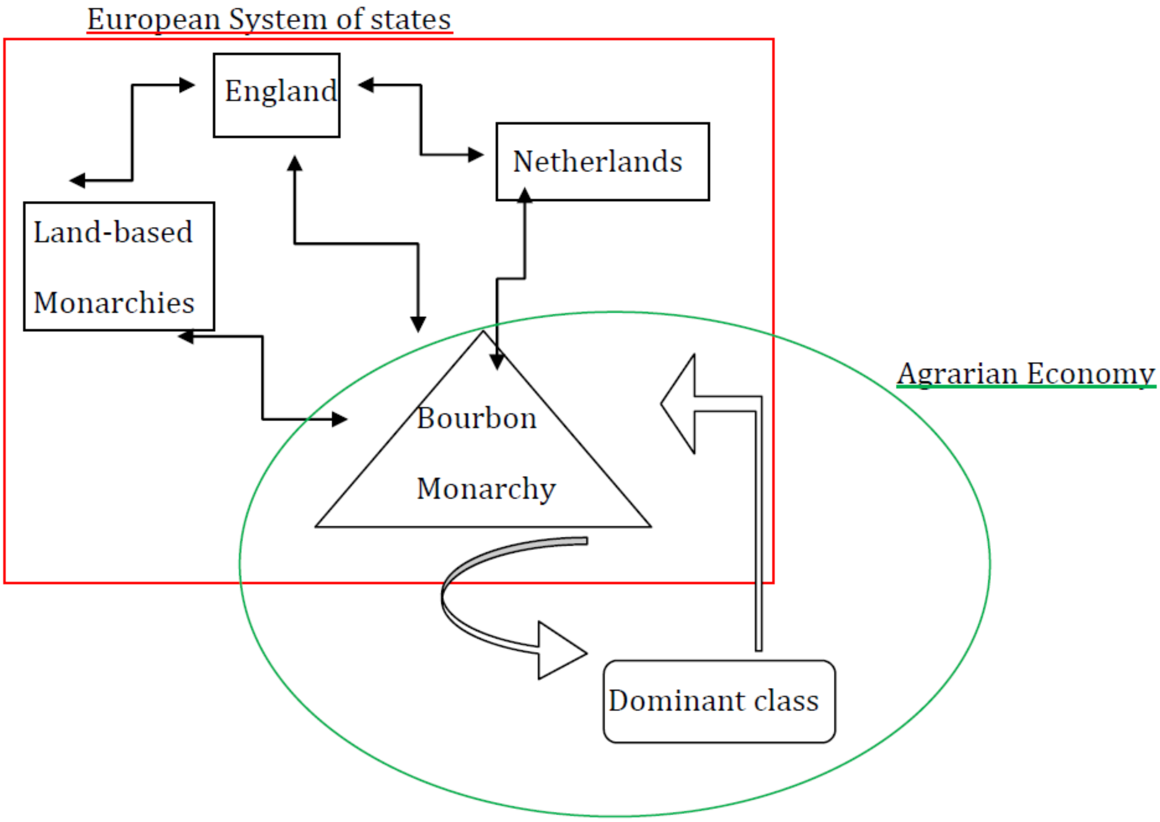


Let's consider the mechanism pertinent to the rectangular section in the causal diagram in fig 2.4. It is represented below in a simplified form (see fig 2.5). The absolutist monarchy is its central component and it is part of two relevant systems. It is a component in the European system of states where it competes with more economically developed powers abroad. In the case of France, the Bourbon monarchy bids for supremacy over continental Europe and on seas (double edged arrows).The absolutist monarchy also is a proto-bureaucratic imperial state embedded in agrarian economies, where it is tied to a dominant class by ties of mutual dependence and conflicting interests. In France, the monarchy relies on the dominant class for collecting dues and taxes among peasants. In return, the latter is granted military protection, fiscal exemption and feudal offices. In virtue of these offices the dominant class has strong political leverage and can oppose monarchy decisions. This set of ties is represented by the thick arrows in the diagram in fig. 2.5.

This complex mechanism explains how events unfolded in France and led to the state breakdown, which eventually caused the social revolution in 1789. Insofar as it is *complete* vis-à-vis the (segment of) the relevant chain, it succeeds in connecting conditions (a), (b) and (c) to (1) in the diagram 2.4. In so doing, it confers the causal order that these conditions would lack otherwise and that cannot be given by the method of comparison which, only detects relevant similarities and differences and cannot tell causes and effects apart. The chain of events this mechanism backs up shows instead how (a), (b), and (c) interact and jointly lead to outcome (1).

Causal order among the relevant factors is not fully achieved whenever the mechanism is incomplete and the causal chain discontinuous. Complete mechanisms are required for filling the diagram with all the links that would be missing otherwise and connect the full set of relevant causal variables. I shall argue in what follows that it is exactly on this ground that process-tracing helps to rule out effectively alternative hypotheses.

Fig. 2.5 Mechanism for breakdown Bourbon Monarchy



Skocpol considers four families of theories of social revolutions as alternatives to her own:

- Marxist theories. This family of theories understands revolutions as class-based movements growing out of objective structural contradictions within historically developing and inherently conflict-ridden societies. Key to any society is its mode of production, or specific combination of socio-economic forces of production (technology and division of labor), and class relations of property ownership and surplus appropriation. The basic source of a revolutionary contradiction in society is the emergence of a *disjuncture* within a mode of production between the social forces and social relations of production. In turn, this disjuncture expresses itself in intensifying class conflicts. Revolution itself is accomplished through class action led by the self-conscious, rising revolutionary class (1979: 7-8).

- Aggregate psychological theories subsume revolutions under the more general concept of political violence. People typically engage in political violence when they live in conditions of relative deprivation that is, when they perceive a discrepancy between the opportunities to which they feel entitled and the ones they actually have (1979: 9). The widespread feeling of frustration generated by this condition of relative deprivation constitutes the fundamental factor that triggers the organization and participation in revolutionary action.
- Systems/value consensus theories understand revolutions as violent responses of ideological movements to severe disequilibria in the social systems. A social system experiences a crisis when values and environment become "dis-synchronized" due to the intrusion of new values or other disruptive factors such as new technologies (1979: 12). If existing authorities fail to respond in a flexible manner to the current crisis and resist change with an intransigent attitude, then revolutionary movements develop in order to re-synchronize societal values and environment. Revolutions thus consist in the purposeful implementation of strategies of violence to implement the necessary change.
- Political-conflict theories understand revolutions as a special case of collective action in which groups fight the government for ultimate political sovereignty over the population and to some extent succeed in displacing existing power-holders (1979: 11). Revolutions develop when

> discontent becomes widespread in the population and leads to the emergence of multiple sovereignties. They are only successful, however, if the contenders, that is, the revolutionary coalition, have control of substantial force (ibid.)

Skocpol rules out rival explanations of social revolutions by means of two strategies:

I. Explanations are ruled out if their modus operandi *fails* to be present.

This strategy thus strictly obeys to the logic of GEM.

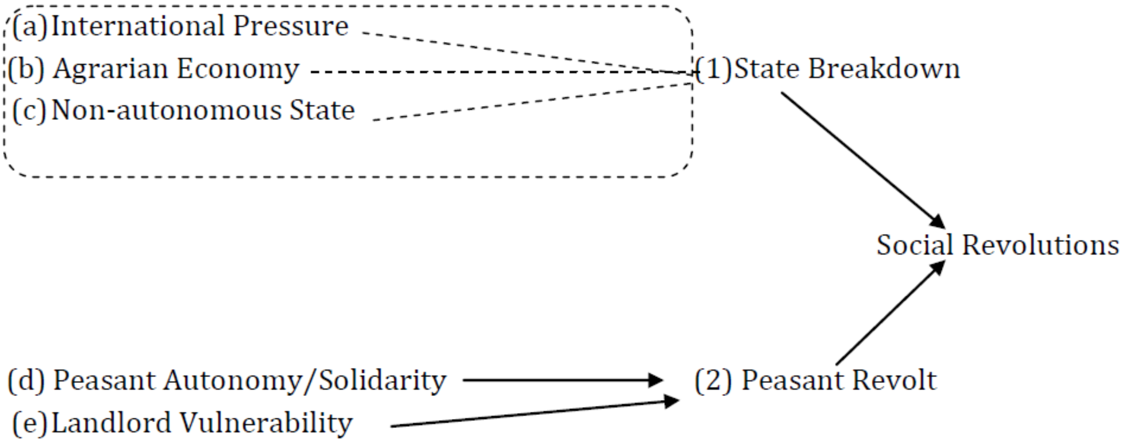
Both Marxist and systems/value consensus theories above posit the presence of a *purposeful* movement directed to overthrow the extant social and political order as an essential component of the revolutionary process. But this factor is found not to be operative in the cases analyzed.

II. Explanations are ruled out if their modus operandi, though present, fails to identify fully the relevant causal chains and thus fail to establish full causal order among the relevant causal factors.

This strategy deviates from GEM as traditionally conceived.

Consider political-conflict theories, which Skocpol regards as relevant to some extent to explain the revolutionary outcomes (1979: 14). Central to their modus operandi are class relations and interests and the availability of resources for class struggle. Even though these mechanisms are both present in the cases at hand – say the class-mechanism and the resource-mechanism – they only explain some segments of the chain of events that unfolded in the three cases (continuous arrows). They fail to account for the sequence of events leading from conditions (a), (b) and (c) to outcome (1). Causal order among *these* conditions is thus not achieved by the modus operandi posited by political-conflict theories. As a consequence they only establish some links in the causal diagram and fail to fill the remaining links in (dotted links).

Fig. 2.6 General hypothesis in political-conflict theories



2.5.2 PROCESS TRACING CONTRIBUTION TO VALID CAUSAL INFERENCE

When process-tracing and the method of comparison are used in historical narratives, they contribute jointly to causal inference by means of a procedure that consists in the gradual elimination of causal hypotheses. Their joint operation conforms to the logic of causal investigation as described in GEM. The method of comparison identifies the relevant causal factors and process-tracing reconstructs the causal chains that in each case connect causally the factors identified by the method of comparison. The eliminative procedure employed consists of two distinct strategies. The first strategy, which is described in GEM, rules out alternative explanations whose modus operandi fails to be present in the cases at hand. GEM does not require the identification of the *actual* modus operandi, but only evidence of the absence of the modus operandi that is posited by the competing explanations. This strategy may not be sufficient to eliminate all alternative hypotheses. As Skocpol's study shows, it might be the case that the modus operandi posited by other rival theories is operating.

The second eliminative strategy narrows further the range of competing explanations in these thorny, but not necessarily uncommon, cases. It helps to rule out those explanations whose modus operandi, even though present, *fails to disclose the whole pattern of causal relations among the relevant variables*. If process-tracing succeeds in identifying causal chains that in each case considered connect *all* relevant variables, it discloses the whole pattern of causal interaction and rules out those hypotheses that fail to do so. Two epistemic considerations justify this eliminative strategy. First, causal order is fully established if all the relevant variables are connected by causal chains. Knowledge of the temporal order in which events occurred is insufficient to establish causal order when the causal structure is complex as in Skocpol's case. Certain circumstances are long-standing causes, such as the agrarian economy, whereas others are triggering factors, such as the increasing international pressure. If one only considers temporal order, it may be plausible to infer that agrarian economy caused the increase in international pressure. Causal chains thus help with telling causes and effects apart.

Secondly, the identification of causal chains connecting the relevant variables provides evidence that corroborates further the general causal hypothesis. In other words, it supports the conclusion that the variables identified by the method of comparison are not spuriously correlated. The method of comparison, does not guarantee that full control of the relevant variables is achieved.[57] Confounding factors may lurk in the background and escape cognizance by the social scientists; and this is indeed one of the most powerful argument in favor of process-tracing as an adjunct method of causal analysis. However, the skeptic may object that this is still insufficient to rule out those rival hypotheses that fail to identify all causal chains among the relevant variables and regard the one hypothesis that is left as valid. After all, it may be the case that the method of comparison identifies variables that are spuriously correlated, and process-tracing posits sequences of events that are spurious as well.

In particular, the skeptic may ask for additional evidence that jointly corroborates the historical and general hypotheses. The identification of *complete* mechanisms provides the evidence required. A mechanism is *complete* vis-à-vis a given historical hypothesis if, and only if, it backs up each link in the chain, which connects in turn the relevant variables in

[57] See section 1.2.1 for a brief discussion of the comparative method and the problem it encounters in providing valid conclusions; see section 1.3.2 for a distinction between the comparative method and the method of comparison.

the general hypothesis. The mechanism thus supports both the causal chain and the general relationship. By describing complete mechanisms process-tracing helps causal inference effectively. Valid causal hypotheses are the ones that are left once rival explanations are gradually eliminated, either because the mechanisms they posit fail to be present, or because it fails to back up fully the relevant causal chains. The criterion of continuity thus discerns the successful causal hypothesis within a narrower range of alternative explanations.

2.6 CONCLUSION

Process-tracing is employed in much practices in the social sciences but there is no shared view yet on when it is a valuable strategy to use, and which criteria it ought to satisfy to be fruitfully applied. Scholars tend to tie its conditions of validity to the use of specific kinds of evidence. Process-tracing is thus regarded as (more likely) to provide valid results if it relies on statistical and experimental evidence, or when its application is governed by fully-fledged theories of social phenomena. However plausible, these conditions are not adequate to assess process-tracing performance in contexts where these favorable epistemic circumstances fail to occur. In these cases, for want of criteria, process-tracing would be in fact acknowledged a "mere" heuristic role. I argued in this chapter that further fruitful distinctions along the continuum between merely heuristic and fully valid evidence can be drawn so as to assess more precisely process-tracing performance.

The criterion I propose, that process-tracing ought to outline *complete* mechanisms vis-à-vis the hypothesis at hand, aims to single out successful applications of process-tracing in historical narratives where the alleged favorable epistemic circumstances tend not to obtain. I argued that when used in tandem with the methods of direct inference process-tracing can help effectively causal inference whenever it fulfills the aforementioned condition. By drawing causal links among the variables that direct inference first identifies, it assists the latter in ruling out the alternative hypotheses. The eliminative strategies on which process-tracing and methods of direct inference jointly rely ideally discard all the rival hypotheses until the point in which a single hypothesis is eventually left. This final theory should be regarded as valid as long as there is no evidence that either disproves the posited mechanisms or identifies new relevant variables so far undetected and thus unaccounted.

3 DRAWING LESSONS FROM CASE STUDIES BY ENHANCING COMPARABILITY

3.1 INTRODUCTION

In an influential book on the principles and practice of the case study method, John Gerring defines the case study as "the intensive study of a single case where the purpose of that study is – at least in part – to shed light on a larger class of cases (a population)" (Gerring 2007a: 20).[58] And, in fact, case studies are often performed with the purpose of "drawing lessons" in the form of conclusions that apply beyond the single case and explain other outcomes in addition to the one studied directly. Obtaining general results from a studied context is a legitimate epistemic purpose; furthermore, it is also instrumental to plan interventions in contexts yet unfamiliar. Case studies, in fact, are also used in fields such as economics, political science or educational research to suggest hypotheses that help informing policy decisions in altogether new situations. The former purpose raises issues of mere generalizability, which regard the range of conditions under which the conclusions of the case study are expected to hold. When policy making is the goal, the concern of generalizability is further complicated by the need to formulate guidelines on how to intervene in unstudied contexts.

In the philosophical literature, issues of generalizability are usually discussed under the name of external validity (Campbell and Stanley 1963, Cook and Campbell 1979, Guala 2005, 2010, Steel 2008, 2010).[59] The concepts of internal and external validity were first introduced by Campbell and Stanley in their work on experimental designs (1963). And indeed, the relevance of the distinction is apparent when thinking of laboratory sciences. In the social sciences, experiments are most often set up with the purpose of teaching lessons about the non-experimental world. Knowing the conditions under which these results are applicable outside the laboratory becomes therefore a prominent concern.

[58] In Chapter 1 I argued that the purpose of generalizing results, though certainly desirable, should not be regarded as a defining feature of case study research as Gerring among others suggests.

[59] See section 1.3.2 for my definition and discussion of internal and external validity.

Though a less pressing concern, generalizing remains a goal of case study researchers. External validity is thus an issue for Case Study Research (CSR) as well.

Philosophers, however, have so far worried little about the generalizability of CSR. Social scientists making use of case studies proved rather timid in addressing this issue. This tendency is changing nowadays. The interest in case studies and their methodological riddles experienced an upsurge in the last decades so that it is appropriate to speak of a new trend. This recent turn has the merit of bringing the problem of generalizability to the forefront.[60] However, even if scholars in the new trend strive for a more autonomous understanding of CSR liberated by the statistical and experimental viewpoint, their attempts are not always coherent and successful. In fact, the contemporary debate on external validity is largely influenced by the heteronomous paradigm of CSR. The currently dominant approach is a hybrid that retains major features of the old heteronomous paradigm in a revised form. To stress the continuity with the old paradigm, I term the currently dominant approach the traditional view on external validity. [61]

The adherents to the traditional view emphasize the specificity of CSR by describing its advantages and disadvantages with respect to other methods of inquiry. In particular, they usually assume the existence of a trade-off between internal and external validity, and find a downside of the case-study design in the lack of external validity. Treated as a comparative weakness of CSR, external validity is, however, only shortly discussed and quickly dismissed.[62] This situation generates an interesting tension and calls for attention: there is, in fact, a gap to bridge between purposes and means. Generalizing is set forth by

[60] I n Chapter 1 I introduce and discuss the new trend which sees case study research as an autonomous epistemic genre.

[61] In section 1.3.2 I distinguish three approaches to generalizability: two of them discuss generalizability in terms of external validity; the first is the "traditional view" and the other is the approach I introduce and defend in this chapter. Whereas the first approach gathers consensus among a rather compact group of scholars in political science, only scattered contributions were made to the second approach. For this reason I regard the "traditional view" as dominant in the debate on the external validity of CSR. The third approach is concerned with the generalizability of theoretical and conceptual results. For the reasons given in section 1.3.2 this approach is not further explored in this thesis.

[62] Gerring (2007a) is an interesting case in point. His book on the principles and practice of CSR devotes two chapters to investigate techniques to strengthen the internal validity of case studies; external validity, after being briefly mentioned in the introductory chapter, is only discussed indirectly in the chapter on case selection.

case study researchers as a prominent goal, but methodological discussions related to it conclude with a gloomy perspective.

In this chapter, I first examine the reasons that led most of the scholars to the conclusion that CSR is a weak methodology with regard to establishing external validity. As we shall see, this conclusion is based on assumptions that are, to say the least, disputable. In section 3.2, I revisit the line of reasoning that underwrites the traditional view on external validity and outline some of the objections raised against it. I shall focus in particular on the role that the concept of *typicality* plays within this view, and argue that the centrality given to it diverts the debate from the real issue of external validity.

One unfortunate result of this has been to lead the debate to a dead end where it stands now. I propose to refocus the debate on external validity in CSR by bringing the concept of *comparability* of case studies to the fore. Whereas typicality characterizes the *case* and its relationship to a set of other cases, comparability is a property of the actual *study*. In what follows, I shall suggest that by making case studies more comparable in the sense to be specified, one has more reliable grounds to judge the external validity of the results. This refocusing has two major beneficial effects. First, my analysis demonstrates why it would be best to situate external validity as a problem of inference rather than mere representativeness, as the traditional view maintains. Second, the approach that I develop suggests strategies for strengthening the generalizability potential of case studies. The goal, in short, is not so much a critique of the traditional view's account of the pitfalls of CSR with respect to external validity, but a shifting of perspective that reveals room for improvement.

3.2 THE TRADITIONAL VIEW ON THE EXTERNAL VALIDITY OF CSR

The use of case studies is common in the social sciences, and apparently increases (Gerring 2007a). Interestingly, CSR starts to be used as an autonomous tool of investigation even in fields that typically relegated it to an at best ancillary position, such as economics (Rodrik 2003, Bates et al. 1998). John Gerring (2004, 2007a) quite surprisingly notices that, even though widely employed across the sciences, CSR is still regarded as a weak methodology, and attributes the low appreciation in which it is held to the general lack of its understanding.

Several scholars have recently tried to rehabilitate this methodology by providing a thorough analysis of its specificity. Becker and Ragin (1992), Brady and Collier (2004), George and Bennett (2005), Gerring (2004, 2007a), Mahoney and Goertz (2006b), and Ragin (2000) all contribute to the methodological reflections on CSR by emphasizing its distinctiveness with respect to the other methods of inquiry. These works find some convergence in their understanding of what CSR is good for. Nonetheless, they tend to agree on the fact that external validity counts as a weakness of the method. This conclusion is supported by a set of assumptions on what external validity is and how it should be evaluated. I term this the traditional view on external validity.

Some of these beliefs are widely shared in the literature on external validity, and are thus not confined within the debate among case study researchers. At the same time, not all scholars above would probably endorse each of these assumptions with the same degree of confidence. Even if it is not fully expressed by any of these authors, I take George and Bennett (2005), Mahoney and Goertz (2006b), and Gerring (2004, 2007a) as holding this view. The assumptions on which it rests are, in fact, traceable in the following excerpts:

> Recurrent trade-offs [in case study methods] include […] the related tension between achieving high internal validity and good historical explanations of particular cases versus making generalizations that apply to broad populations. The inherent limitations include a relative inability to render judgments on the frequency or representativeness of particular cases (George and Bennett 2005: 22).

> Questions of *validity* are often distinguished according to those that are *internal* to the sample under study and those that are *external* (i.e. applying to a broader -unstudied- population). The latter may be conceptualized as a problem of representativeness between sample and population. Cross-case research is always more representative of the population of interest than case study research […] Case study research suffers problems of representativeness because it includes, by definition, only a small number of cases of some more general phenomenon. Are the men chosen by Robert Lane

> typical of white, immigrant, working-class American males? Is Middletown representative of other cities in America? These sorts of questions forever haunt case study research. This means that case study research is generally weaker with respect to external validity than its cross case cousin. The corresponding virtue of case study research is its internal validity (Gerring 2007a: 43).

> In qualitative research, it is common for investigators to define the scope of their theories narrowly such that inferences are generalizable to only a limited range of cases. Indeed, in some qualitative works, the cases analyzed in the study represent the full scope of the theory. By contrast, in quantitative research, scholars usually define their scope more broadly and seek to make generalizations about large numbers of cases. Quantitative scholars often view the cases they analyze simply as a sample of a potentially much larger universe (Mahoney and Goertz 2006b: 237).

Even though not fully developed and thoroughly discussed, the adherents to the traditional view share a set of underlying assumptions that enables the conclusion that the lack of external validity is a comparative weakness of CSR. This conclusion, in turn, fits comfortably an account of CSR that essentially consists in fleshing out the comparative advantages and disadvantages of this method with respect to the other research designs. Specific normative implications are then derived regarding situations where the case study design is the appropriate method to use and how to make it stronger.

The set of beliefs that constitutes the traditional view is the following:

1. External validity is a property of the scientific results *and* of the research designs that delivers them.
2. Internal and external validity stand in a trade-off relation.
3. External validity is a matter of representativeness.
4. External validity is a quantifiable property. Whether it is high or low depends on the scope (breadth) of the population to which the results of the study apply.

Besides sharing the four assumptions on external validity, these scholars agree that CSR encounters major problems in selecting cases that are representative of a large population. This fact together with the assumptions that external validity is a matter of representativeness (3), and the degree of external validity depends on the scope of the population to which the results from a study apply (4) entail the conclusion that case studies provide results with very limited external validity. Furthermore, since external validity is regarded as a property of the results of a given study *and* the method by means of which the results were obtained (1), it follows that CSR is *low* in external validity. Finally, since a trade-off is assumed between internal and external validity (2), CSR turns out to be weaker in external validity *and* stronger in internal validity than methods that are more successful in selecting cases that are representative of broader populations.[63]

I will discuss the four assumptions and their normative implications below. The traditional view on the external validity of CSR borrowed assumptions 1 and 2 from the debate on the external validity of the experimental designs. In this context, the aforementioned assumptions have been independently discussed and challenged. I will examine assumptions 1 and 2 only shortly by referring to this discussion and the related criticisms, and by giving some additional reasons why they are further unjustified when carried over to CSR. I will then turn to assumptions 3 and 4, which have been smuggled into CSR from the statistical discourse. These assumptions seem to be more challenging, and their normative implications, to the best of my knowledge, have not been examined yet. I will then focus on those at length.

3.2.1 THE ASSUMPTIONS CARRIED OVER FROM THE EXPERIMENTAL CONTEXT

Assumption 1 treats external validity as depending on intrinsic features of the research design: a given method is thus characterized as good or bad at providing generalizable results. This is at odds with the original formulation of Cook and Campbell (1979), where external validity is used to qualify solely the *result* of an experiment. In this formulation, an

[63] In regarding CSR as low in external validity and *high in internal validity* the traditional view differs from the heteronomous paradigm of CSR. The old paradigm, in fact, maintained that CSR also faces serious problems of internal validity.

experiment is externally valid if, and only if, its *results* can be generalized to a broader population. More generally, the recent literature commonly treats external validity as a property of a *whole design* rather than of a *particular application* of it (Lucas 2003).

Assumption 2 asserts a trade-off between internal and external validity. It has been discussed by Jimenez-Buedo and Miller in relation to the use of experiments in the social sciences (2010). They explain the asserted trade-off with the underlying intuition that in the experimental context "the more we ensure that the treatment is isolated from potential confounds in order to make certain that the observed effect is attributable to the treatment, the more unlikely it is that the experimental results can be representative of phenomena of the outside world, since typically, in the outside world, many factors interact in the production of events that we are interested in" (2010: 302).

Assumptions 1and 2 have been carried over to the case study context without clear arguments in their support. In particular, the use of the notion of external validity as an undistinguishable property of the scientific results *and* of the design where those results are obtained has slipped silently into common parlance as if it had no important normative implications. The normative implications are, however, remarkable. It follows from assumption 2 that a design described as having a comparative advantage in one respect is, in virtue of the trade-off, comparatively weaker in the other. Assumptions 1 and 2 together have thus given the scholars above license to qualify CSR as high in internal validity and low in external validity.[64] These assumptions rationalize the methodological prescription that recommends the use of CSR when the main goal is achieving internal validity and of other designs, such as the statistical methods, when the goal is deriving broad generalizations instead.

The soundness of this methodological principle is certainly disputable once assumptions 1 and 2 are also disputed. Assumption 1 has it that external validity is a property of scientific results and of the research method by means of which they have been produced. Remember from section 1.3.2 that scientific findings are externally valid

[64] They are not, by themselves, sufficient for this conclusion.

if, and only if, they are correctly extrapolated from the studied context to a context yet unstudied. Given this definition, it seems illegitimate to use the concept of external validity to refer to the research method as well. One establishes whether results are externally valid by empirical testing on a case-by-case basis; however, it is absurd to regard a research design as externally valid on a case-by-case basis.

Alternatively, on a more charitable reading, scholars speaking of a research design *as being low* in external validity can be interpreted as meaning that the design in question *tends to provide results* that lack external validity. This requires, however, an additional methodological argument for why this would be the case. The argument which case study researchers typically provide appeals ultimately to assumption 3: they assume that external validity is a problem of representativeness, and argue that CSR faces problems at selecting representative samples. *Hence*, it tends to provide results that are poor in terms of external validity. This argument will be discussed in detail when I address assumption 3.

Assumption 1 has been challenged by Jeffrey Lucas (2003) in relation to the experimental design. Although the experimental method is criticized by several scholars for being poor in external validity, Lucas rejects this criticism as essentially misdirected. Specifically, he responds that "critiques of investigative techniques as being low in external validity because findings cannot be generalized quite often should be directed at the theory under test, rather than at the methodology employed to test it" (2003: 238). If findings fail to hold in naturally occurring situations while the theory is supported in well-designed experiments, this suggests that the theoretical framework should be submitted to scrutiny rather than the design itself. The theory might in fact be in need of revision so as to include those variables impacting the phenomenon that it might have overlooked.

Assumption 2 has been addressed by Jimenez-Buedo and Miller (2010). They notice a tension between the widely held beliefs that internal and external validity stand in a trade-off relation and, at the same time, that the former is a prerequisite for the latter. As said above, the general idea behind the trade-off is that creating the conditions to reach higher internal validity in the studied context necessarily leads to results that are less likely to be valid outside of it. This is generally explained by referring to the *artificiality* of the experimental situation. Upon analysis of the methodological literature and the experimental practice, Jimenez-Buedo and Miller conclude that "the notion of artificiality is a rather more elusive concept than we normally acknowledge" (2010: 307); and that the alleged trade-off relation is far less cogent than the common understanding would have one believe.

This argument has a strong, though indirect, bearing on the discussion on CSR. The reasons given for the emergence of a trade-off, in fact, speak against its existence in the context of case studies. If one pays heed to Jimenez-Buedo and Miller, the trade-off between internal and external validity is borne out of the fact that in the experimental context internal validity is obtained *by increasing the isolation* of the experimental system from outer influences. It is then hard to see why such a trade-off should arise in CSR in the first place, where the researcher has very poor control, or no control whatsoever, over the system under study. In general in CSR, and in particular in historical narratives, internal validity *cannot* be achieved, in fact, by increasing the isolation of the factors of interest.[65] If it is true that artificiality trade-offs against external validity, this should be all in favor of CSR.

These criticisms suggest that the methodological norm based on assumptions 1 and 2 does not have the self-evident status that the traditional view presumes.

3.2.2 THE ASSUMPTIONS CARRIED OVER FROM THE STATISTICAL CONTEXT

Whereas assumptions 1 and 2 have been addressed in the literature, the remaining assumptions have not yet been addressed. The discussion that follows therefore focuses on assumptions 3 and 4 which are carried over to the case-study context in the following form:

> 3* A case study is externally valid if, and only if, the case is *representative* of a broader population.

[65] See Chapter 2 for a discussion on the internal validity of historical narratives.

4* The broader the population, the higher the external validity of the case study.

Assumption 3* is a condition for the generalizability of scientific results in CSR. Results that are obtained within a study apply outside of it if, and only if, the studied case *represents* the population[66] in some sense to be specified. The traditional view borrows its idea of representativeness from statistical discourse. The external validity inference is here conceived as an inference from sample to population, legitimized by the former being a statistical representative of the latter.

Translated into a qualitative framework, a case is said to stand in a sample-to-population relation with a universe of cases when it is a *typical* case within that universe. Typicality is therefore understood as the key requirement for ensuring external validity to the case study within the traditional view. Methodological precepts oriented to strengthen the external validity of case studies would all go in the direction of giving rules for the selection of the cases. The following excerpt is an example:

> The *typical case* study focuses on a case that exemplifies a stable, cross-case relationship. By construction, the typical case may also be considered a *representative* case, according to the terms of whatever cross-case model is employed […] One may identify a typical case from a larger population of potential cases by looking for the smallest possible residuals […] for all cases in a multivariate analysis. In a large sample, there will often be many cases with almost identical near-zero residuals […] Thus researchers may randomly select from the set of cases with very high typicality (Seawright and Gerring 2008: 299). [67]

[66] The terms population and/or universe (of cases) are borrowed from the statistical discourse to refer to the set of cases to which the results from a given study apply. I prefer the term target case/s (target for short) because it does not smuggle in the assumption that the hypothesis of external validity necessarily concerns a broader set of cases that relates to the studied case through a sort of sample-to-population relationship. I will switch to the term target below when I introduce my own conditions and criteria for external validity.

[67] See also Mahoney and Goertz (2006b).

Reduced to a matter of representativeness, the problem of external validity thus amounts to adopting the selection procedure that maximizes the probability of choosing the case most typical of the target of interest.

In the traditional view two major problems threaten the external validity of CSR. The first lies in the difficulty of establishing the typicality of the selected case, the solution of which consists in further refining the selection procedure of the case to study. The second is the intrinsic limitation to the degree of external validity CSR can reach. According to assumptions 4 and 4*, the degree of external validity depends on the size of the population to which the results are generalizable. In virtue of assumptions 3* and 4*, CSR is low in external validity because its capacity for selecting representative samples of a sizable population of cases is very limited indeed. Even if one succeeds in identifying typical cases, so the argument goes, their typicality is always confined to a small population. CSR, in fact, studies intensively either one case or a small set whose degree of representativeness is not only hard to establish but also limited. Representativeness, it is said, increases with the size of the sample, and so in turn does external validity.

3.3 EXTERNAL VALIDITY IN CSR: FROM TYPICALITY TO COMPARABILITY

The traditional view treats the problem of external validity as a problem of mere representativeness. This has two major normative implications. First, its methodological precepts as regard to external validity are mainly oriented towards guiding the selection of the "right" case, understood as a typical case. Secondly, the traditional view describes the difficulty that CSR has in putting together a representative sample as the source of the method's incapacity, or the extreme weakness, in achieving external validity.

This reasoning errs already at the first step, and this makes its negative conclusion problematic. External validity is *not*, as I shall argue, essentially a problem of representativeness but rather one of inference, and so a problem of which the representativeness of the case may offer one possible solution. The challenge of external validity actually consists in identifying the circumstances under which the results of a study can be generalized to other cases. The inference from the studied case to some new contexts needs to be justified by some factors that give us reason to believe that what was found true of the former is most probably true of the latter as well. Typicality might be *one* of these factors. The claim that the case at hand is typical backs up the inference to conclude that what is true of the case is also true of entire the population.

Finding the typical case is therefore a practical solution to what is truly an epistemic problem. Typicality is a solution to the problem of generalizability; typicality per se is not the ultimate problem to solve. The traditional view conflates these two concepts—the typicality of a case and its generalizability, and in so doing, not only fails to capture the essential distinction, but also confuses one solution with the entire problem. As a consequence, the methodological norms it imparts to guide the selection of the case cannot respond to the epistemic challenge of external validity.

The traditional view confines the methodological discourse on external validity to the stage of the selection of the cases and in so doing implicitly suggests that the problem of external validity is fully solved by singling out the representative case from a population of cases. Representativeness, however, only offers a solution if the strategy used to establish it properly responds to the epistemic challenges posed by external validity. That is, the typical case cannot be identified by presupposing knowledge that its identification is expected to deliver in the first place. The problem with the strategy described by Gerring in the excerpt above is exactly this one. What he suggests for the selection of the typical case presupposes knowledge of the cases that we are not supposed to possess when the problem at hand is correctly described as one of external validity. If we already know the causal relationship we are interested in generalizing, there is nothing left to generalize in the first place. Gerring's strategy probably solves successfully issues of representativeness, but cannot double up as a solution to an inferential problem.

The scholars subscribing to the traditional view fail to respond to this challenge, because they fail to distinguish conditions for the external validity of the results and epistemic criteria that help establish whether these conditions hold. The conditions for external validity are the circumstances that justify the generalization; typicality is the one explicitly acknowledged by these scholars:[68]

> Condition for External Validity (CEV): If the case is typical of a broader universe of cases, then the case-study result is generalizable to the universe of cases.

Typicality, however, cannot double up as an epistemic criterion for the assessment of external validity. Once the conditions for the generalizability of the results have been defined, independent strategies should be devised that help establish whether those conditions hold. These are epistemic criteria that inform us about the representativeness of the case and, at the same time, do not presuppose the knowledge of the universe of cases that we are expected to extract from the case study itself. This criterion is comparability.

> Epistemic Criterion (EC):[69] Comparability of the study is required to establish whether the case is typical of the universe of cases and the result hence generalizable.

If the case study is *comparable* in the appropriate respects, it enables us to elicit both the information that is to be generalized *and* the information that is required to decide about the generalizability of the same results from the case.

The notion of comparability has been introduced by LeCompte and Goetz (1982) in a work on the validity of the ethnographic methods. LeCompte and Goetz understand external validity in terms of typicality *and* comparability.

> The fieldworker's problem is to demonstrate what Wolcott conceptualizes as the typicality of a phenomenon, or the extent to which it compares and contrasts along relevant dimensions with other phenomena. Consequently, external validity depends on the identification and description of those characteristics of phenomena salient for

[68] Typicality might be taken as a restrictive condition. In the literature on external validity, several scholars endorse similarity instead, which is a broader concept. I discuss the relation between typicality and similarity below.

[69] The epistemic criterion serves to assess whether the conditions of external validity indeed hold.

> comparison with other, similar types. Once the typicality of a phenomenon is established, bases for comparison may be assumed (1982: 51).

LeCompte and Goetz have the merit of hinting to the epistemic problem at the core of external validity but still fail to disentangle fully conditions for validity and criteria of assessment. *Typicality* refers to the condition the case has to satisfy for having results from the study that are generalizable. Typicality, however, cannot be established a priori, nor it can be inferred from knowledge of the universe of cases the generalization itself is expected to provide. The typicality of the case and the generalizability of the results are established upon comparison with other/new cases.

The hypothesis of external validity is truly empirical and has to be settled on a case-by-case basis (Guala 2005, 2010, Steel 2008, 2010). Whether the results of the original study apply outside of it, in fact, depends on matters of fact about the target which, as such, have to be inquired empirically. The idea that external validity is a fixed feature of a given design or can be established a priori, that is without engaging with the empirical investigation of the target, is misled. Claims of external validity always presuppose empirical assumptions about the target contexts even if they are sometimes left implicit or not subjected to scrutiny. Comparability is therefore the epistemic requirement to be imposed on the design of the study in such a way that, by contrasting its results with what we observe in other situations, we are capable to adjudicate the typicality of the case at hand and the generalizability of its results.

By rendering the case study comparable in the appropriate respects, the problem of its external validity becomes decidable. This does not mean that external validity is in this way guaranteed; only that it can be established reliably. The discourse on external validity so far developed among case study researchers is misdirected and therefore not helpful to this end. Required in addition are strategies that are epistemically viable for assessing the typicality of the case and the generalizability of the findings. Disentangling the two issues by distinguishing between typicality and comparability is a first step in this direction.

A second step involves making the notion of comparability more precise. I return to this point below. Before turning to this aspect, however, I want to emphasize a final point in relation to the traditional view. Its focus on representativeness as if it was the ultimate challenge to establishing external validity has biased the debate and led to its current dead end. That is, the consensus view is that the lack of external validity constitutes an irremediable weakness of CSR, attributed to the fact that it studies a very limited number of cases. This conclusion, however, links high external validity to a capacity for offering *broad* generalizations. This assumption about the requisite breadth of the generalization, however, is disputed in the literature.

3.3.1 RADICAL LOCALISM AND EXTERNAL VALIDITY

Influential scholars in the debate on experimental design hold the view that the generalizations that science allows are *always* very limited in scope. In a discussion on the external validity of experiments, Francesco Guala (2002) mentions Bruno Latour, David Gooding, and Andrew Pickering as promoting a form of radical localism, and Ian Hacking and Nancy Cartwright as defending milder positions in a similar spirit. In its extreme version, radical localism denies any external validity to scientific hypotheses except when the outside world can be carefully engineered and made alike the laboratory such that the experimental results can be exported directly (2002: 1196). Guala defends a less skeptical position that admits of several ways to solve the problem of external validity.

According to Guala,[70] the problem amounts to minimizing the error in the inference from the laboratory to the outside world; this is achieved by making the two contexts as similar as possible. One way to this end is what he calls "engineering the world." Another strategy is adapting the experimental setting to the outside conditions. For instance, the former can be modified as to reproduce more accurately non-experimental settings. Even though various strategies exist to generalize reliably from experiments to the outer world, external validity is bound to remain a "local" matter. That is, the generalizations that

[70] Guala mentions Deborah Mayo to be a defender of this view.

science allows never travel too far and never apply too broadly. Case studies would then pose no special problem, as all type of studies possesses only limited generalizability.

Moreover, from the perspective of this alternative approach to external validity, *similarity* between the two contexts is the condition that grants generalizability to the results. CEV can be thus relaxed as follows:

> (CEV)*. If the case is *similar* to the target case/cases, then the results obtained in the former are generalizable to the latter.

Similarity is a broader concept than typicality. Typicality presupposes similarity between the case and its target but further requires a sample-to-population kind of relation between the two. This idea, originating as it does in a statistical context, badly fits CSR, where random sampling is often not a feasible strategy. Furthermore, it is restrictive in that it asks that the relevant population be clearly defined before engaging in the study of the case that is supposedly representative of it. This led to the type of discourse on external validity I discussed above.

Similarity instead does not require any a priori definition of the relevant target and leaves the issue of generalizability open to the empirical analysis that follows the case study. Unlike experiments, in CSR, the case studied and its target cannot be "made" similar as an experiment and the non-experimental setting are. The case, in fact, cannot be adapted in any meaningful sense to the target, whereas, as Guala suggests, the experimental settings can be slightly modified to fit some features of the outside world. And, in general, given the complexity of the phenomena that case studies examine, the idea of engineering the outside world as to reproduce the study conditions is simply not practicable. The way that is open to CSR is finding similarity between the studied system and the target in vivo.

I argued above that the hypothesis of external validity is empirical and to be settled on a case-by-case basis, and it is justified by the similarity between the studied case and the target, which in CSR can only be found in vivo. General conclusions about the external validity of case studies are thus hard to come by: whether the studied case and the target are similar enough is a contextual issue that depends on the epistemic purpose of interest. That is, it depends on the kind of results we are interested in extrapolating from the former to the latter. In this perspective, methodological prescriptions should be directed to inform contextual decisions about the external validity of the hypothesis of interest. In particular, they should help to decide whether upon empirical investigation the studied case and the target are similar enough so that what was true of the former can be reasonably extrapolated to the latter. I shall argue in the rest of the paper that by making case studies *comparable,* one facilitates this type of assessment.

3.4 IMPROVING THE EXTERNAL VALIDITY OF CSR BY ENHANCING COMPARABILITY

On the basis of what is said above, we can thus relax EC as to encompass the broader condition of similarity:

> (EC)*: Comparability of the study is required to establish whether the case is similar to the target case/cases, and therefore whether the result is generalizable.

Even though intuitively appealing, comparability as described by (EC)* is too vague to be useful. We need to refine the criterion further to distinguish what qualifies as a comparable case study and what does not. Furthermore, one needs to specify what makes a case study comparable and what detracts from it. In this way we would propose some principles that might help strengthen the external validity of CSR by making its assessment more reliable.

To judge comparability, it is worth keeping in mind that when assessing the generalizability of a result, one faces severe epistemic constraints. The external validity hypothesis is, in fact, empirically settled by the comparison between the case studied and the target case, of which we know very little. If we knew of the target what we already know of the case, there would be no worries for external validity in the first place. Certainly we do not know whether the result or hypothesis that is true of the case is also true of the target, since this is what one aims to establish. But, in general, any inference of external validity is bounded by the limited knowledge of the target case (Steel 2008). Hence, provided that it is correct, the inferential strategy that requires a minimum amount of information about the target case to establish whether the scientific results at hand are externally valid is to be preferred.

With this proviso in mind, I suggest that comparability requires that the study renders available the information necessary to establish whether, upon comparison, the case is sufficiently similar to the target so as to justify the generalization of the results obtained. Take, for instance, the most common scenario in which the result to be generalized is a causal relationship. The case and the target have to be similar in the respects that are causally relevant to the hypothesis for this to be valid in the latter as well (Guala 2010, Steel 2010). The study then needs to inform us about the respects that matter to the causal relation in the studied case such that we can proceed to the comparison with the target and eventually to the inference.

In CSR, where neither engineering the world is an option nor is adapting the case to the target, knowledge of the relevant causal factors is what enables the inference. [71] Without it, even a tentative assessment of external validity would not be possible. Complete knowledge of the relevant causal factors is, however, an epistemic ideal. Setting the ideal aside, one supposes that the more complete this knowledge is, the more reliable the inference will be.

Provided that comparability is satisfied in the sense described above, we can then distinguish between high and low comparability depending on whether the inference from the study is more or less epistemically efficient: when one can economize on the information needed about the *target case* to make the inference reliable. However, as I shall argue below, this presupposes that the *studied case* be examined more thoroughly in the relevant causal respects. The extent to which causal relationships are investigated thoroughly in the *studied* case determines in turn the degree of comparability of the study itself. Hence, the more information about the *target* case is required in order to establish the external validity of the hypothesis, the lower the comparability of the case study will be. If the case study only describes the causal factors that are relevant for the outcome of

[71] Guala says that in experimental economics this causal knowledge can be sometimes black boxed without implications for the external validity inference (Guala 2010: 1080).

interest, then the comparison between the studied case and the target needs to be fully articulated – that is, contrasted along all relevant respects before the hypothesis can be generalized to the new case. Since epistemic efficiency is a virtue when external validity is at stake, a case study that requires full comparison is low in comparability.

If the study describes the complete causal mechanism instead, then comparison with the target can be partial: it needs only to involve a limited number of features to support the hypothesis of external validity. Recall from Chapter 2 that a causal mechanism consists of the system components and their relationships when associated in such a way as to give rise either to macro-level regularities or to unique series of events that are causally connected[72]. If the mechanism is complete, it specifies the causal links among the relevant factors and identifies the set of conditions that are jointly sufficient to produce the outcome in question: the difference in a subset of these factors can thus disrupt the mechanism. If the study describes the complete causal mechanism, its comparability is higher because partial comparison between the case and the target in a subset of causal factors is sufficient to draw conclusions about the behavior of the whole mechanism. If instead the studied mechanism is underspecified, it becomes much harder to economize on information of the target that is needed in order to settle the hypothesis of external validity.

The examples that follow illustrate the distinction between high and low comparability of case studies.

The first case study on Botswana is an instance of the high type. It provides a characterization of the causal mechanisms in the studied case besides the relevant causal factors. As such, this study would enable us to draw conclusions about the generalizability of the results upon partial comparison between the studied case and the targets. And indeed, the evidence seems to be in favor of this hypothesis. As we shall see, the authors of the study attempt to draw conclusions about target cases on the basis of a limited comparison between the latter and Botswana.

[72] See sections 2.2.2 and 2.4 above.

The second example is meant to illustrate how a case study low in comparability would look like instead. Despite the effort to single out the factors causally relevant to the outcome of interest, there is no description of the related mechanisms. If used to draw conclusions about cases yet unstudied, this work would thus demand a much more comprehensive comparison between the studied cases and possible targets. In particular, such comparison would have to be comprehensive of the full array of causal factors that matter to the outcome of interest before confirming any hypotheses of external validity.

3.5 HIGH AND LOW COMPARABILITY OF CASE STUDIES

3.5.1 THE STRANGE CASE OF BOTSWANA

In a case study of Botswana, Daron Acemoglu, Simon Johnson, and James Robinson (2003) explain the unusually good economic performance observed in the country in the last decades. If compared with the average in sub-Saharan Africa, in fact, Botswana performed outstandingly in terms of per capita income growth rate in the last 35 years. The authors start with the assumption based on previous studies that proximate determinants of its economic success are the institutions and the related policies that the country developed over time. Institutions are conducive to growth when they correspond to a social organization that ensures effective property rights to a broad cross-section of the society. The authors refer to this cluster as *property right institutions*.

What the authors aim to explain is, however, why Botswana was able to develop the institutions it now possesses and thus search for what can be defined as the deep determinants of growth (Rodrik 2003: 3). To this end, they adopt a case-oriented methodology. They eventually offer a country narrative in which they reconstruct the processes through which Botswana developed its institutions.

The examination of the country's history suggests five (structural) features as plausibly responsible for its property right institutions and good economic policies:

1. Botswana is very rich in natural-resource wealth.

2. It had unusual precolonial political institutions that enabled an unusual degree of participation in the political process and placed restrictions on the political power of elites (*Kgotla*).[73]
3. British colonial rule in Botswana was limited. This allowed the precolonial institutions to survive to the independence era.
4. Exploiting the comparative advantage of the nation after 1966 directly increased the incomes of the members of the elite.
5. The political leadership of Botswana Democratic Party (BDP),[74] particularly that of Seretse Khama,[75] inherited the legitimacy of these institutions, which gave it a broad political base.

The causal influence of each of these features on Botswana's modern institutions is established by describing the mechanisms through which this influence is conveyed. These mechanisms, theorized in the background literature on development economics to which the authors refer, are used to explain why these factors are causally relevant to the emergence of property rights institutions (Acemoglu and Robinson 2000). Consider, for instance, the mechanism of *political losers*:

> An institutional setup encouraging investment and adoption of new technologies may be blocked by elites when they fear that this process of growth and social change will make it more likely that they will be replaced by other interests - that they will be political losers. Similarly, a stable political system where the elites are not threatened is less likely to encourage inefficient methods of redistribution as a way of maintaining power (Acemoglu, Johnson, and Robinson 2003: 103).

According to the mechanism of political losers, political elites do not oppose the adoption of institutions and policies favorable to growth if they feel that their power is not threatened by the change; that is, they fear not being political losers. In Botswana

[73] *Kgotla* is an assembly of adult males in which issues of public interest are discussed (Acemoglu et al. 2003: 93).
[74] BDP is the dominant party in the country.
[75] Seretse Khama was BDP political leader and president of Botswana from 1965 until 1980.

(feature 2) precolonial institutions ensured some degree of political stability and went almost unaffected by the imposition of British colonial rule (feature 3). In addition, the legitimacy of Seretse Khama and the broad coalition he formed further strengthened it (feature 5). The authors conclude that features 2, 3, and 5 influenced the building of property right institutions by setting the mechanism of political losers at work and thus ensuring a high degree of political security to the existing elites. A similar use is made of the other two theoretical mechanisms: they trace the causal relations among the factors that jointly determine the emergence of property rights institutions and the ensuing economic growth.

In the final section of the work, Botswana is compared with four African countries - namely, Somalia, Lesotho, Cote d'Ivoire, and Ghana. The authors apply the lessons drawn from Botswana to these countries by way of a partial comparison with each target case. They reason along the following lines: the difference in one feature is regarded as sufficient to infer the disruption of the related mechanism and explain the failure to develop property right institutions (and eventually experience long term growth) in the target. On this assumption, each country is compared to Botswana with respect to one or two of the features above, and never along all five dimensions before drawing a conclusion about why property right institutions failed to develop in the circumstances. As an illustration of this form of reasoning, consider the mechanism of *constraints* that Acemoglu, Johnson, and Robinson (2003) describe as follows.

> When (precolonial) institutions limit the powers of rulers and the range of distortionary policies that they can pursue, good policies are more likely to arise (see Acemoglu and Robinson 1999). Constraints on political elites are also useful through two indirect channels. First, they reduce the political stakes and contribute to political stability (mechanism of *political losers*), since, with such constraints in place, it becomes less attractive to fight to take control of the state apparatus. Second, these constraints also imply that other groups have less reason to fear expropriation by the elites and are more willing to delegate power to the state (ibid.: 104).

The mechanism of *constraints* is set into operation by feature 2 that is, by the type of precolonial institutions. If these institutions have the right properties, such as *kgotla* had in Botswana, then they are effective in placing constraints on rulers. These, in turn, affect the emergence of property right institutions both directly and indirectly. At the same time, feature 3 can have the opposite effect of inhibiting the mechanism of constraints. In fact, strong British colonial rule alters precolonial institutions and disrupt the mechanism of constraints if it was in place before.

Acemoglu, Johnson, and Robinson thus compare Botswana and Somalia with respect to features 2 and 3 and find them similar in feature 3 but different in feature 2. Similarly to Botswana, British government had in fact only marginal interest in Somalia and imposed very soft colonial rule. Somalia had, however, precolonial institutions that induced intense factional conflict and were therefore incapable of placing constraints on political elites. From this partial difference in the type of precolonial institutions between the two countries (feature 2), Acemoglu, Johnson, and Robinson infer that the mechanism of *constraints* was not operating properly in the country and it thus impeded also the proper working of the mechanism of political losers. As a consequence, property right institutions did not emerge and in turn its economic performance faltered.

A similar line of reasoning is then used in the comparison with the other countries. The comparison is limited to one or two structural features whose difference is taken as evidence that the corresponding mechanism is not operating properly in the target and thus explains the absence of property right institutions and the ensuing bad economic performance. Features 1 to 5 are thus treated as INUS conditions: non-redundant components that together produce the outcome.[76] If one of the conditions is missing, the mechanism is disrupted and the outcome will not obtain. The reverse is also true. If one factor has to be absent for the outcome to obtain and is instead present, the mechanism is equally disrupted. If the case study successfully identifies INUS conditions, a minimal comparison is sufficient to draw causal conclusions about the target.

[76] See section 1.3.3 for a discussion of INUS conditions as relevant evidence for policy-making.

It seems that Acemoglu, Johnson, and Robinson reason in a similar way to what Daniel Steel (2008) calls *comparative* process tracing. Whereas process-tracing is a within-case strategy of causal inference,[77] *comparative process tracing* serves the purpose of generalizing findings from a studied context to a context the knowledge of which is indeed very limited. In particular, comparative process tracing consists in comparing the mechanisms in the study and the target at the "critical junctures". From this limited comparison, one can establish the presence (or absence) of the mechanism in the target and, thus, the presence (or absence) of the causal relationship that ensues. In this case study, Botswana is compared to each target case in only a limited number of features involved in the working of the mechanisms that are conducive to growth through the emergence of property right institutions.

Partial comparison in the relevant features between Botswana and the target cases is used by the authors to explain that bad institutions and poor growth are caused in the latter by the incorrect operation of the related mechanism. Whether this inference is ultimately justified, however, depends on whether the following assumption is also vindicated. The extrapolation of the mechanism in fact presupposes causal homogeneity across the contexts: causal factors that are alike would operate together in similar manners in the studied case and in the target. Is this assumption justified? Even though the authors do not explicitly discuss it, they hint to some facts they seem to regard as evidence for it. They seem to suggest, in fact, that cultural, geographical, and social affinities among the countries considered are such that had similar factors been in place in the studied and the target cases, they would operate together in similar ways.

Whether this assumption justified or not, this study can be said to attain a higher degree of comparability than the one I will discuss below where equally reliable conclusions could only be achieved at a higher epistemic cost. Provided that the assumption of causal homogeneity is satisfied, in the study above the comparison across cases can be performed in a limited number of respects, because the mechanisms identify the causal links among

[77] Cf. Chapter 2 for a discussion of process-tracing as a method of within-case causal inference.

relevant variables and thus isolate those that if absent would be sufficient to disrupt the mechanism.

3.5.2 COMMUNITY BASED PROGRAMS TO DEFEAT MALNUTRITION

The second study formulates a policy hypothesis about the effectiveness of community-based programs to defeat malnutrition. It is a report by the World Bank in the early Nineties on successful nutritional programs in Africa (Kennedy 1991). The study aims to identify the factors that make programs against malnutrition work. To this end, it combines the use of two qualitative methodologies --namely, large sample survey and case-oriented studies. Eileen Kennedy (1991) motivates the study by appealing to the fact that the literature on malnutrition in the 1970s and 1980s focused only on the types of interventions implemented. The 1989 meeting of the International Nutritional Planners in Seoul, however, concluded that "how" a program is implemented is as important as, or maybe more important than, the type of intervention for successful programming (1991: 1). In line with the recommendations of nutritional planners in Seoul, Kennedy uses survey and case studies jointly to identify what factors matter for the effective implementation of programs against malnutrition. The ultimate goal is to learn lessons that can be generalized to other African contexts (1991: 2). The evidence is combined in the following way.

The survey offers *prima facie* evidence[78] of what factors are required for successful implementation. It combines the findings from 110 answers received from policy makers and program implementers. The respondents answered two types of questions: whether the program was successful and what factors they thought were key to its success.[79] Furthermore, six programs among the 66 found to be successful were selected for an in-depth analysis: the Macina Child Health Project, in Segou Region, Mali; the Infant Feeding Project, in Togo; the Imo State Child Survival Project, in Nigeria; the Applied Nutrition

[78] Following Reiss, *prima facie* evidence is relevant to the hypothesis at hand but still defeasible, and hence not decisive for it (Reiss 2008).

[79] 110 were the individuals/institutions who responded to the mail survey out of the 330 initially contacted. In addition to this first pool, 78 individuals involved in various ways in the implementation of programs against malnutrition were interviewed.

Program, in Ghana; the Mali Institutional Development Enterprise and Nutrition Program, in Mali; and the Nutrition Project, in Kinshasa, (formerly) Zaire. These projects were selected out of the 66 because they represent different types of community-based programs that can succeed in combating malnutrition. The survey and additional interviews singled out seven factors as important for successful implementation: community participation, program flexibility, institutional structure, recurrent cost recovery, multifaceted program activities, training and staff qualifications, and infrastructure (Kennedy 1991: 7). The same factors were found to be present in almost all six cases. The case studies then focus on how the seven factors were actually implemented in the specific context.

The main conclusion drawn from the report is that different types of programs can succeed in defeating malnutrition. Whether they succeed depends on a set of conditions regarding how the programs are in fact implemented. The study, however, cannot be considered in itself sufficient to establish which of the factors identified are necessary for the effectiveness of the program: it offers at best *prima facie* evidence that these factors are causally relevant to the outcome in question. It thus gives some support to the idea that there is not such a thing as one-size-fit-all intervention, but, rather, different ways can be pursued.

Has the ultimate goal of this study been achieved? That is, are we learning lessons that are generalizable to the other African contexts (1991: 2)? This is an empirical question: whether these causal hypotheses are justified in other situations not yet explored is a question to be settled on a case-by-case basis, upon comparison between the studied context and the target cases. Such comparison is not attempted by the author; however some observations can be made as regards the comparability of this study.

The strength of this study derives from the effort it makes to single out factors that matter to the successful implementation of the programs against malnutrition. We can consider the study comparable in the sense described above. By listing the set of causal respects that matter to the phenomenon of interest, it enables the comparison with new cases along the relevant dimensions. In this way it would help formulate conclusions about what outcome would occur in the target case. Whether these conclusions are justified depends on assumptions about causal homogeneity between contexts similar to those discussed in the case study above. Even if these assumptions were justified, however, this study does not enable the type of efficient comparison that the case of Botswana allows. There is no evidence, in fact, that the factors identified are INUS conditions for the outcome. The complete description of the relevant mechanisms would facilitate the identification of the set of non-redundant factors that are jointly sufficient for the phenomenon of interest.

Kennedy's case studies, however, fail to characterize the mechanisms that explain why the factors identified matter to the success of the program and how. Their contribution is limited to a more or less exhaustive list in which the relevant factors are described in detail, and so are the modifications in the specifics of implementation that occurred over time. An analysis of the causal mechanisms at work is however lacking, despite the fact that this was set as a research goal at the very beginning of the paper (Kennedy 1991). Different from the case of Botswana discussed above, the causal relationships among the factors of implementation are not specified and the underlying causal mechanisms are not even hinted at. As a consequence, even though we can consider these studies to some extent comparable to new target contexts, they have low epistemic efficiency.

In order to decide about the generalizability of the causal findings obtained in this study, one should then perform a fully-fledged comparison of the relevant causal factors between the studied contexts and the target cases. There would be no other way, in fact, to establish whether the causal result is likely to obtain also in the new situations. Knowledge of the mechanisms responsible for the causal relationships of interest would avoid this cumbersome epistemic strategy and render the comparison more efficient in this respect. One could in fact examine the mechanisms at the critical junctures in the target cases and infer on this ground whether the causal relationship of interest is likely to obtain. The study by Kennedy does not allow this more efficient comparison, even if it reveals information that is ultimately crucial to formulate judgment of external validity, namely knowledge of the relevant causal factors.

3.6 CONCLUSION

The debate on the external validity of CSR is stunted, and, so far, developed under the influence of the statistical perspective. The resulting approach was biased in an unfruitful direction that ultimately led the debate to the dead end, where it seems to stand now. Reaching external validity in CSR was, in fact, essentially regarded as a hopeless endeavor and the downside of CSR. This conclusion stands in a stark contrast with the struggle for generalizations in which case study researchers engage at the same time. In this chapter, I argued that this conclusion is unjustified once we examine more carefully the assumptions on which it is based. The bottom line of this discussion is not that CSR has high rather than low external validity as the traditional view maintains. Rather, in the perspective I defend here this type of judgment is not very appropriate in the first place.

First, external validity is not a property of a research design. Furthermore, if one accepts some form of localism, the generalizations that science allows never travel too far and never apply too broadly. Moreover, whether certain results have external validity *in fact* is an empirical matter that can only be settled on a case-by-case basis. The problem to solve then is not how to increase the external validity of CSR; rather, it is making external validity a decidable issue. This can be done once we abandon the old paradigm, the traditional view, and its focus on representativeness as *the* problem of external validity. External validity becomes decidable only if the study first renders available the evidence that is necessary to circumvent the epistemic impasse in which any inference of this kind finds itself. Making case studies stronger in external validity therefore means strengthening the design in such a way that it helps researchers reach judgments of external validity with a higher confidence and epistemic efficiency. One way to this end is by enhancing its comparability.

External validity should be right on the agenda of philosophers and methodologists who worry about making of case study research a design better understood and better used. Generalizing is not only a valuable goal in itself for scientific practice but also the middle step on the way to sound policy making. Drawing the right lessons from the contexts we know already is to some extent presupposed by planning effective interventions in contexts with which we are not acquainted yet. One way to learn how to use the knowledge obtained from studied contexts in unknown situations is by discussing issues of external validity. The inferential problem that it underpins is in fact the first the scholar encounters when transferring the knowledge gained from epistemically privileged systems to less privileged ones. Experimentalists were the first to worry about it, and they still reflect upon it extensively. Case study researchers should do the same.

4. ON THE EXTERNAL VALIDITY OF CAUSAL EFFECTS AND THEIR RELEVANCE FOR POLICY MAKING

4.1 INTRODUCTION

According to many commentators development economics is in a dismal state (e.g. Cohen and Easterly 2009; Deaton 2010). The recent upsurge in the use of randomized controlled trials (RCTs) across the social sciences raised the hope for an improvement of the allegedly infelicitous state of this field. RCTs, its defenders believe, are likely to deliver the recipes for growth that eluded research on development so far (Banerjee 2007). RCTs would provide the hard evidence required for sound policy making because by measuring accurately causal effects, RCTs supply valuable information on the effectiveness of development policies. More cautious scholars, however, object to what they regard as a burst of unjustified optimism. They contend that RCTs' capacity to cure the ills of development economics and promote effective policy making is in fact very limited, one of the main reasons being the lack of external validity of their results.[80]

Several strategies have been proposed to overcome this problem. They aim at improving the external validity of causal effects so as to render them relevant for purpose of policy making. In this chapter, I examine and assess these strategies. Whereas it can be conceded that they succeed to some extent in addressing the issue of external validity, they fail to fulfill the demand for relevance. The proposal made by Nobel Laureate James Heckman stands out as an exception in this regard. By defining a broader range of causal effects, the model Heckman outlines can address some of the questions that policy makers actually face. As I shall argue, however, it is still inadequate for the most pressing problems policy makers face in the context of development. In these cases, in fact, problems of planning trump the problems of prediction that Heckman's model is designed to address.

This chapter is organized as follows. In section 4.2, I discuss the problem of relevance that RCTs face in relation to their lack of external validity. In section 4.3, two strategies to

[80] RCTs are also criticized in other respects, e.g. the fact that they are gold standard in most evidence hierarchies.

increase the external validity of causal effects are examined, and found lacking in terms of relevance. In section 4.4, I present Heckman's model for causal effects as a response to the Rubin-Holland model which is the template for the analysis of every RCT. In section 4.5, I argue that Heckman's model constitutes a more promising alternative to the other strategies on offer, even though its relevance is still too limited to respond adequately to policy makers' needs. In section 4.6, I draw a distinction between problems of prediction and problems of planning, argue that in developing contexts policy makers confront first and foremost problems of planning, and Heckman's model cannot be of help in addressing those. In section 4.7, I present a sketch of the planning process and the type of evidence it draws upon, and argue that case studies provide relevant evidence in its earlier phases. Section 4.8 concludes.

4.2 ON THE EXTERNAL VALIDITY AND RELEVANCE OF RCTs

The widespread and increasing popularity of randomized controlled trials (RCTs) in the scientific fields where practical concerns are most prominent is to a large extent driven by the expectation that RCTs can improve policy making remarkably. Development economics is an interesting case in point. It is in fact a field directed to a large extent to further the solution of practical problems which according to many commentators have so far suffered from a long series of failures (Cohen and Easterly 2009).

RCTs recently spread in this area thanks to a team of young scholars which founded in 2003 the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute for Technology. Research at the Poverty Lab and in its regional centers consists in planning and performing RCTs to evaluate the impact of programs and policies in areas relevant to the economic development such as agriculture, microfinance, health, education, the labor market, governance, and so on.[81] Abhijit Banerjee, one of the directors of the Poverty Lab, expresses the goals and motivations of their enterprise as follows. He argues that the aid that flowed into developing countries in the past decades failed to meet its goals because

[81] In 2014 the Action Lab has almost 500 completed or ongoing evaluations in 56 countries around the world. See http://www.povertyactionlab.org/

policy makers were clueless as to which projects are effective in promoting growth and erasing poverty. He further adds that, by evaluating policies and programs rigorously, RCTs finally provide the hard evidence that is needed for, and which was lacking so far, "making aid work" (Banerjee 2007).

Banerjee's appeal to the relevance of RCTs for effective policy making rests on two related claims. First, RCTs are taken to provide the *type* of information policy makers need in their practice. Secondly, (only) the information RCTs provide is regarded as sufficiently reliable for the goals at hand. RCTs aim at establishing whether a given program is effective and would thus supply the policy maker with evidence of *which* program works. They do so by evaluating the impact of a given policy on the outcome of interest.

4.2.1 MEASUREMENT OF CAUSAL EFFECTS IN RCTs

The quantity RCTs aim at measuring is typically known as *causal effect*. The notion of causal effect was first defined formally in the Rubin-Neyman statistical model and was subsequently popularized by Paul Holland (Holland 1986). The Rubin-Holland model (RH) posits that, in a population of units Ω, two causes[82] denoted by t (treatment) and c (control) are potentially administered on each unit ω in $\boldsymbol{\Omega}$. The model defines:

- S is the variable that indicates the cause to which each unit in Ω is exposed. Hence, $S = t$ indicates that the unit is exposed to t and $S = c$ indicates that the unit is exposed to c.
- Y is response variable that measures the effect of causes.
- Since the value of the response variable is potentially affected by the particular cause, t or c, to which the unit is exposed, Y_t and Y_c are introduced as the potential outcome variables to measure each effect.[83] $Y_t(\omega)$ is the value of the response that would be observed if unit ω were exposed to t; $Y_c(\omega)$ is the value that would be observed were ω exposed to c.

[82] Holland uses the terms cause and treatment interchangeably (Holland 1986: 946).

[83] Holland calls Y_t and Y_c *response* variables.

RH defines the *individual causal effect* of t (relative to c) on ω (as measured by Y) as follows:

$$Y_t(\omega) - Y_c(\omega) \tag{1}$$

The causal effect (1) is the difference in the values that the response variable Y would take upon administration of treatments t and c on unit ω. Thus, (1) measures an *individual* causal effect. The causal effect so defined faces a crucial problem of measurability. The "fundamental problem of causal inference" (1986: 947) is that the individual causal effect cannot be measured due to the impossibility of observing the effect of two distinct treatments on the *same* unit ω. A strategy has to be devised to evaluate an acceptable approximation to (1).

Holland suggests two solutions to this empirical problem. The "scientific" solution consists in formulating assumptions that allow treating different units as if they were the same. Typical assumptions are:

- Temporal stability;
- Causal transience;
- Unit homogeneity.

Temporal stability asserts that the value $Y_c(\omega)$ does not depend on the moment in which c is administered on unit ω and Y measured. Causal transience asserts that the value $Y_t(\omega)$ is not affected by ω being previously exposed to c and Y_c measured on ω. If these two assumptions hold, it is possible to estimate the individual causal effect of t (relative to c) on ω (as measured by Y) by exposing ω sequentially to c and t and measuring Y after each exposure (648). Alternatively, assuming unit homogeneity amounts to asserting that $Y_t(\omega_i) = Y_t(\omega_j)$ and $Y_c(\omega_i) = Y_c(\omega_j)$ for each i and j in Ω. The causal effect of t is then $Y_t(\omega_i)$ - $Y_c(\omega_j)$. By assuming temporal stability and causal transience or, alternatively, unit homogeneity the individual causal effect can thus be estimated.

The "statistical" solution, adopted in RH and the ensuing statistical literature,[84] consists in redefining the causal effect at the population level and evaluating its expected value.[85] RH thus defines the *average causal effect* T of t over Ω as the expected value of the difference $Y_t(\omega) - Y_c(\omega)$ over the ωs in Ω (1986: 947):

$$T = E(Y_{\mathrm{t}} - Y_c) \qquad (2)$$

(2) can also be expressed as follows by applying the usual rules of probability:

$$T = E(Y_t) - E(Y_c) \qquad (3)$$

The observed data (S, Y_s)[86], however, can only give us information about:

$$E(Y_s/S = t) = E(Y_t/S = t) \qquad (4)$$

and

$$E(Y_s/S = c) = E(Y_c/S = c) \qquad (5)$$

(3) is correctly estimated only if $\mathrm{E(Y_t)}$ is equal to $\mathrm{E(Y_t/S}=t)$ and $\mathrm{E(Y_c)}$ is equal to $\mathrm{E(Y_c/S}=c)$. It is possible, however, that $\mathrm{E(Y_t)}$ differs from $\mathrm{E(Y_t/S} = t)$ and $\mathrm{E(Y_c)}$ differs from $\mathrm{E(Y_c/S} = c)$ since some sub-population in Ω is exposed to t, and the same is true of c. The evaluation of (3) requires the assumption of *independence*, which asserts that S is independent from any other variables over Ω. In particular, this means that the value $S(\omega)$

[84] RCTs belong to this tradition.

[85] In the second section of the article Holland introduces probabilistic concepts: "A probability will mean nothing more nor less than a proportion of units in Ω. The expected value of a variable is merely its average value over all of Ω. Conditional expected values are averages over subsets of units where the subsets are defined by conditioning in the values of variables" (1986: 945).

[86] The causal indicator variable S determines which value, Y_t or Y_c, is observed for a give unit. If $S(u) = t$, then Y_t is observed, and if $S(u) = c$, then , then Y_c is observed. Thus the observed response on unit ω is $Y_{s(u)}(u)$. The *observed* response variable is therefore Y_s. Even though the model contains three variables, S, Y_t, and Y_c, the process of observation involves only two, S and Y_s.

does not depend on Y_t and Y_c. If an adequate implementation strategy to assign the units to *t* and *c* such as a hypothetical randomization is carried out correctly, *independence* is satisfied and the average causal effect (3) can be estimated.

Causal effects can be measured via a variety of methods (Morgan and Winship 2007), for instance causal graphs, regressions, matching estimators, and RCTs. RCTs are experiments in which subjects are typically divided into two groups, one of which is administered a treatment and the other a control. In particular, RCTs measure *average* causal effects. Arguably, they do so quite reliably and on this further ground Banerjee, among others, defends their relevance.

Indeed, it is often claimed that the attractiveness of RCTs largely resides in the rigor of the method (Cartwright 2007a, 2007b). RCTs are experiments of a special sort where subjects are assigned to the treatment and control groups by way of a randomizing procedure. Hence, if the probability of the outcome is higher in the treatment than in the control group, under certain conditions one can safely attribute causal efficacy[87] to the policy implemented. These conditions comprise:

a) A metaphysical premise asserting that probabilistic dependence calls for causal explanation;

b) The assumption that all factors causally relevant to the outcome (other than the treatment) are distributed identically between the treatment and control groups.

c) The probability of the outcome in the treatment and control groups is inferred correctly from observed frequencies.

According to Nancy Cartwright ideal RCTs, RCTs that are carried out in ideal conditions where assumptions a) to c) are met, are *clinchers*: the positive results in the experiment deductively imply the causal conclusion (2007a, 2011). Whether the assumptions are met in practice, and a real RCT clinches its conclusion, depends on the

[87] I follow Cartwright in the use of the terms efficacy and effectiveness. Efficacy refers to whether a treatment, or policy, causes a given outcome in the studied population under the selected circumstances; effectiveness refers to whether the treatment, or policy, causes a given outcome in the target population (2009a, 2009b).

success of the implementation procedure. The additional feature that according to Cartwright renders RCTs more attractive than other clinchers is that RCTs are *self-validating*. A clincher is self-validating if, and only if, the justification for its assumption is provided *within* the design itself rather than outside of it.[88] Most clinchers, such as econometric modeling, are not self-validating in this sense since the causal assumptions on which they are based are justified by knowledge obtained outside the study design. The RCTs design *envisages* procedures that, if properly carried out, maximize the probability that the assumptions are met in the practice: support for the assumptions is thus "built in" the study design (2011: 1400).[89]

4.2.2 EXTERNAL VALIDITY AND RELEVANCE OF CAUSAL EFFECTS IN RCTs

Critics challenge the usefulness of RCTs from various angles. A first set of criticisms points to the fact that using RCTs, even if desirable, is not always a viable methodological strategy. One objection is thus directed at their limited applicability: RCTs cannot be used to assess the impact of certain kinds of programs such as monetary policies, trade policies, a different political system, and so on. RCT design mandates that the treatment, or program, be administered randomly to the experimental subjects so as to render the treatment and control groups causally homogeneous. The random administration of national or regional programs such as monetary policies, trade policies, political reforms is nearly always unfeasible either because of practical constraints, or because it might undermine the program efficacy in the first place. Furthermore, RCTs can be impeded by ethical concerns related to the withdrawal of the treatment/policy from part of the population involved in the experiment (Banerjee 2007).

[88] The metaphysical assumption is not subject to self-validation. It depends on the theory of causality one endorses.

[89] The procedures comprise the use of statistics to retrieve probabilities from frequencies, the use of randomizing devices to allocate subjects in the two groups, quadruple blinding, and so on (Cartwright 2011: 16-17).

A second set of criticisms reacts to the claim which is often made that *only* RCTs provide the hard evidence needed for effective policy making. Critics suggest that other methodological options equally reliable are at hand. Methods that provide hard evidence of the kind RCTs supply would include the econometric methods, derivation from established theory, Galilean experiments, and so on (Cartwright 2007 a, b).[90] As an example consider the reconstruction of instrumental variables by Reiss (Reiss 2008, 2013). Instrumental variables are used for causal inference in econometric regressions when the relationship between two variables of interest is potentially confounded by other variables (2013: 129). Reiss proves that if a set of rather demanding assumptions is satisfied, the correlation between the variables of interest entails causation. Imagine that we are interested in whether X causes Y, and choose variable Z as an instrument for (X, Y). The instrumental variable approach delivers valid causal inference if the following assumptions are satisfied:

a) Z causes X

b) Z causes Y if at all only through X (that is, not directly or via some other variable)

c) Z is not itself caused by Y or by a factor that also affects Y (Reiss 2013: 169).[91]

Instrumental variables so defined are *clinchers* in Cartwright's sense: if the assumptions are met, a correlation between Z and Y deductively implies the causal conclusion.

A third set of criticisms holds that, even if applicable, RCTs do not constitute an effective strategy for improving policy making, the (main) argument being that RCTs are lacking in external validity. This charge deserves attention for two main reasons. First, not only the opponents but also the promoters of RCTs agree that their external validity is severely constrained (Banerjee and Duflo 2009). Moreover, it hits at the heart of the

[90] See Deaton for a critique in a similar spirit (annual-conference-2012-debates-in-development). For a critique of the alleged superiority of RCTs over the other methods of causal inference see Scriven (2008).

[91] In addition to a) to c) above, three general assumptions need also to be satisfied. They comprise the "Reichenbach principle", which concerns the relationship between probabilities and causality, the assumption of transitivity, and functional correctness (for a discussion, see Reiss 2008: 132-3).

arguments of Banerjee and others who defend the prominence of RCTs on account of their alleged usefulness for combating poverty in developing countries.

The notion of external validity was introduced and discussed in Chapter 1 (see 1.3.2). Recall that the general idea behind this notion is that a scientific hypothesis has external validity if, and only if, it is correct in asserting that the results that hold in the sample where they were first established also hold outside of it. To illustrate the notion in the present context, imagine that the RCT measures the impact of using bed nets on the probability of contracting malaria among hospitalized women in Northern Kenya. Imagine further that the RCT delivers a positive result (that is, there is a positive difference in outcome between the treatment and the control group) and that we trust the procedure. The RCT thus establishes that bed nets are a preventer of malaria; it does so, however, in a circumscribed and very specific context. The hypothesis of external validity has it that bed nets are a preventer of malaria in the experimental population, the hospitalized women in North Kenya, *and* in other populations as well (for instance shepherds in Swaziland and tribes of hunters in New Guinea).

It might, however, be the case that the result obtained depends on features of the experimental setting and the specific way in which the program was implemented in the experimental context that might fail to hold elsewhere. Bed nets might work among Kenyan women because they were instructed on how to use them and their compliance monitored in the hospital; once freely distributed among fishermen in Kerala, they might, however, turn ineffective because they are used as fish nets instead. If settings and strategies of implementation are dissimilar as they indeed often are, the causal result might be bounded to the experimental context and fail to travel outside of it.

External validity and relevance are issues tightly connected. Relevance was defined in Chapter 1 as the *adequacy* of scientific results to further a given purpose, solve a specific problem, or fulfill a goal of interest. While discussing the relation between relevance and validity, I suggested that it is reasonable to argue that the former *presupposes* the latter.[92] If

[92] See section 1.3.3 above.

this is true, relevance would also presuppose the external validity of scientific results that, established in a context different than the context of interest, would be adequate if valid for the task at hand.

If the results obtained by means of RCTs lack external validity, and thus fail to hold outside the experimental context, it is hard to see how they can be of any use to the policy maker, at least in an obvious and straightforward manner. If the goal is fighting malaria in Kerala, it is not enough to know that bed nets proved to be a preventer in Kenya: this fact *per se* has no bearing on the task in question, unless some kind of connection is established between what was observed in Kenya and what will happen in Kerala. The objection often raised to RCTs is that they offer poor help in this regard. According to many the evidence they provide is insufficient to justify the hypothesis of external validity that is, the claim that the result obtained therein would also hold elsewhere. It is thus understandable that, when RCTs are waved by their promoters as the most effective means to combat poverty world widely, the RCT-skeptic feels somehow uneasy.

4.3 IS EXTERNAL VALIDITY A PROXY FOR RELEVANCE?

The argument above should give an intuitive feeling of why external validity might well be a concern if one worries about the relevance of RCTs for policy making. And, indeed, both the RCT-enthusiasts and the skeptics regard external validity as an inescapable issue, try to confront it and possibly improve the use of RCTs with respect to their external validity. It will emerge from the discussion that follows that common assumption in both camps is that relevance is a sort of "spill-over" effect of external validity. In both camps, in fact, strategies are proposed to strengthen the external validity of causal effects under the presupposition that relevance will automatically ensue. In other words, the shared view has it that causal effects that are externally valid are *ipso facto* relevant to policy makers.

External validity and relevance remain, however, distinct concerns. It is not external validity *as such* that confers relevance onto given results, even though it might be a helpful, and maybe even necessary, ingredient for effective policy making. The issue of what evidence is relevant for a given task should be tackled independently, and prior to, the question of external validity of the same evidence.[93] Failure to discern this fact led to the belief that, by improving the external validity of scientific results, one also provides what policy makers need. As I shall argue throughout this chapter, this is a gross mistake.

In the RCT-enthusiasts camp, the proposal was made that *replication* is the strategy that solves the problem of external validity that plagues RCTs (Banerjee 2005, 2007, Banerjee and Duflo 2009). The general idea is that repeating a given experiment across different settings is liable to increase the external validity of the causal conclusion. More precisely, if positive results about the efficacy of the treatment are repeatedly obtained when the same experiment is run in settings that differ in some (relevant) respects, the hypothesis that the causal effect thereby measured has external validity is (eventually) corroborated.

The proposal remains vague in several respects. One would assume that the settings are to be different in "meaningful" ways and the experiment run a "sufficient" number of times before the external validity hypothesis is regarded as corroborated. On intuitive ground, it is plausible that one might be interested in running the experiments in new settings that approximate the context of intervention; or in settings where changes are made in features that are likely to be causally relevant. Replications, however large, that do not encompass this sort of changes seems not to be very helpful, in the end, to corroborate the hypothesis of external validity. No clarification is, however, provided in these regards by the promoters of this strategy (Cohen and Easterly 2009).

Furthermore, these scholars do not clarify what hypothesis of external validity they specifically have in mind. It is legitimate, in fact, to try to export from RCTs *qualitative* causal conclusions about the causal effectiveness of a given program. These conclusions state that the program causes the outcome in the target population as it does in the experimental population.[94] It is also legitimate to try to export *quantitative* causal conclusions about some measure of improvement in the average effect that is expected to hold in the target population as it was in the experiment. Each type of conclusion is

[93] See section 1.3.3 for the difference between validity claims and relevance claims and the type of considerations on which they are based.

[94] It is more accurate to say that the causal conclusion one elicits from an RCT is that the program causes the outcome at least in some member of the experimental population (Cartwright 2007a: 15).

justified by different, and quite demanding, assumptions about the causal features in the target population, information that RCTs on their own do not deliver. To be clear on what causal conclusions one wants to export from an RCT is therefore important because different hypotheses of external validity require different types of justification.[95]

More to the point, this strategy is *haphazard* if seen from the viewpoint of the policy maker. It certainly cannot be denied that results that happen to be persistent across contexts increase somehow the confidence in the generality of the phenomenon. The generality of a given phenomenon offers, however, evidential ground for the expectation that it will also hold in the context of interest, if justified by further assumptions about the similarity between the (set of) contexts where the phenomenon was first established and the target.[96]Moreover, the results this strategy delivers are not perspicuously relevant. Replications of the RCTs are typically implemented by changing the location and/or modifying some feature of the experimental setting. The kind and degree of modifications are however steered and constrained by the experimental setting itself, rather than the policy maker's purpose and interest. Feasible modifications are only the ones that fall under the experimental control. Even if general results are thus obtained,[97] they are likely to satisfy more the experimenter's curiosity than the policy maker's needs.[98] Even if the experimenter gives priority to the policy maker's needs, she might be interested in modifications of the environment that cannot be implemented in the experimental design.

On the other hand, if the causal result fails to be replicated, this framework simply treats the failed occurrence as evidence against the hypothesis of external validity. A failed replication, however, might be clue to the fact that one "stumbled" into a salient variation of some contextual features. As such, the evidence of a failed occurrence is perfectly compatible with the treatment being efficacious but counteracted by other circumstances;

[95] For a discussion of the type of claims we can extrapolate from RCTs and the related assumptions see Cartwright (2007a).

[96] See my discussion of the notions of similarity and comparability in Chapter 3.

[97] How often the replication of an RCT is liable to deliver the same result is disputable, and indeed very much disputed (see Deaton at 2012 Annual Conference, NYU Development Research Institute). This point is, however, not directly relevant to my argument.

[98] The risk is thus that of searching for the key where the lamp sheds its light.

or with it being efficacious were other circumstances in place.[99] It is plausible that the policy maker would like to distinguish the case in which the treatment efficacy is an ephemeral and idiosyncratic phenomenon from the case where the phenomenon is robust *and* eclipsed in the circumstances. The replication of RCTs as such, however, is opaque to this sort of considerations. A different strategy is required if one wants to tell the two cases apart.

In the opponents' camp the suggestion was made that, to obtain causal effects with high external validity, one should turn to an altogether different strategy of identification and estimate. Cross-country regressions have been proposed as such an alternative (Rodrik 2009). According to this view cross-country regressions would outperform RCTs in terms of external validity because they identify and estimate causal effects with a broader geographical and temporal coverage.

The charge from this camp against RCTs is that they license conclusions that, even though highly reliable, are only valid for the experimental population and, thus, typically very narrow in scope. The solution proposed then amounts to broadening the scope of the causal claim by broadening the population on which it is tested. Cross-country regressions would do just that. They in fact measure average causal effects that "cover" a long list of countries (all countries included in the sample) over a long temporal frame (the temporal interval covered by the regression).

Unlike RCTs, cross-country regressions face notorious problems of internal validity, however.[100] It is thus a legitimate question whether they can indeed constitute a "real" solution to the problem of external validity that RCTs face. It has been argued that internal validity is a prerequisite to external validity, rather than standing in a trade-off relationship with it.[101] It seems reasonable in fact that results that are incorrect in the studied population cannot apply to populations not yet studied. It, thus, seems that if the study

[99] Another way to put it, adopting Cartwright's language, is that this evidence is compatible with the corresponding capacity claim being true (see section 4.5 for a brief discussion of the notion of capacity).

[100] Problem of internal validity in correlation analysis were introduced in section 1.2.1 and discussed more extensively in Chapter 2.

[101] See section 3.2.

fails to identify correctly the causal effects in the sample, it hardly can do so for out-of-sample populations. There are certainly cases, though, where one has reasons to trust the validity of a cross-country regression. Furthermore, if the regression measure causal effects correctly, it does so for a large pool of countries. What then about relevance?

This seems to be the case of a trade-off. In at least one respect cross-country regressions ensure broader coverage at the expense of relevance. Measuring the causal effects of a given policy in a pool of countries requires that the policy be treated as a homogeneous variable across the whole set. For this to be possible some process of abstraction is typically required. That is, the concrete properties of the program are stripped away for constructing a variable that abstracting away from the particulars of the case becomes broadly applicable. This operation gives rise to two types of problems.

First is the problem of construct validity, which has to do with whether the so-called homogeneity assumption is justified. Generally speaking, constructs and variables are valid if they succeed in measuring what they actually purport to be measuring. In cross-country regressions, this requirement is complicated by the fact that the variables ought to measure the *same* phenomenon across contexts. This means that one needs warrant for the operation of measuring particulars in different countries by using one and the same variable; in other words, one need to be justified for regarding different particulars as *instances* of the same phenomenon. In the context of interest, one would have to make sure that the programs implemented in the various countries are sufficiently alike to justify the conceptual operation of treating them as instantiations of one and the same policy variable. If this operation is unjustified the measurement is invalid.

The second problem has to do with the level of abstraction being *adequate* to the policy maker's interest. It is plausible that policy makers aim at fine-grained control on the outcome of interest; and fine-grained control depends on the nitty-gritty of the program at hand. As one climbs the ladder of abstraction in concept construction, one achieves greater generality at the expense of *details* concerning the phenomenon of interest.[102] In

[102] Giovanni Sartori calls the process of constructing more abstract concepts "climbing the ladder of abstraction", and distinguishes proper and improper ways of doing it. The proper way leads to the

the case of medical protocols the details might regard the treatment dosage, duration, the strategy of administration, and so on. In the case of nutritional policies like the ones discussed in Chapter 3, the program is a *package* of activities stretching from growth monitoring and nutrition surveillance, nutritional first aid, distribution of food, nutritional training of volunteers, food supplementation demonstration, and so on. Each activity can be defined in turn at a more concrete level: nutritional first aid can in fact comprise vitamin A distribution for children every 6 months, iron supplementation, oral rehydration, and so on.

Details such as the ones described above need often to be sacrificed in climbing the ladder of abstraction. A given program or policy variable can apply justifiably at the large scale typical of cross-country regressions only if very few details of its specific instantiation in each country are retained. If it is plausible to assume that nutritional programs are to a large extent standardized across countries, this is hardly true of more complex policy packages, such as institutional reforms concerning political, social and economic systems. However, these are the details in which the policy maker might be ultimately interested. Knowing that institutions have a positive impact on growth rate might not be very helpful for policy making, even if causal effects are correctly measured in a large pool of countries. The policy maker might in fact want to know the effect of tinkering with the "how" and "what" of specific institutional set-ups.

These proposals are both driven by the quest for generality. The strategies they suggest, whether the replication of RCTs or the use of cross-countries regressions, rest on the assumption that external validity is a synonym for generality, and causal effects are relevant if they are general enough. Relevance is not, however, a matter of scope. Policy makers worry only incidentally about the generality of the causal claims they use because they typically deal with specific contexts. Causal effects are of interest to them only if they respond to questions about the impact that given programs or policies would have were

formulation of general concepts that correctly apply to large number of cases because their empirical content (their intension or connotation) has been properly reduced in the process of abstraction. The improper way gives rise to mere generalities, whose application to a large number of cases is unwarranted. Generalities arise from concept-stretching. For a thorough and illuminating discussion see Sartori (1970).

they implemented in a specific way in a specific context; and insofar as this is the type of problem the policy maker faces. The higher the generality of the causal effect, the higher the chance that it will also travel to the specific context of use. From the point of view of the policy maker this is, however, hardly sufficient.

The mere replication of RCTs is blind to the modifications of interest to the policy makers, either because they are likely to fail to fall under the experimenter's control, or because they modify the effect in such a way as to render it opaque to detection. On their side, cross-country regressions privilege concerns about the external validity of the causal effects over their relevance. To allow identification, causal effects are defined for variables that have been stripped away of all concrete details of the particular program, details that arguably are of major interest to policy makers.

An alternative approach to the problem of external validity that plagues RCTs is hinted at by Angus Deaton (Deaton 2009, 2010). Deaton argues that the best strategy for justifying any hypothesis of external validity consists in establishing *why* the hypothesis holds in the experimental setting in the first place. In this spirit, he suggests that theory development and mechanistic theorizing are the type of evidence that eventually grant the inferential ticket one needs.

While Deaton's proposal remains suggestive, the same intuition is further developed and formalized by James Heckman. Heckman outlines a model for causal/counterfactual effects as a framework where policy makers address and solve the policy questions of interest. His contribution foreruns and encompasses the more recent debate on RCTs in development economics: it directly confronts the statistical model for causal effects on which RCTs are in fact based. In what follows, I shall consider how Heckman's proposal refines and outperforms the HR model for causal effects.

4.4 A NEW MODEL FOR CAUSAL EFFECTS: MARSHALLIAN CAUSAL FUNCTIONS

Heckman's proposal can be regarded as a *refinement,* rather than a *replacement,* of the RCT program (Heckman 2010). It questions the relevance of the Rubin-Holland model for causal effects, which informs the RCT template, and offers a more sophisticated version of it that promises to take adequate care of policy problems (Heckman 2001, 2008). Heckman maintains that policy makers typically face and try to solve the following problems:

> (P1) Evaluating the impact of historical interventions on outcomes including their impact in terms of the well-being of the treated and of society at large.
> (P2) Forecasting the impacts of interventions implemented in one environment in other environments, including their impacts in terms of well-being.
> (P3) Forecasting the impacts of interventions never historically experienced to various environments, including their impacts in terms of well-being (Heckman 2008: 7-9).

Heckman regards P1 as a problem of internal validity and P2 and P3 as problems of external validity (2008: 8). He maintains that RH can only answer questions of the first type and fails to address problems of the second and third type: RH can only define causal effects for the impact that policies already implemented had on outcomes. It does not allow the construction of counterfactual states in which the causal effect of old and new policies in new contexts can be evaluated. This shortcoming derives from RH's inadequacy for understanding and modeling the *causes of effects* (2001: 29).

Consider that RH posits two causes (t and c) potentially acting on each unit ω in population Ω.[103] Notice further that t, c and ω are not defined as variables but as parameters within the model. As such they are single-valued; that is, no variation is definable over them. The only variables in this model are Y, which is the outcome to be measured, and S, which indicates the relevant state of the world. The variation we observe in Y is therefore solely determined by the units being exposed either to cause t or c. No variation in Y can obtain as a consequence of variation *in* t, c or ω. This setting constrains dramatically the range of causal effects definable within the model, and thus measurable eventually outside of it.

[103] See section 4.2.1 above.

Heckman points out that, having been collapsed into a parameter, the intervention is essentially treated as a black box (2008: 6). As it stands, the model therefore delivers no information on the interventions performed and the difference between them except for the fact that they do, indeed, differ. The causally relevant aspects of each intervention are, in fact, neither characterized, nor are the respects in which the interventions actually differ. Nor is any information available on the characteristics of the population that undergoes the interventions. In sum, the model cannot evaluate the outcomes of new policies and of old policies on a new population simply because there is no way to define a new policy or a new population therein. If the corresponding causal effects are not definable, problems (P2) and (P3) cannot be addressed.[104]

Heckman concludes that RH is not helpful for solving the problems of major concern to policy makers. He proposes a deep structural model in which he defines what he calls Marshallian causal functions (2008: 29-30):

$$Y(s) = g(\boldsymbol{Q_s}, \boldsymbol{X}, \boldsymbol{U_s})$$

The function g maps the vectors of generating characteristics $\boldsymbol{Q_s}$, $\boldsymbol{X}$, $\boldsymbol{U_s}$ into outcomes. These vectors represent the whole range of causes that contribute to bring about outcome Y. Define $S = \{s_1 ... s_n\}$ as the set of treatments which are possibly implemented on unit ω. In this framework each treatment s_i[105] is defined as a bundle of characteristics which are the components of vector $\boldsymbol{Q_s}$. If s_i (i=1) is a medical protocol, then the components of $\boldsymbol{Q_s}$ might indicate the type of drug (Q_1), the months of treatment with the drug (Q_2), the quality of physicians (Q_3), and so on.

I mentioned above unit ω upon which the treatment is hypothetically implemented. Notice that ω does not appear in the causal function. It is instead modeled by means of the vector $\boldsymbol{X}$ which includes the determinants of Y that are observable. Units are thus

[104] The choice of not modeling the causal relationship can well be motivated by empirical worries. This fact, however, does not undermine Heckman's objection that in such a model questions of interest to policy makers such as P2 and P3 cannot be raised.

[105] i= {1...n}.

characterized in terms of the causal factors of *Y* and the treatment they are exposed to. In other words, ***X***'s components fully identify the unit: units are equivalent if ***X***'s components are. Individual causal effects can thus be evaluated. Finally, vector ***U*** collects the determinants of *Y* that are unobservable, both regarding the treatment and the unit as well. This framework allows ***U*** to include different components as the treatment implemented varies (***U*** is indexed to *s*).

In this model causal effects can be defined by varying either a single component of one vector, or the vector as a whole, while keeping the remaining variables fixed. For instance, we might define the causal effects of $\boldsymbol{Q_s}$ on *Y* as follows:

$$g(\boldsymbol{q_s}, \boldsymbol{x}, \boldsymbol{u_s}) - g(\boldsymbol{q_s}', \boldsymbol{x}, \boldsymbol{u_s})$$

This framework is an "all causes" model which supposedly makes explicit the whole range of factors relevant to the outcome *Y* (2008: 28). Consider that the range of causal effects we define within a model depends on the factors we can hypothetically vary. In RH, the only causal effect definable is the difference in outcome consequent to the application of treatment *t* and *c* on unit ω. As I remarked above, interventions are treated as parameters rather than variables. This explains why Heckman asserts that in RH they remain black boxes to the investigator. Marshallian causal functions model instead interventions as packages of measures, in which each single components (or single aspects of the treatment) can vary. Furthermore, the features of the environment relevant to the outcome are also made explicit.

In this way, it is possible to observe variation in the outcome as a result of changes in the context. By modeling the causes of effects, Marshallian causal functions multiply the range of counterfactuals open to exploration, and the causal effects that are definable therein. According to Heckman, this operation allows the range of questions that can be addressed to be much broader. In particular, this model would enable us to evaluate the impact that a given policy has on a new population, and thus address problem (P2). This is obtained by varying single components in $\boldsymbol{x}$. We might further evaluate the impact of newly envisaged policies and thus address problem (P3), by varying either $\boldsymbol{q}_s$, $\boldsymbol{q}_s$, and $\boldsymbol{x}$ or the three vectors all together (or single components in them).

4.5 HOW *RELEVANT* ARE MARSHALLIAN CAUSAL FUNCTIONS?

As I shall argue, Heckman's model for causal effects deals with the problem of external validity in such a way that it also renders the causal effects defined therein more relevant to the policy maker. It cannot be denied, however, that this gain comes at a huge cost. The model relies extensively on "some" theoretical background, which should allegedly provide the massive causal knowledge required.

One of the basic assumptions for the model to work, in fact, is that it characterizes the whole array of causes of the outcome in question and the corresponding functional form. Heckman is confident that "some theory" would take care of this demand (2001: 12-3). He remains vague, though, as to *which* theory can adequately meet such an epistemic requirement.

Furthermore, having such an ambitious range of causal effects correctly estimated is also very demanding in empirical terms as to the extent and accuracy of data collection, measurement, and analysis. These concerns only matter, however, for an overall assessment of the advantages and disadvantages of Heckman's approach with respect to the alternative models and strategies. Such a comprehensive assessment is not of interest here. Rather, the question at stake is whether Heckman's proposal addresses more adequately policy makers' questions than the strategies to which it reacts.

Heckman's strategy to improve the external validity of causal effects consists in modeling the causal structure for the outcome in question. It is widely acknowledged by philosophers of science that *some* knowledge of the causal structure is indeed required for formulating claims of external validity more reliably, even though there is disagreement as to the *type* of knowledge that one should use. For instance, some scholars maintain that knowledge of *capacities* licenses reliable inference (Cartwright 1989, 2009a). Capacities are causal powers endowed with *potentiality* and *stability*. That is, capacities produce a characteristic effect when impeding factors are not operating and they *contribute* to the overall outcome even when counteracted. Furthermore, their ability to do so persists across some range of contexts. In virtue of their stability across changes in the background conditions, it seems that knowledge of capacities can help external validity inference. Some scholars disagree, however.

Daniel Steel among others maintains that capacities are not helpful to solve the most pressing problem of external validity (2008: 83-4): it is not at all clear how we are to obtain knowledge of capacities where we are more likely to need it. Issues of external validity typically arise in those circumstances where we have very limited knowledge of the contexts to which we are interested in extrapolating causal claims. Knowledge of capacities would be helpful in these cases, wasn't it for the fact that it presupposes that we already have substantial knowledge of the target contexts; namely, that the power in question is stable across the relevant changes in the background conditions. In other words, using capacities to solve the problem of external validity seems to presuppose knowledge that we do not possess when issues of external validity arise (one is trapped in what Steel defines the *extrapolator's circle*). Several scholars propose the use of mechanistic knowledge as an alternative (Guala 2005, 2010, Steel 2008, 2010).

Disagreement aside, the intuition shared by these approaches is that understanding *why* a certain outcome obtains is a fairly reliable guide for predicting whether a similar outcome will also occur were circumstances to differ in some respects. Heckman's approach shares this intuition. His proposal is motivated by the insight that policy makers address counterfactual questions that, as such, can only be answered properly in well-developed causal models.[106] Counterfactuals are conditional statements that describe what would happen in hypothetical states of affairs. Heckman has in mind counterfactuals of interest to policy makers that describe the impact that given policies would have on outcomes of interest in given populations. In well-developed causal models counterfactual states (potential outcomes) and the related counterfactual effects (the difference between the outcomes) are precisely defined. Such models, in fact, render the whole causal structure

[106] In this intuition he is not alone (see Reiss and Cartwright 2004, Reiss 2011).

explicit and thus make possible to assess how the outcome would change were the relevant circumstances modified in some relevant respects.

Both the scholars who propose RCT replication and those who defend the use of cross-country regression as strategies to improve the external validity and relevance of causal effects fail to take this point fully on board. Causal modeling is there either implicit, as in the case of RCT replication, or too coarse, as in the case of cross-country regressions. These accounts are based on the assumption that relevance is a by-product of external validity, and thus set the generality of the causal claim as their primary task. As argued above, however, external validity *per se* is not guarantee that the relevant questions are addressed. And, indeed, the upshot of these two accounts is that the range of counterfactual effects they define is highly constrained.

RCTs are based on the Rubin-Holland model, which avoids explicit modeling and thus only defines counterfactual effects for potential outcomes in the alternative states of the cause it features. Any replication of RCTs that is not based on rendering the underlying model explicit cannot improve this situation just by extending the range of additional treatments *in practice*. Cross-country regressions are bound to treat the program as homogeneous across the countries and, to this end, they are likely to strip away significant details from the policy variable. Thus, provided that the regression equation counts indeed as a causal model (Deaton 2010), the potential outcomes defined therein do not involve tinkering with the program, the details of which remain black-boxed to the policy maker.

Heckman's approach, by way of modeling explicitly the causal structure, is superior to the strategies examined above. That is, it can be more easily used to address the questions of concern to policy makers. Its suppleness, in fact, permits to address both a *broader* range of counterfactual questions and more *fine-grained* counterfactual questions. Breadth has to do with the range of modifications that can be implemented within the model and, thus, with the variety of contexts it potentially encompasses. Breadth is maximized in Heckman's model as the range of modifications spans the whole causal structure which is here fully deployed. Thus, a much broader range of potential outcomes can also be defined.[107] Furthermore, fine-grainedness has to do with the possibility of exploring the effect of minute modifications in the causal structure. These are of interest to policy makers because they often aim at a surgical control of the outcome. Tinkering with the components of a treatment or policy, rather than the whole package, is a means to achieve it. Furthermore, their decisions often involve the nitty-gritty of program implementation. They ought to decide about the program details, such as its timing, duration, scale, and amount, and other qualitative aspects regarding the forms and contents of the implementation. It is therefore plausible that having a model that defines counterfactual effects for minute interventions is highly desirable for policy makers concerned with this type of decisions.

How relevant is, though, Heckman's model *in fact*? Reiss and Cartwright (Reiss and Cartwright 2004, Reiss 2008) argue not very much. According to them, the range of policy counterfactuals it assesses is still too limited for the purpose of policy making. Causal effects in Marshallian causal functions are, in fact, very "special" quantities: they measure the variation in outcome that follows variation in individual components of the causal structure *while the other components are held fixed*. The upshot is that causal effects can only be defined for those components that are variation-free; that is, can be modified independently from the others the relation with which should not be constrained by functional restrictions. Reiss and Cartwright aptly call counterfactual effects so-defined *Galilean effects* as they are "just the kind of effect we look for in a Galilean experiment" (Reiss 2008: 207). Their argument for lack of relevance stands on metaphysical grounds. It is simply not the case that systems in which control variables of interest can be modified, while letting the rest of the causal structure unaffected, are a common feature of our social world. Galilean systems, though epistemically convenient, are very special arrangements. In this sense, thus, the range of causal effects defined in Marshallian causal functions is still too constrained. Even though policy makers might be in principle interested in

[107] The range of causal effects definable within the model is however constrained by further assumptions which I shall discuss below.

producing Galilean effects, this might just not be an actual possibility in the systems in which they work.

The relevance of Heckman's approach can be further challenged. His proposal gets off the ground insofar as an additional, and fundamental, assumption is justified. That is, the counterfactual of interest to the policy maker is of the following type:

> (**PH**) What would be the impact on the outcome, were a policy implemented in a given way in a given environment?

There are contexts of intervention, typically developing contexts, in which the counterfactual of prior concern to the policy maker is of the following type instead:

> (**P***) How should a given environment be modified and/or a program implemented therein, were a *given* impact to be produced?

I understand **PH** as a problem of prediction, and **P*** as a problem of planning. In contexts where the causal structure is *defective*, in a sense I will describe below, counterfactual questions such as **P*** trump, that is override, **PH**. Heckman's model is designed to address problem **PH**, and is inadequate, as the approach to which it reacts also is, to confront problem **P***. In the contexts where **P*** trumps **PH**, the relevance of this model to policy makers is therefore very limited.

4.6 A SALIENT DISTINCTION: PREDICTING AND PLANNING

Contexts such as the ones developing economists worry about are typically characterized by *defective* causal structures. For a program to work, a given set of background conditions needs to be in place. A program of children immunization in hospitals is effective only if the hospital is within reach for villagers and mothers have a means to get there. Educational programs raise the literacy rate in the community if classrooms are available and big enough to allow students to attend the class. Causal structures are defective when the set of conditions that need to be present for a given policy to produce the intended effect are either inadequate or plainly absent. In these contexts, the local circumstances demand that the policy maker be concerned with *how* to adapt the policy so as to render it effective, rather than whether a given policy is effective as such. Issues of *adaptation* thus arise, and counterfactual questions such as **P*** trump **PH**.

This point is illustrated by the following case. Abhijit Banerjee tells the story of a project by the World Bank described in the Sourcebook on "Empowerment and Poverty Reduction". The Gyandoot program in Madhya Pradesh, India, consisted in providing computer kiosks in rural areas. Banerjee ironically remarks that the World Bank enlisted the program among the most successful in combating poverty, while at the same time acknowledging that the project was "hit hard by lack of electricity and poor connectivity" (Banerjee 2007). From this patent failure Banerjee concludes that the World Bank should have considered firm evidence about the likely impact of the proposed action before having the program implemented. In particular, since RCTs are the simplest and best way of assessing the impact of a program, "one would not want to spend a lot of money on an intervention without doing at least one successful randomized trial if one is possible" (ibid.)

If the World Bank had estimated the causal effect of the program in question before implementing it, as Banerjee indeed recommends, it would probably have predicted that the program was unlikely to produce the desired effect in the target context.

This piece of evidence, however, is not what the World Bank needs in order to confront the kind of problem at hand. In the first place, evidence about the likely causal effect is not necessary, at least at this stage of program implementation. Had the World Bank first gathered evidence on the required background conditions, and checked their presence or adequacy in the target context, the "hard evidence" of the RCT would have become redundant. The regular and reliable provision of electricity is obviously one such condition. Upon realizing that the electrical system in place is defective, and the energy supply irregular and unreliable, the need to measure the likely impact of installing computer kiosks in the villages would just vanish.

Moreover, knowledge of the background conditions, the ones *required* by the program and the *actual*, is not delivered by an RCT, or, for that matter, by any method for the measurement of causal effects, as they are not designed to provide it. This knowledge, however, is crucial in the case at hand. It testifies, in fact, that if the required conditions are not in place the program cannot just be implemented in the target context as it was in the context of origin. *Adaptation* is required. The policy maker has thus to make the program work by adjusting it to the local conditions.

Adaptation is an issue for the policy maker whenever she considers whether to carry a given program over to a new context whose causal structure is dissimilar from the original context. Thus, it is a more general concern than that which is being discussed here.[108] It is hardly the case in fact that two contexts are alike in such a way as to allow for the same policy to be implemented in exactly the same way in both cases without the need at least of some minor adjustments. Adaptation, however, is an even more challenging issue when the causal structure in the target context is defective, as it is often the case in the developing contexts of interest here.

Development programs typically depend for their successful implementation on the presence of adequate infrastructures, the production and distribution of resources such as water and energy, production and distribution of food, and so on. More ambitious programs consist in the implementation of institutional changes. Successful privatization, for instance, depends according to many on a broad range of background circumstances that span from the adequate protection of property rights, to the rule of law, an independent judiciary, a democratic political system, and so on. It is hardly disputable that similar conditions are often insufficiently developed, or patently absent, in many developing contexts. Adaptation in these cases might consist in rendering the conditions adequate for the program to work; that is, setting the conditions "right". However, it need not be.

If the causal structure is *defective* in the sense described above, the policy maker actually faces a whole range of alternatives. She might in fact decide:

1. To modify the background conditions so as to render the program effective;

[108] See Pawson and Tilley (2004: 33).

2. To modify the program so as to adapt it to the actual background conditions;

3. Some strategy between options 1 and 2.

This situation is characterized by the fact that multiple courses of action are open to the policy maker. She has thus to devise the alternative strategies that enable her to fulfill the desired goal, and assess them comparatively before coming to the decision of which one to implement. The solution to this problem is far from obvious. One concern, among several, that arises at this stage is that interfering with the extant causal structure in order to set the conditions right might disrupt, or simply affect, local processes in unintended ways. In similar situations the policy maker confronts what is more accurately construed as a problem of *planning* rather than mere *prediction.*

Planning consists in formulating strategies once the objectives have been defined and information about the context taken into account. Prediction instead translates information about the environment and the chosen strategy into statement about future results (Armstrong 1983). In a planning process what program to implement, how to implement it, which other modifications of the context to make are open to question, and genuinely devised along the process upon analysis of the actual situation. Whereas prediction is concerned with the outcome of given policies, planning consists in designing the intervention *in such a way* that the intended outcome would eventually follow.

How do we approach planning then? Let's first consider to what extent a model such as Heckman's would help. Heckman defends the relevance of his model on the grounds that it aims at estimating the most ambitious counterfactuals policy makers face. These counterfactuals describe what would be the impact on outcomes of *new* policies, policies that have never been tried before, in *new* populations, populations that, it goes without saying, never underwent this policy before.

The model is thus designed to answers questions such as **PH** above. Can it be used also to address problems like **P***? It seems as if, by enabling to estimate the impact of *novel* programs, it could also *design* novel interventions, and not only assess their effects. In other words, the model seems equipped for planning policies. This is, however, not the case. Consider how the model works.

$$Y(s) = g(\boldsymbol{Q_s}, \boldsymbol{X}, \boldsymbol{U_s})$$

It displays the causal structure of the outcome by describing its determinants and the corresponding functional relations. These determinants comprise the policy specified as a vector of components ($\boldsymbol{Q}_s$) and the contextual factors relevant to the outcome ($\boldsymbol{X}$). A "new" policy is characterized here by means of variations of the components of $\boldsymbol{Q}_s$. Similarly, a "new" population is defined by modifications of the components of vector $\boldsymbol{X}$. The "new" policy and population look just like the "old" ones, except for the fact that the *values* of the relevant variables, or their components, are set at *new levels*, levels that have not been observed before.

In this model the policy maker tinkers (theoretically) with the program and context, and observes the consequences of her tinkering on the outcome of interest. She manipulates a target context the causal structure of which is presupposed, whereas the impact of her ideal manipulation is actually inferred. This model, thus, is not adequate to address *genuine* planning. Planning is characterized by the policy being the epistemic output of the decisional process, and not its input.

More specifically, planning consists in the inference *from* the actual causal structure and the intended outcome *to* the alternative strategies that help achieve the desired goal. In planning, knowledge of what variables one is going to manipulate is thus delivered and not assumed. The inferential process *from* actual causal structure and intended outcome *to* alternative strategies is based on a much broader set of considerations than the standard approaches to policy making considered in this paper are willing to concede. These considerations include:

I. The background conditions necessary and sufficient for the alternative programs to work;

II. The presence or absence of these sets of conditions in the target context;

III. The causal structure in the target context relevant to the outcome of interest.

Concerns about III are particularly prominent. The local context, with its causal field and causal laws, is the place where to start to plan the policy. It is also, therefore, the place where to start with the modeling process, the output of which should be the program to implement. The answer to whether the best strategy to undertake is option 1, 2, or 3 above, therefore, crucially depends upon the circumstances in place. Any consideration about the effectiveness of a given program needs thus to *follow* the analysis of the local circumstances and the conditions that ought to be present for the program to work.

The upshot is that one should indeed worry about the causal effects of a given policy; which policy this is, however, becomes only apparent upon careful study of the target context. In what follows, I shall spend some words on the planning process in general and on the type of analysis of the target context that is relevant to it and why. In particular, I shall argue that case studies play a prominent role in this analysis and that, in so doing, they provide relevant evidence for the very early stages of the planning process.

4.7 WHAT ELSE DO WE NEED FOR PLANNING?

That planning rests on a much broader set of considerations than those enabled by the approaches examined in this chapter is easily proved by a sketchy, and admittedly approximate, description of the planning process. One can distinguish several phases within the process, each demanding distinct analysis and evidence (Armstrong 1983)[109]. The first consists in the *specification of the objective*. Setting the desired goal presupposes a diagnosis of the problem at hand, which requires background theory and empirical analysis

[109] Armstrong actually distinguishes four steps in planning rather than three as I describe below (the fourth stage is monitoring results). My purpose here, though, is illustrating the complexity and distinctiveness of the process rather than being exhaustive in its characterisation.

to define and explain the relevant problem. The analysis of the problem affects how the objective is formulated.

The second phase consists in the *generation of alternative strategies* to achieve the desired goal. The analysis of the causal structure in the target context is central here. It ideally includes the alternative causal paths through which the outcome is affected (each corresponding to a potential strategy of intervention), and the background conditions required for each strategy to be effective (the INUS conditions).[110]

The third phase consists in the *selection of the strategy* to be implemented. This decision depends on different strands of causal evidence and considerations of various character: whether the potential strategies can be *actually* exploited to fulfill the policy objective; the estimate of the causal effect each strategy is likely to produce; the comparison of the alternative strategies in other relevant respects (ethical, financial, practical, environmental, social), and so on.

In what follows I will briefly discuss by way of an example how case studies can help in the first and second phase by providing relevant evidence for the specification of the policy objective and the formulation of alternative strategies of intervention. I will further suggest that pieces of causal evidence other than case studies are relevant in the third phase of strategy selection.

In 1981, Amartya Sen published "Poverty and Famines", an essay on the causes of starvation. He presents the entitlement approach as an alternative to the Food Availability Decline (FAD) theory which was dominant then. Whereas the FAD theory explains starvation as caused by a decline in the availability of food per head, Sen understands it as resulting from a person's failure to be entitled to a bundle with enough food, which is determined in turn by legal, social, economic, and political factors of the relevant society and by the person's position in it. Sen's work can be understood as a defense of the empirical superiority of the entitlement approach over the FAD theory in terms of predictive accuracy and explanatory power. The FAD theory, in fact, leads to the

[110] See sections 1.3.3.1 and 3.5.1.

formulation of expectations about the occurrence of starvation that turn out to be incorrect in cases where the entitlement approach would deliver correct predictions instead. Moreover, even in the cases in which FAD's predictions happened to be accurate the theory fails to account for the actual pattern of starvation: "*who* died, *where*, and *why*?" (1981: 120 – *italic in the original*)

At the core of Sen's approach lies the notion of *maximum food entitlement*:[111]

For occupation group *j*, define maximum food entitlement as:

$$(F_j) = q_j \, p_j / p_f$$

where q_j is the amount of commodity each member of group *j* can sell or consume, p_j is the price of commodity *j*, and p_f is the price of food per unit (1981: 50). Entitlement failures are "direct" when one produces less food for own consumption; when one obtains less food through trade by exchanging one's commodity for food one experiences a "trade entitlement failure" (1981: 51). Trade entitlement failure, in turn, can be caused by a decrease in q_j due to an autonomous production decline (e.g. when livestock is destroyed by a draught) or to a fall in the demand for the good in question. Alternatively, it can be due to a decline in occupation *j*'s food exchange rate (p_j/p_f). Sen explains that groups can suffer both direct and trade entitlement failure when they produce a commodity that is both directly consumed and exchanged for some other food. In this perspective, people failure's to command enough food as represented by a fall in F_j is the cause of starvation.[112]

Sen's theoretical introduction is followed by the analysis of four cases of starvation: the Great Bengal famine (1943), the Ethiopian famine (1972-1974), the famine in Sahel (1973),[113] and that in Bangladesh (1974). The four case studies play a prominent role in his

[111] The model specified below is a very simplified version of a more complex model that Sen discusses in the appendix to his book.

[112] Sen specifies that there are cases of starvation for which his model does not account like cases where falls in the entitlement set are determined by illegal transfers such as brigandage and looting.

[113] Sen distinguish three alternative definitions of "Sahel". In what follows I adopt the political definition according to which Sahel refers to six countries in the Sahelian area, namely Mauritania, Senegal, Mali, Upper Volta (now Burkina Faso), Niger, and Chad.

argumentative strategy. One purpose of introducing actual case studies of starvation is, in fact, proving the empirical superiority of the entitlement approach over the FAD theory. Upon the analysis of available data, for instance, Sen remarks that neither Bengal nor Ethiopia experienced a dramatic decline in food availability per capita in terms of either food production or calories consumption. The FAD theory would thus have failed as a predictor in these cases. It would have succeeded instead in the case of Sahel where a dramatic decrease in food availability was indeed experienced. However, the theory could not account for the specific pattern of starvation observed in this area which caused a sharp increase in the mortality rate of specific segments of the population.

The case studies do not only provide evidence for the higher empirical adequacy of the entitlement approach over the FAD theory, however. Through the lenses of the entitlement approach, they identify local patterns in the actual causation of starvation that vary from case to case. The analysis of how entitlement relations shifted in the years of the famines brings to light specificities of each case that in the absence of a case-based analysis would have gone unnoticed. In Bengal, the starvation was driven by a sharp increase in the price of rice typical of an economy of war which was further fuelled by the following speculation and panic hoarding.[114] In this context the segments of society that suffered the most where people in those occupations that, being not subsidized by the government, faced a dramatic deterioration in their terms of trade and a drastic fall in the demand for the goods they actually produced (e.g. agricultural laborers, fishermen, craftsmen, and other productive occupations). The massive starvation that occurred in Ethiopia in 1972-1974 was triggered by the severe drought in the Sahel belt in the preceding years. The sharpest decline in food entitlement was here experienced by pastoral nomads and agriculturists from the Eastern regions who underwent dramatic losses of their own crops and livestock and tried to find relief by migrating to the capital Addis Ababa.

[114] In 1942 Japanese troops occupied Burma (now Myanmar) which was then part of the British Empire. Being part of the British Empire itself, Bengal was directly involved in the conflict.

Shedding light on the specificities of actual causal patterns in each case is certainly interesting on its own. More importantly, it is a preliminary to actual planning.

STAGE I: IDENTIFYING THE POLICY OBJECTIVE

The first stage in planning consists in the specification of the objective: it depends on how the problem is analyzed with the help of background theory and empirical analysis, performed in Sen's work by means of case studies. Sen frames the problem of starvation within the entitlement approach. As discussed above this framework proves empirically successful in the cases at hand; however, it is too *abstract* to help specifying adequately a policy objective. The entitlement approach to starvation as such would in fact suggest as a goal that of preventing a failure in the food entitlement of the relevant population. An adequate specification of this policy objective would require that one clarifies what the relevant segments of the population are and what induced the failure in their food entitlement. Case studies can help at this stage by *re-situating* the problem in the context at hand: in this way, they render possible the identification of a more *concrete* policy objective.[115] The case study of the famine in Sahel can illustrate this fact.

In Sahel starvation is typically engendered by occasional draughts and hits specific segments of the population, namely pastoral nomads and agriculturists that inhabit the northern regions within these countries. In the years of severe draught (1972-74), pastoral nomads and agriculturists in the dry region of Sahel had their crops and livestock either destructed or severely damaged. Both these segments of the population suffered direct and trade failure in their food entitlement due to a sharp drop in the amount of commodity produced and its price. In particular the exchange rates between animals and grain collapsed despite the severe loss in livestock.[116] The sharp fall in their command over food

[115] See section 1.3.3.1 for a discussion of why *concrete* causal knowledge might be relevant for policy making.

[116] This unusual combination of decline in the quantity and price of the same good, namely livestock, is explained by Sen in several ways. First, consumption in times of crisis tend to move away from animals that are "superior goods" towards grains which are cheaper source of nutrition; furthermore, animals are often saved as stock and in times of difficulty the owner might need to "dissave". Finally grains constitute

explains why precisely the nomadic pastoral population and agriculturists from this area in each country had the highest mortality rate, and were thus much more affected than other segments of the population by the draught.

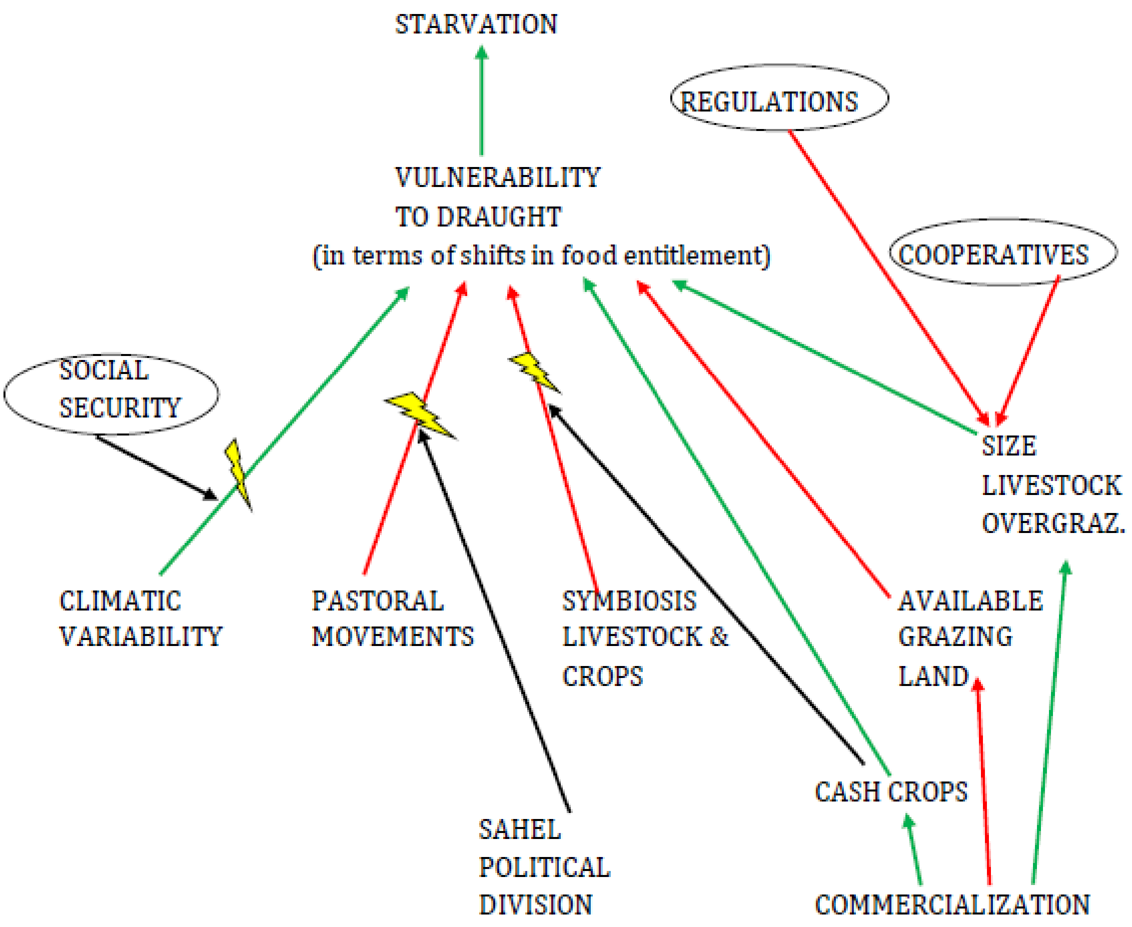
The case study, thus, identifies the context in which the Sahelian famine exploded, the factor that triggered it (the occasional draught), and the segments of population that experienced the most dramatic shift in food entitlement. On the basis of this evidence the policy objective is identified in *reducing herders' and agriculturists' vulnerability to draught*.

STAGE II: FORMULATING STRATEGIES OF INTERVENTION

Once the objective is specified, alternative strategies of intervention are formulated on the basis of the local causal patterns of starvation retrieved by means of the case-based analysis of the target context. This strategy consists in back-tracking from the outcome of interest (the policy objective) to causes through the analysis of the processes connecting the latter to the former (a sort of process-tracing from outcomes to causes). The following diagram represents the causal structure Sen reconstructs in the case of Sahel (1981: 126-129):

a more adjustable (and thus desirable) stock than livestock and its products when the time pattern of consumption requires flexibility.

Fig. 4.1 Starvation in Sahel



In figure 4.1 red arrows stand for negative causal influence, green arrows stand for positive causal influence, and black arrows pointing into other arrows represent factors that disrupt existing mechanisms (the disruption is represented by the yellow flash).

This type of analysis consists in the identification of multiple causal paths leading to the same outcome. The obvious benefit is that multiple courses of action are actually identified. Each path is a causal process that could be potentially exploited to modify the outcome. One can regard each process, in fact, as corresponding to an alternative strategy of intervention, and the causal variables that belong to it as potential levers that could be in principle manipulated to affect the outcome. But there is more to it. This type of analysis leads also to the identification of the processes that *used to be operative and had been disrupted* at a certain point in time.

There is evidence that traditional insurance systems were in place against draught which consisted in the movements of herds across the Sahel region (i.e. across the actual borders). The system was disrupted by the political division of Sahel which limited the freedom of movements of pastoral nomads. Similarly, commercialization disrupted the symbiotic relationship between livestock and traditional crops that used to grant a stable production of food. Reactivating these processes might, or might not be, a viable strategy. However, it seems that the option should not be excluded a priori, but only after an accurate empirical analysis of the processes that used to be in place, and of the factors that led to their disruption.

There are moral reasons that might justify a similar course of action, but they do not concern us here. There is also at least one practical reason of why this epistemic strategy might be recommendable. Having the background conditions in place is one of the factors that should matter in the criteria one uses when choosing among strategies of intervention. If the cause to manipulate belongs to processes that are *local* it might be the case that the required background conditions are already in place. If these processes were disrupted at some point in time, it is prima facie plausible that the background conditions might be easier to set right because they are part of local mechanisms that used to be at work in the target context. Implementing similar strategies would thus require less adaptation to local conditions than strategies conceived and imported from outside.

Case studies thus help formulate alternative strategies by providing processual and contextual evidence.[117] In particular, if they outline the complete causal mechanisms they can fulfill the two epistemic requirements of the second phase of planning. First, mechanisms are the alternative causal paths through which the outcome can be affected: each of them thus constitutes a potential strategy of intervention. Furthermore, the description of complete mechanisms individuates the background conditions that together produce the outcome. If the mechanisms are actually operating in the context the background conditions are necessarily in place and their identification is not strictly

[117] I discuss the relevance of processual and contextual knowledge for policy making in section 1.3.3.1.

speaking required for strategy implementation.[118] If the mechanisms have been disrupted, and thorough investigation of the background conditions involved cannot be performed in the target context, one should turn to cases where similar mechanisms are actually in place. If these cases are epistemically accessible, they can be used to extrapolate knowledge of the missing background factors upon comparison with the target of interest.[119]

STAGE III: POLICY SELECTION

The analysis of the causal processes leading to the outcome of interest, the actual and the disrupted, is not sufficient on its own for the selection among the alternative strategies of intervention. Letting aside considerations that have to do with the side-effects of policies or with concerns of other nature (financial and ethical considerations for instance), for purposes of strategy selection knowledge of causal processes needs to be supplemented by other pieces of causal evidence. Two certainly seem prominent here. First, one needs evidence that the alternative strategies are invariant under (some range of) intervention. Invariance under intervention is also called stability or autonomy in the econometric discourse. The underlying idea is that the causal relationship one intends to exploit to reach the policy objective does not break when one intervenes on it by manipulating some of the factors in the relevant causal process. Second, one needs evidence of causal effects. The quantitative estimate of the impact that each strategy would have on the outcome in question is in fact relevant to choose among the possible strategies.

These two pieces of evidence, invariance under intervention and causal effects, are both underwritten by counterfactual claims. Single case studies, however, cannot support counterfactuals: to evaluate counterfactual claims one need to turn instead to causal models. Two methodological proposals are promising in this respect. Qualitative claims that certain causal relations are invariant under a range of interventions can be assessed within models such as those proposed by Reiss (Reiss 2011). The causal model Reiss describes is devised to evaluate token claims. In the light of the previous discussion it is

[118] It might be, however, useful when exploring the side effects of the policy at hand.

[119] See Chapter 3 for a discussion of how making case studies comparable.

reasonable to think of policy claims as token causal claims: strategies of interventions that are specified at a high level of concreteness because they are formulated on the basis of local causal processes and fit outcomes that also are.[120] The strategies that prove to be invariant under intervention can then be compared in terms of the causal effect each of them is likely to produce by using models such as Heckman's.[121]

Heckman's model for causal effects should thus be regarded as one epistemic input, among others, for planning. It is one of the premises in a complex inferential process that delivers as its epistemic output the strategy to implement in order to fulfill the policy objective. In contexts where planning comes to the forefront, the inference from intended outcome and actual causal structure to alternative strategies is *prior* to the inferences licensed by models for causal effects like Heckman's. Only once the range of alternative strategies is specified, in fact, policy makers can engage with the actual selection of the strategy to implement. Causal effects become a relevant piece of evidence, one among many, at this stage of the decisional process. The selection from a range of alternatives, in fact, should also be based on the consideration of possible side-effects as well as financial, environmental, and social concerns. Furthermore, ethical considerations play a crucial role at this stage: what concerns, and whose concerns, give priority to in selecting the strategy to implement is matter that requires the policy maker to take a normative stance.

Approaches that give central stage to causal effects in the analysis of policy making forgo a crucial dimension which amounts to formulating alternative strategies of intervention. This dimension of policy making should also be subject to methodological scrutiny. It seems particularly important in contexts where policy making cannot, and should not, be reduced to carrying over strategies implemented elsewhere. Contexts where the necessary factors for the success of policies are lacking or defective make room for deep reformulation of policies and adaptation to local conditions. The epistemic starting

[120] For a discussion of the notion of fitness of causal claims and the requirement that causes fit outcomes (and the other way around) see Woodward (2010).

[121] On intuitive ground, I would further suggest that Heckman's model measure causal effects for causal relationships whose invariance under intervention is presupposed rather than inferred within the model.

point for this inquiry then needs to be the inference from local causal conditions to possible alternative strategies.

4.8 CONCLUSION

In this chapter I defended the idea that problems of external validity and relevance are, indeed, tightly connected. Relevance, however, should not be regarded as a spill-over effect of external validity. Rather, it should be regarded as a distinct, and prior, concern. I suggest in fact that the concern for relevance should drive the search for evidence, and thus dictate whether and when concerns about external validity should also be raised. Giving relevance priority over external validity means raising first the question of what kind of problems the policy maker addresses, what evidence she might need to solve them, and how this evidence can be provided.

Due to defective causal structures and need for adaptation, intervening in developing contexts is better understood as a problem of planning. Planning differs from prediction in that the policy to implement is treated as the epistemic output, rather than the input, of a complex inferential procedure in which the knowledge of the target context is upfront. I submit that case studies are relevant at the early stages of planning by helping the specification of the policy objectives and the formulation of alternative strategies of intervention. Causal effects, such as those Heckman's model can measure, are a valuable input at the later stage of strategy selection. Even at this stage, however, they are but one relevant piece of evidence among others.

Factual matters are by no means the only important concern when planning policies in developing contexts. Ethical concerns are at least equally prominent and, as such, they deserve a separate and thorough discussion.[122]

[122] For a discussion of the complex relation between values, science, and policy making see Douglas (2005).

5. GENERAL CONCLUSIONS

Studying the uses and the forms of case-based reasoning is a broad and intriguing endeavor for the philosophical speculation, and the narrow and partial discussion in the previous chapters can hardly do justice to it. Case-based reasoning in its multiple forms permeates our scientific, moral, and practical thinking; from this pervasiveness and its apparent simplicity much of its fascination as well as its challenges derive. The lack of formal sophistication and clear-cut results probably contributed to make it for a long time a forgotten method for the philosopher of science, although it never was for the social scientist. Even in the times in which case-based reasoning was subject to the harshest critiques and fell into discredit, social scientists could hardly ignore the simple fact that case studies continued to be a persistent feature of their practices. Thus, the case study never truly deserted the social scientist's toolbox and her methodological discussions.

This thesis is an attempt to bring case studies back to the fore of philosophical discussion. It does so in a moment of raising awareness among philosophers of science of the oblivion in which case studies fell for long time and of their importance. I addressed case-based reasoning from a very specific, and arguably narrow, angle. First, I confined my analysis to methodological issues, eschewing other fruitful perspectives such as the historical, the sociological, the epistemological, and the ethical, to name just a few. Furthermore, I restricted the disciplinary fields of investigation to *some* social sciences, to the exclusion of other fields where case-study reasoning arguably play a prominent role, for instance medicine, biology, psychology, psychiatry, and business studies. Moreover, I looked at case-based reasoning as an epistemic strategy for inquiring into social phenomena. In particular, I conceived of case study research as a machinery for establishing hypotheses, and of case studies as the output of its operation in which raw evidence, its analysis, and the ensuing conclusions intermingle.

To some extent my work is a reaction to two challenges to which case study research has been repeatedly subjected. On the one hand, case studies have been typically regarded as providing limited justification to the scientific hypotheses. This idea can be rephrased in a more modern, and quite fashionable, vocabulary by saying that case studies are too soft a piece of evidence for scientific inquiry. Case study research has also been often dismissed for its lack of generalizing power. Understandably, this charge becomes particularly serious when the generation of generalizations is regarded as the main aim of science. Only to some extent, however, was my work directed at addressing these challenges: it had also a more constructive goal, which is exploring the potential and specificity of case studies. In the methodological framework I adopted this amounted to understand the ways in which case study research finds, validates, and generalize knowledge, what kind of knowledge this is, and to clarify thus what case studies can offer that the other methods cannot provide.

This line of inquiry was also backed up by the conviction that the social scientific practice can be conducted in such a way as to be of use to the policy-maker, and that case studies possess features that render them interesting from this point of view. This ultimate preoccupation was, however, also influenced by the belief that the relationship between scientific practice and policy making is very complex. On the one hand, as other scholars previously acknowledged, scientific results are just one input into policy decision making, where a much broader set of considerations is at stake. On the other hand, there is no obvious and straightforward way for scientific epistemic practices to be of use to the policy makers. The spillover of scientific activity onto policies is various and, to some extent, unforeseeable. For what concerns my methodological investigation, I regarded the issues of internal validity, external validity, and relevance as crucial for a scientific practice that counts being of use to the policy makers among its objective.

Historical narratives are the output of case study research when employed for causal investigation. The primary objective of my inquiry into historical narratives has been to define the conditions under which the results of a causal inquiry so conducted can be regarded as valid. I focused in particular on process-tracing as the most widespread and less understood technique of causal analysis in case studies. Whereas previous proposals have been made to assess the contribution of process-tracing as such, these conditions prove to be very narrow in terms of applicability. In particular they fail to apply to historical narratives that are not deductively derived. On an intuitive ground I argue that this is most often the case, and devise conditions for the circumstances in which historical narratives are the outcome of a markedly inductive inference. It remained to be ascertained whether the validity conditions that I propose are applicable across the board that is, can also operate in those contexts where process-tracing relies substantially on the use of theoretical, experimental, or statistical evidence.

A reasonable demand to make on causal hypotheses thus obtained is that they be valid not only for the cases in which they have been established, but also in other cases yet unexplored, or studied only to a limited extent. This demand is reasonable because it is hard to deny that the scientific practice is *to some extent* a generalizing endeavor, and because scientific conclusions that fail to travel to other contexts can be hardly of use when less proximate purposes, like for instance policy making, come to the fore. In the philosophical debate external validity is much discussed and yet riddled with conceptual confusions. This is even truer of case study research. First of all, it is not indisputable that external validity, a concept coined in the experimental context, is the most suitable way to treat the problem of generalizability from case studies. Furthermore, the way the concept is used by a case study researcher is very much imbued of statistical talk and influences. This had the effect of leading the discussion on the external validity of case studies to what I regard as a dead end. Thus I confined my attention to the latter issue and to what I would describe as conceptual cleansing. The former question, however, is so far unanswered and calls for attention.

When one considers the aims of scientific practice, and evaluates the methods and the results they deliver also for their capacity to fulfill those aims, validity is a necessary but insufficient concept. Relevance is an equally important concern. It measures the adequacy of the scientific results for the purposes we value and the problems we try to solve. It thus requires thorough discussion of which aims are a legitimate goal of scientific inquiry, and which ones are not, and draws to some extent on factual considerations to justify one's point of view. Relevance, then, is what is also at stake in the growing debate on the use of randomized controlled experiments in development economics, whose aim is noble and hardly objectionable, that is erasing poverty from that part of the world that still suffers from it. I rejoin this debate because in the first place it offers the occasion to draw a neater distinction between external validity and relevance, which were conflated so far. While the two are connected, it is not only fruitful but also necessary to regard them as distinct concerns. Furthermore, it allows me to emphasize that the relevance of a given piece of evidence, namely causal effects in the debate in question, strongly depends on how one conceptualizes the type of problem policy makers actually face.

I characterized the problems policy makers face in developing contexts as problems of planning, and outlined this notion within an admittedly sketchy framework. What emerges from that discussion, however streamlined and incomplete, is that multiple source of evidence are required when one confronts purposes as embedded in their concrete context of reference. I pointed out that case studies are relevant evidence for planning because they help to specify the policy objectives and to identify alternative strategies of intervention; I also suggested, however, that this is but one piece of evidence among the many the policy maker needs. Future methodological investigations might help clarify further what are the distinct pieces of evidence that serve those purpose, what methods can provide them, and how they can be integrated. Future avenues of research might then also be directed to understand more precisely what the role of case studies is among the various sources of evidence on which we can draw; and to elucidate further their specificity and the respects in which they differ from other evidence. Finally, one might investigate what sorts of complementarities can arise by the conjoined use of different methods.

What I regard as the most challenging and seductive of the possible pursuits in this line of research is the attempt to gain deeper philosophical insight into the nuances and reaches of the case-based forms of reasoning. That is having a firmer grasp on what it means to reason in cases, and to reason with cases. When is something turned into a case, and what does it mean to look at something as a case? How does one reason from case to case, and when is this reasoning sound and fruitful? What does one carry with herself along the travel? This I regard as the most intriguing, and probably the most elusive, of all quests along this research path. Most probably, these are questions to which multiple answers can be given, and be all valid. If I had to pursue such an endeavor, however, I would do it equipped not only with the analytical lenses of the philosopher of science but with those of the historian and the sociologist as well.

REFERENCES

Acemoglu, D., Johnson, S., & Robinson, J. A. 2003. "An African success story: Botswana". In D. Rodrik, eds. *In Search of Prosperity. Analytic Narratives on Economic Growth*. Princeton University Press: Princeton.

Acemoglu, D., & J.A. Robinson. 2000. "Political losers as a barrier to economic development." *The American Economic Review* 90: 126-130.

Acemoglu, D., and J.A. Robinson. 1999. "The political economy of institutions and development." *Background Paper for World Development Report 2001*.

Alexandrova, A. 2009. "When analytic narratives explain."*Journal of the Philosophy of History* 3:1-24.

Al Rubaie, T. 2002. "The rehabilitation of the case study method". *European Journal of Psychotherapy, Counselling and Health* 5:31-47.

Armstrong, J.S. 1983. "Strategic planning and forecasting fundamentals". In Kenneth Albert, eds. *The Strategic Management Handbook*. McGraw Hill: New York

Banerjee, A.V. 2007. *Making Aid Work*. MIT Press:Cambridge

Banerjee, A.V. 2005. "'New Development Economics' and the challenge to theory". *Economic and Political Weekly* 40: 4340-4344.

Banerjee, A.V. and E. Duflo. 2009. "The experimental approach to development economics". *Annual Review of Economics* 1:151-178.

Bates & al. 1998. *Analytic narratives*. Princeton University Press: Princeton.

Becker, H. and C. Ragin. 1992. *What is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press: Cambridge.

Bennett, A. and C. Elman. 2006. “Qualitative research: recent developments in the case study methods”. *Annual Review of Political Science* 9:455-476.

Bennett A. and A.L. George. 1997. “Process Tracing in Case Study Research”, *MacArthur Foundation Workshop on Case Study Method.*

Brady, H. and D. Collier, eds. 2004. *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Rowman and Littlefield Publishers: Lanham.

Bunge, M. 1997. “Mechanisms and explanation”. *Philosophy of the Social Sciences* 27:410-465.

Bunge, M. 2004. “How does it work?: The search for explanatory mechanisms”. *Philosophy of the Social Sciences* 34:182-210.

Campbell, D. T. 1975. “Degrees of freedom and the case study.” *Comparative Political Studies* 8:178-193.

Campbell, D. T. and J.C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Rand McNally: Chicago.

Cartwright, N. 2011. “The Art of Medicine”. *The Lancet* 377:1400-1401.

Cartwright, N. 2009b. “Evidence-based policy: what's to be done about relevance?” *Philosophical Studies* 143: 127-136.

Cartwright, N. 2009a. “What is this thing called efficacy”. In C. Mantzavinos, eds. *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*. Cambridge University Press: Cambridge.

Cartwright, N. 2007b. *Hunting Causes and Using Them.* Cambridge: Cambridge University Press.

Cartwright, N. 2007a. “Are RCTs the Gold Standard?” *BioSocieties* 2:11-20.

Cartwrigth, N. 1989. *Nature Capacities and Their Measurement.* Oxford University Press: Oxford.

Cartwright, N. and J. Hardie. 2012. *Evidence-based Policy. A Practical Guide to Doing it Better.* Oxford University Press: Oxford.

Cohen, J., and W. Easterly, eds. 2009. *What Works in Development? Thinking Big and Thinking Small.* The Brookings Institutions: Washington.

Coleman, J.S. 1986. “Social Theory, social research, and a theory of action”. *American Journal of Sociology* 91:1309-1335.

Cook, T. D., and D. T. Campbell, eds. 1979. *Quasiexperimentation: Design and Analysis Issues for Field Settings.* Rand McNally: Chicago.

Creager A.N.H., Lunbeck E., and M.N. Wise, eds. 2007. *Science Without Laws. Model Systems, Cases, and Exemplary Narratives.* Duke University Press: Durham.

Deaton, A. 2010. “Instruments, randomization, and learning about development”. *Journal of Economic Literature* 48:424-455.

Deaton, A. 2009. “Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development”. The Keynes Lecture, British Academy.

DeWalt, K. and B. DeWalt. 2002. *Participant Observation: A Guide for Fieldworkers.* AltaMira Press: Walnut Creek.

Donohue J.J. and S.D. Levitt. 2001. “The Impact of Legalized Abortion on Crime.” *Quarterly Journal of Economics* CXVI: 379-420.

Douglas, H. 2005. "Inserting the public into science." In Maassen, S. and P.Weingart, eds. *Democratization of Expertise?* Springer: Dordrecht, 153-169.

Eck, J. 2006. "When is a bologna sandwich better than sex? A defense of small-n case study evaluations." *Journal of Experimental Criminology* 2:345-362.

Eckstein, H. 2000 [1975]. "Case study and theory in political science." In R. Gomm, M. Hammersley & P. Foster, eds. *Case Study Method: Key Issues, Key Texts*, 119-164.

Elster, J. 1998. "A Plea for Mechanisms", in Hedström*,* P., and R. Swedberg, eds. *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press: Cambridge.

Elster, J. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge University Press: Cambridge.

Eysenck, H.J. 1976. Introduction. In H.J. Eysenck, eds. *Case Studies in Behaviour Therapy*. Routledge and Kegan Paul: London.

Flyvbjerg, B. 2006. "Five misunderstanding about case study research". *Qualitative Inquiry* 12:219-245.

Forrester, J.1996. "If *p*, then what? Thinking in cases". *History of the Human Sciences* 9:1-25.

Geertz, C. 1973. "Thick description: toward an interpretive theory of culture." In *The Interpretation of Cultures: Selected Essays*. Basic Books: New York.

George, A.L. 1997. "Knowledge for statecraft: the challenge for political science and history." *International Security* 22:44-52.

George, A.L. 1994. "The two cultures of academia and policy making: bridging the gap." *Political Psychology* 15:143-172.

George, A.L. 1976. "Bridging the gap between theory and practice." In J. Rosenau, eds. *In Search of Global Patterns*, 114-119.

George, A. L., and A. Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. The MIT Press: Cambridge.

Gerring J. 2007b. "The mechanismic worldview: thinking inside the box". *British Journal of Political Science* 38:161-179.

Gerring, J. 2007a. *Case Study Research. Principles and Practices*. Cambridge University Press: Cambridge

Gerring, J. 2004. "What is a case study and what is it good for?" *American Political Science Review* 98: 341-354.

Grϋnbaum, A. 1988. "The role of the case study method in the foundations of psychoanalysis". *Canadian Journal of Philosophy* 18:623-658.

Guala, F. 2010. "Extrapolation, analogy, and comparative process tracing." *Philosophy of Science* 77:1070-1082.

Guala, F. 2005. *The Methodology of Experimental Economics*. Cambridge University Press: Cambridge.

Guala, F. 2002. "Experimental localism and external validity." *Philosophy of Science* 70:1195-1205.

Hacking, I. 1992. "Style' for historians and philosophers". *Studies in History and Philosophy of Science* 23:1-20.

Heckman, J.J. 2010. "Building bridges between structural and program evaluation approaches to evaluating policy". NBER Working Paper Series.

Heckman, J.J. 2008. "Econometric causality". Discussion Paper n. 3525. Iza, Bonn.

Heckman, J.J. 2001. “Econometric Counterfactuals and Causal Models”. Unpublished Keynote Address, International Statistical Institute, Seoul.

Hedström*,* P. and R. Swedberg. 1998. *Social Mechanisms. An Analytical Approach to Social Theory*, Cambridge University Press: Cambridge.

Hedström P. and P. Ylikoski. 2010. “Causal Mechanisms in the Social Sciences”. *The Annual Review of Sociology* 36:49-67.

Holland, P. 1986. “Statistics and causal inference”. *Journal of the American Statistical Association* 81:945-960.

Hoyningen-Huene, P. 2006. “Context of discovery and context of justification and Thomas Kuhn”. In J. Schickore and F. Steinle, eds. *Revisiting Discovery and Justification*, 119-131.

Jimenez-Buedo, M., and L.M. Miller. 2010. “Why a trade-off? The relationship between the external and internal validity of experiments.” *THEORIA* 25:301-321.

Jones, T. 1999. “FIC Descriptions and Interpretive Social Sciences: Should Philosophers Roll their Eyes?” *Journal for the Theory of Social Behaviour* 29:337-369.

Jonsen, A.R. and S. Toulmin. 1988. *The Abuse of Casuistry*. University of California Press: California.

Kennedy, E. T. 1991. *Successful nutrition programs in Africa: What makes them work?* World Bank Publications.

Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences*. Cambridge University Press: Cambridge.

King, G., Keohane, R.O., and S. Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press: Princeton.

LeCompte, M. D., and J.P. Goetz. 1982. “Problems of reliability and validity in ethnographic research.” *Review of Educational Research* 52:31-60.

Lieberson, S. 1991. “Small N's and big conclusions: an examination of the reasoning in comparative studies based on a small number of cases.” *Social Forces* 70: 307-320.

Lijphart, A. 1971. “Comparative politics and the comparative method”. *The American Political Science Review* 65:682-693.

Lindblom, C.E. and D.K. Cohen. 1979. *Usable Knowledge: Social Science and Social Problem Solving*. Yale University Press: New Haven.

Little, D. 1998. *Microfoundations, Method and Causation: On the Philosophy of the Social Sciences*. Transaction Publisher: New Brunswick.

Little, D. 1995. “Causal explanation in the social sciences.” *The Southern Journal of Philosophy* 34: 31-56.

Little, D. 1991. *Varieties of Social Explanation. An Introduction to the Philosophy of Social Science*. Westview Press: Oxford.

Lucas, J. W. 2003. “Theory-testing, generalization, and the problem of external validity.” *Sociological Theory* 21: 236-253.

Mackie, J.L. 1988. *The Cement of Universe: A Study in Causation*. Clarendon Press: Oxford.

Mahoney, J. 1999. “Nominal, ordinal, and narrative appraisal in macrocausal analysis.” *American Journal of Sociology* 104:1154-1196.

Mahoney, J and G. Goertz. 2006b. “Scope in case study research.” Unpublished working paper.

Malinowski, B. 1935. *Coral Gardens and Their Magic.* American Book Co: New York.

Mayntz, R. 2004. “Mechanisms in the Analysis of Social Macro-Phenomena”. *Philosophy of the Social Sciences* 34:237-259.

McNair, M. 1954. *The Case Method at the Harvard Business School.* McGraw-Hill Book Company, Inc: New York.

Mill, J.S. 1858. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and Methods of Scientific Investigation*. Harper: New York.

Morgan, M. 2012. “Case studies: one observation or many? Justification or discovery?” *Philosophy of Science* 79:667-677.

Morgan, S.L. and C. Winship. 2007. *Counterfactuals and Causal Inference. Methods and Principles for Social Research.* Cambridge University Press: Cambridge.

Morse, J.M. et al. 2002. “Verification strategies for establishing reliability and validity in qualitative research”. *International Journal of Qualitative Methods* 2:13-22.

Pawson, R. 2006. *Evidence Based Policy. A Realist Perspective.* Sage: London.

Pawson and Tilley. 2004. “Realist Evaluation.” Unpublished paper.

Pawson, R. and N. Tilley. 1997. *Realistic Evaluation.* Sage: London.

Ragin, C. 2000. *Fuzzy-set Social Science.* University of Chicago Press: Chicago.

Reiss, J. 2013. *Philosophy of Economics.* Routledge: New York.

Reiss, J. 2011. “Counterfactuals” in Kincaid, H. (ed.), *The Oxford Handbook of Philosophy of Social Science.* Oxford University Press: Oxford.

Reiss J. 2008. *Error in Economics. Towards a More Evidence-Based Methodology*. Routledge: London.

Reiss J. and N. Cartwright. 2004. "Uncertainty in Econometrics: Evaluating Policy Counterfactuals" in Mooslechner, P., H. Schubert and M. Schütz, eds. *Economic Policy Making under Uncertainty: The Role of Truth and Accountability in Policy Advice*. Edward Elgar: Cheltenham, 204-32.

Rodrik, D. 2009. "The new development economics: we should experiment, but how shall we learn?" in Cohen, J. and W. Easterly, eds. *What Works in Development? Thinking Big and Thinking Small.* The Brookings Institutions: Washington.

Rodrik, D. 2003. *In Search of Prosperity: Analytical Narratives on Economic Growth.* Princeton University Press: Princeton.

Runyan, W.M. 1982. "In defence of the case study method". *American Journal of Orthopsychiatry* 52:440-446.

Sartori, G. 1970. "Concept misformation in comparative politics". *The American Political Science Review* 64: 1033-1053.

Saxenian, A. L. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press: Cambridge.

Scriven, M. 2008. "A summative evaluation of RCT methodology and an alternative approach to causal research". *Journal of MultiDisciplinary Evaluation* 5:11-24.

Scriven, M. 2007. "The logic of causal investigations". Unpublished paper. Western Michigan University.

Seawright J. and J. Gerring. 2008. Case selection techniques in case study research. *Political Research Quarterly* 61: 294-308.

Sen, A. 1981. *Poverty and Famine. An Essay on Entitlement and Deprivation*. ILO.

Shadish, W.R., Cook, T.D and D.T Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company: Berkeley.

Skocpol, T. 1979. *States and Social Revolutions. A Comparative Analysis of France, Russia, and China*. Cambridge University Press: Cambridge.

Steel, D. 2010. "A new approach to argument by analogy: Extrapolation and chain graphs." *Philosophy of Science* 77:1058-1069.

Steel D. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science.* Oxford University Press: Oxford.

Steel, D. 2004. "Social Mechanisms and Causal Inference". *Philosophy of the Social Sciences* 34:55-78.

Stinchcombe, A.L. 1991. "The conditions of fruitfulness of theorizing about mechanisms in the social sciences". *Philosophy of the Social Sciences* 21:367-388.

Vennesson, P. 2008. "Case studies and process tracing: theories and practices". In della Porta, D. and M. Keating, eds. *Approaches and Methodologies in the Social Sciences. A Pluralist Perspective.* Cambridge University Press: Cambridge.

Waldner, D. 2003. "Inferences and explanations at the K/T boundary...and beyond". Unpublished paper. University of Viriginia.

Walker, H. 1940. "Degrees of freedom". *Journal of Educational Psychology* 31:253-269.

Woodward, J. 2010. "Causation in biology: stability, specificity, and the choice of levels of explanation". *Biology and Philosophy* 25:287-318.

Yin, R.K. 2003. *Case Study Research. Design and Methods*. Sage Publications: Thousand Oaks.

SAMENVATTING

Ondanks modetrends in de academische wereld heeft case-based reasoning bewezen een blijvende vorm van analyse te zijn in de sociale wetenschappen, de geesteswetenschappen en het morele denken. Algemeen gesproken houdt case-based reasoning in dat de uiteindelijke bron van onze kentheoretische en morele intuïties gevonden kan worden in de tastbaarheid van specifieke gevallen. Hoewel ze een gemeenschappelijke oorsprong hebben, bestaan er verschillende tradities in het academische landschap die gebruik maken van case studies en case-based reasoning. Deze dissertatie richt zich primair op het gebruik van case studies in de sociale wetenschappen die als epistemisch doel hebben om causale hypotheses te formuleren, vast te stellen of te generaliseren. Een tweede doel is het onderzoeken van het nut en belang van causale bevindingen verkregen in en door middel van case studies voor beleidsvorming in de sociale praktijk.

De dissertatie bestaat uit 4 hoofdstukken. In hoofdstuk 1 beschrijf ik wat kan worden gezien als twee verschillende zienswijzen op case studies en de achterliggende wetenschapsfilosofie. De eerste benadering floreerde in de jaren zeventig en zag case studies als een bijzondere - en vaak minder overtuigende – vorm van experimenteel, statistisch of vergelijkend onderzoek. Aangezien deze benadering geneigd is case studies te evalueren met standaarden uit andere methodologische tradities, kan het gezien worden als een heteronoom paradigma. De tweede, alternatieve zienswijze die in het afgelopen decennium werd ontwikkeld begint langzaam aan een vorm aan te nemen en is bij lange na niet volledig gearticuleerd. Deze benadering beoogt case studies los te koppelen van het idee dat ze slechts gebruikt moeten worden als andere methodes niet beschikbaar zijn. Case studies worden hierbij gezien als een *autonoom epistemisch genre* (Morgan 2012).

In hoofdstuk 2 bespreek ik de interne validiteit van case studies in de vorm van historische narratieven. Dergelijke case studies beogen causale hypotheses te formuleren en te onderbouwen door reeksen gebeurtenissen te beschrijven die tot een bepaald resultaat leiden. Doorgaans maken ze gebruik van het nalopen van processen (*process-tracing*) om tot causale gevolgtrekkingen te komen en wenden zich daarnaast tot vergelijkend onderzoek. Ondanks het belang van historische narratieven in de sociale wetenschappen is de rol van process-tracing in deze narratieven nog slecht begrepen. Feitelijk is het debat over process-tracing dat de afgelopen jaren flink is toegenomen in academische publicaties nog onduidelijk en onderontwikkeld. Meer specifiek zijn er nog geen gedeelde criteria om de epistemische waarde van academische bijdrage van process-tracing te beoordelen. Bovendien binden de voorwaarden die worden geopperd vaak de geldigheid van de bevindingen aan het gebruik van specifieke vormen van bewijs, en bieden dus geen nuttige inzichten als deze specifieke bewijzen er niet zijn.

Ik betoog dat de voorgestelde voorwaarden onnodig beperkend zijn en de werkelijke waarde die process-tracing te bieden heeft als het om causale gevolgtrekking gaat niet erkennen. Ik formuleer nieuwe voorwaarden om de prestaties van process-tracing te beoordelen wanneer de omstandigheden niet gunstig zijn en bestaande criteria niet van toepassing zijn.

In hoofdstuk 3 bespreek ik het generaliseerbaarheidsprobleem. Ik geef een overzicht van wat ik zie als de traditionele zienswijze op externe validiteit. Deze benadering is geworteld in een statistische visie op case study-onderzoek en ziet externe validiteit slechts als een probleem voor representativiteit. Hiermee wordt het debat over generaliseerbaarheid afgekapt, omdat externe validiteit simpelweg als een nadeel van case study-onderzoek wordt gezien. Tegelijkertijd wordt interne validiteit beschouwd als een relatieve kracht van case study-onderzoek. Op grond hiervan wordt in deze benadering case study-onderzoek geadviseerd als interne validiteit het belangrijkste onderzoeksdoel is, en worden andere methodes aangeraden als men in plaats daarvan generalisaties nastreeft. Dit is een ongelukkige zienswijze omdat case studies feitelijk vaak worden gedaan met het impliciete of expliciete doel om lessen te trekken met betrekking tot nog onbekende gevallen en contexten.

geadviseerd als interne validiteit het belangrijkste onderzoeksdoel is, en worden andere methodes aangeraden als men in plaats daarvan generalisaties nastreeft. Dit is een ongelukkige zienswijze omdat case studies feitelijk vaak worden gedaan met het impliciete of expliciete doel om lessen te trekken met betrekking tot nog onbekende gevallen en contexten.

Ik poog deze spanning op te lossen door de aannames te onderzoeken achter de traditionele zienswijze over externe validiteit van case study-onderzoek. Sommige van deze aannames zijn al besproken én betwist in het huidige debat. In hoofdstuk 3 bespreek ik die aannames die –naar mijn beste weten nog niet geëvalueerd zijn en verantwoordelijk lijken te zijn voor de impasse waarin sociale wetenschappers in kwestie zich bevinden. In het bijzonder beargumenteer ik dat het debat zich moet richten op de vraag hoe we case studies kunnen vergelijken in plaats van wat een prototypische case study is. Deze twee aspecten zijn weliswaar gerelateerd, maar moeten onderscheiden worden. De traditionele zienswijze doet dit niet en schept daardoor verwarring over wat externe validiteit is en hoe men dit probleem het beste het hoofd kan bieden. Ik verwacht dat het verbeteren van vergelijkbaarheid leidt tot nieuwe ruimte voor het verbeteren van de criteria voor externe validiteit van resultaten uit case study-onderzoek.

In hoofdstuk 4 bespreek ik kwesties met betrekking tot relevantie in de context van beleidsgerelateerd onderzoek. In het bijzonder richt ik mij op het debat rondom het gebruik en het nut van gerandomiseerd onderzoek met een controlegroep (*Randomized Controlled Trials;* RCT’s) om inzicht te krijgen sociaal-economische ontwikkeling. Deelnemers in het debat zijn het er over eens dat RCT’s te kampen hebben met problemen van externe validiteit en dat dit hun nut voor beleidstoepassingen beperkt. Ze verschillen echter van mening over de beste strategieën om dit probleem het hoofd te bieden. Ik analyseer drie alternatieven uit de economische literatuur: replicatie van RCT’s, voorgesteld door de voorstanders van RCT’s; regressie analyse op het niveau van landen, een benadering aangehangen door RCT-sceptici; en causale modellen, geopperd door James Heckman. Ik betoog dat deze strategieën - in beperkte, en verschillende, mate - succesvol zijn in het aanpakken van externe validiteit.

Voorstanders van de eerste twee strategieën erkennen onvoldoende dat er een belangrijk verschil is tussen relevantie en externe validiteit, en benaderen de eerste als een spillover van de tweede. Hun zienswijze is gericht op het verbeteren van de externe validiteit van causale verbanden met als aanname dat hiermee relevantie vanzelf volgt. Ik stel dat dit niet terecht is omdat relevantie en externe validiteit onderscheiden concepten zijn die apart van elkaar moeten worden beschouwd. Het voorstel van Heckman slaagt erin causale verbanden te ontwaren die, feitelijk, relevanter zijn voor de doelen van beleidsmakers. Ik betoog echter dat zijn model geen bevredigend antwoord vindt op de problemen die beleidsmakers tegenkomen in de ontwikkelingscontext. Heckmans model kan weliswaar omgaan met problemen die te maken hebben met *voorspellingen*, maar in de ontwikkelingscontext hebben beleidsmakers vooral te maken met *planning* en problemen die daarbij komen kijken. Planning is een complexe procedure waarbij een verscheidenheid aan relevante data gebruikt moet worden en diverse problemen komen kijken. Causale verbanden vormen slechts één epistemische input in deze procedure, terwijl case study-onderzoek tevens relevant is in andere cruciale fases.

SUMMARY

Despite fads and fashions in the academic culture, case-based reasoning has proved to be a persistent form of analysis in the social sciences, in the humanities, and even in moral thinking. Broadly understood, case-based reasoning locates the ultimate source of our epistemic and moral intuitions in the concreteness and idiosyncrasy of particulars. Even though they can be traced back to a common root, different traditions of reasoning with cases and of using case studies coexist in the academic landscape. This thesis focuses primarily on the use of case studies in the social sciences as an epistemic strategy to formulate, establish, and generalize causal hypotheses. A secondary goal is an investigation into the use of causal findings generated within and by means of case studies to inform policy making in the social realm.

The thesis is organized in four chapters. In chapter 1 I characterize what can be regarded as two alternative views of case studies and the understanding of science in which they are embedded. The first approach flourished in the 70s and looked at case studies as a special, and typically weaker, form of the experimental, statistical, or comparative methods. Since this approach tends to evaluate case studies by criteria belonging to other methodological traditions it can be said to present a *heteronomous* paradigm. The second, alternative view, which developed in the last decades, is taking shape gradually and is still far from being fully articulated. This approach strives for an understanding of case studies liberated from the narrow mindset that caricatures case studies as the method of last resort. In particular, it sees case studies as an *autonomous epistemic genre* (Morgan 2012).

In chapter 2 I address internal validity in historical narratives. Historical narratives are case studies that aim to formulate and substantiate causal hypotheses by articulating descriptions of the sequences of events leading to the outcome of interest. They typically make use of process-tracing to draw causal inference, and often rely on the additional use of the methods of comparison. Despite the important role of historical narratives in the social sciences, how process-tracing operates in the narratives is still poorly understood. The debate on process-tracing in fact, even though it is growing thanks to a number of recent contributions, is still muddy and under-developed. In particular, there are no shared criteria to assess its epistemic contribution; moreover, the conditions proposed so far tend to tie the validity of the findings to the use of specific kinds of evidence and are thus unhelpful when this specific evidence is not available.

I argue that the proposed conditions are unduly restrictive and fail to acknowledge the actual contributions process-tracing can offer to valid causal inference. I formulate new conditions to assess process-tracing performance in cases in which the favorable evidential circumstances do not occur and existing criteria fail to apply.

In chapter 3 I address the problem of generalizability. I provide an outline of what I define as the traditional view on external validity. This approach is conditioned by a statistical viewpoint on CSR and reduces external validity to issues of mere representativeness. In so doing it leads the debate on the generalizability of case-study results to a dead end as it quickly dismisses external validity as the downside of CSR. At the same time, it suggests that CSR is comparatively stronger in providing results internally valid. On this ground this approach recommends the use of case studies when internal validity is the main research goal of interest, while turning to other methods when one pursues generalizations instead. This outcome is unfortunate because, as a matter of fact, case studies are often performed with the explicit or implicit purpose of drawing lessons from the studied case to be carried over to new contexts yet unstudied.

I attempt to solve this tension by examining the assumptions behind the traditional view on the external validity of CSR. Some of these assumptions have already been addressed, and actually disputed, in the current debate. In chapter 3 I focus instead on those assumptions that, to the best of my knowledge, have not been addressed yet and seem to be responsible for the dead end in which the discussion among social scientists seems to be trapped now. In particular, I suggest that the debate should focus on how make case studies comparable rather than how select the typical case. Typicality and comparability are concepts closely related but distinct. The traditional view conflates the two and thus run into confusion about what external validity is really about and how it can actually be confronted in a fruitful manner. I surmise that by enhancing the comparability of studies unnoticed room for improvement is made for formulating more reliable assessment of the external validity of results obtained in case studies.

In chapter 4 I discuss issues of relevance when policy making purposes are at stake. In particular, I focus on the debate on the use and usefulness of randomized controlled trials (RCTs) to find the key to economic and social development. The participants to this debate agree that RCTs are affected by limited external validity, and that this impinges on their usefulness for policy making. They diverge, however, on the strategies to overcome this problem. I analyze three alternatives that are found in the economic literature: replication of RCTs, which has been proposed by the promoters of RCTs; cross-country regressions, which have been typically endorsed by RCT-skeptics; and the causal models proposed by James Heckman. I argue that these strategies succeed in their attempt to a different, and limited, extent.

Proponents of the first two strategies fail to take into adequate consideration the distinction between external validity and relevance, and treat the latter as a spill-over of the former. Their strategies, in fact, aim to improve the external validity of causal effects on the assumption that relevance will automatically follow. I argue that this is not the case because external validity and relevance are distinct concerns and should thus be confronted separately. The proposal by Heckman succeeds in delivering causal effects that are, as a matter of fact, more relevant to policy makers' purposes. I argue, however, that his model cannot adequately address the type of problems policy makers are likely to confront in developing contexts. Whereas Heckman's model is equipped to face problems of *prediction*, in developing contexts policy makers face problems of *planning*. Planning is a complex procedure that depends on various pieces of evidence and raises several concerns. Causal effects are but one epistemic input in this procedure; case-study evidence is also relevant to the crucial phases of planning.

ATTILIA RUZZENE

• E-MAIL: ruzzene@fwb.eur.nl

CURRENT POSITIONS

LECTURER

Faculty of Philosophy (EUR)

DOCTORAL STUDENT

Erasmus Institute for Philosophy and Economics (EUR)

EDUCATION

2010 **PhD in Economics of Complexity and Creativity**, University of Torino
Dissertation: Causality in Economics. A methodological Inquiry
Thesis advisor: Prof. Roberto Marchionatti
Committee: Prof. Franco Donzelli, Prof. Stefano Fiori, Dr. Julian Reiss

2009 **Research Master in Philosophy of Economics**, Erasmus Institute for Philosophy and Economics Erasmus University Rotterdam, The Netherlands
Dissertation: Causal Knowledge in Case-study Research. A Contextual Analysis
With distinction

2006 **MS in Economics**, Coripe Piemonte, University of Torino, Italy

2003 **BA&MA in International and Diplomatic Sciences** (Branch: Economics of Developing Countries), University of Torino, Italy
Summa cum laude - Awarded Premio Optime to the best dissertations 2003/2004

PUBLICATIONS

Policy making in developing countries: from prediction to planning. ***Journal of Economic Methodology***, forthcoming.

Review of K.I. Wolpin *The Limits of Inference without Theory*. MIT University Press. ***Journal of Economic Methodology***, forthcoming.

Process-tracing as an Effective Epistemic Complement. ***Topoi***, 33 (2): 61-72, 2014.

Drawing Lessons from Case Studies by Enhancing Comparability. ***Philosophy of the Social Sciences***, 42 (1): 99-120, 2012.

Review of Y. Chantal-Gagnon *The Case Study as Research Method: a Practical Handbook*. Presses de l'Université du Quebec. ***International Studies in Philosophy of Science***, 25 (3): 293-296, 2011.

Meccanismi Causali nelle Scienze Sociali, **APhEx**– Portale Italiano di Filosofia Analitica, 5, 2012

MISCELANNEOUS

"Reasoning with Cases in the Social Sciences, 11-12 November 2011", The Reasoner 5(12): 214-5

"Mechanisms and Causality in the Sciences, 9-11 September", The Reasoner 3(11): 10.

RESEARCH TALKS

2013

- "On the External Validity of Causal Effects and their Relevance for Policy Making"

  ENPSS – Philosophy of Social Science Roundtable – Venice, September 2013

- "On the External Validity of Causal Effects and their Relevance for Policy Making"

  International Network of Economic Method (INEM) conference – Rotterdam, June 2013

- "Making knowledge usable by means of case studies"

  Philosophy of Science in a Forest (PSF), Leusden, May 2013

2012

- "Making knowledge usable by means of case studies: are institutions really for growth?"

  Symposium *Causes and Comparability in Cases: the Human and Social Sciences*

  The Philosophy of Science Association Conference, San Diego, November 2012

- "On the epistemic autonomy of process-tracing in historical narratives"

  Workshop Social Mechanisms and Social Explanation, EIPE (EUR), Rotterdam, May 2012

2011

- "Process-tracing without a theory: why it works and how"

  Rotterdam-Tilburg Graduate Workshop in Philosophy of Science (TiLPS and EIPE) - Tilburg, November 2011.

- "Drawing lessons from case studies by enhancing comparability"

  International Network of Economic Method (INEM) conference - Helsinki, September 2011.

- "Ethnographic research in policy-making. Does it take us from here to there?"

  POP Faculty of Philosophy (EUR) - Rotterdam, June 2011.

- "Pluralism in extrapolation: making room for ethnographic research"

  PhD seminar, Erasmus Institute for Philosophy and Economics (EUR) - Rotterdam, May 2011.

- "Case-study research: low in external validity, high with extrapolation"

  13th Philosophy of Social Science Roundtable - Paris, March 2011.

2010

- "On the (limited) relevance of causal effects for evaluating interventions"

  Causality in the Biomedical and Social Sciences, EUR - Rotterdam, October 2010.

- "Causal inference in the social sciences. Why mechanisms matter and how to find out about them"

  Work in Progress in Causal and Probabilistic Reasoning, Kent University – Paris, June 2010.

- "Causal inference in the social science. Why causal mechanisms matter and how to find out about them"

  Understanding and the Aims of Science, Lorentz Center - Leiden, May 2010.

2009

- "The relevance of mechanistic evidence in case-study research",

  Mechanisms and Causality in the Sciences, University of Kent (Poster session) - Canterbury, September 2009.

2008

- "Case-based methodology. What evidence for economics?"

  Convegno Nazionale Storep, Università LUISS Guido Carli – Rome, June 2008.

TEACHING EXPERIENCE

- Erasmus School of Economics (ESE), Erasmus University Rotterdam, 2014
  Philosophy of Economics, 3rd year undergraduate course (with Dr. Constanze Binder and Dr. Conrad Heilmann)
- Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam, 2014:
  Topics in the Philosophy of Science, ERIM Research Master's course (with Dr. Conrad Heilmann)
- Faculty of Philosophy, Erasmus University Rotterdam, 2013-14
  Ethical Aspects of Economics, interdisciplinary 3rd year undergraduate course (Full responsibility),
- Erasmus School of Economics (ESE), Erasmus University Rotterdam, 2013
  Tutorials Philosophy of Economics (Dr. Conrad Heilmann and Dr. Constanze Binder)
- Erasmus School of Economics (ESE), Erasmus University Rotterdam, 2011
  Tutorials Philosophy of Economics (Dr. Julian Reiss)

SERVICE TO THE PROFESSION

Referee for: Topoi, Social Studies of Science, APhEx, Erasmus Journal for Philosophy and Economics (EJPE).

Co-organiser of the Formal Ethics 2014 conference at Erasmus University Rotterdam, Rotterdam 30-31 May 2014, jointly with Constanze Binder and Conrad Heilmann.

Co-organizer of the Rotterdam Graduate Conference in Philosophy of Science, Rotterdam 8-9 March 2012, jointly with Sine Bagatur and Francois Claveau.