Texts in Quantitative Political Analysis
Series Editor: Justin Esarey

Alessia Damonte
Fedra Negri *Editors*

Causality in Policy Studies

A Pluralist Toolbox

OPEN ACCESS

Springer

Texts in Quantitative Political Analysis

Series Editor

Justin Esarey, Dept of Politics, FM Kirby Hall 319
Wake Forest University
Winston Salem, NC, USA

This series covers the novel application of quantitative and mathematical methods to substantive problems in political science as well as the further extension, development, and adaptation of these methods to make them more useful for applied political science researchers. Books in this series make original contributions to political methodology and substantive political science, while serving as educational resources for independent practitioners and analysts working in the field.

This series fills the needs of faculty, students, and independent practitioners as they develop and apply new quantitative research techniques or teach them to others. Books in this series are designed to be practical and easy-to-follow. Ideally, an independent reader should be able to replicate the authors' analysis and follow any in-text examples without outside help. Some of the books will focus largely on instructing readers how to use software such as R or Stata. For textbooks, example data and (if appropriate) software code will be supplied by the authors for readers.

This series welcomes proposals for monographs, edited volumes, textbooks, and professional titles.

Alessia Damonte • Fedra Negri
Editors

Causality in Policy Studies

A Pluralist Toolbox

Springer

Editors
Alessia Damonte
Social and Political Sciences
University of Milan
MILANO, Milano, Italy

Fedra Negri
University of Milan
Milano, Italy
University of Milan- Bicocca
Milano, Italy

Fondazione
Compagnia
di San Paolo

ISSN 2730-9614 ISSN 2730-9622 (electronic)
Texts in Quantitative Political Analysis
ISBN 978-3-031-12981-0 ISBN 978-3-031-12982-7 (eBook)
https://doi.org/10.1007/978-3-031-12982-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

How can we think of causation in policy research? Almost any research tradition provides a different answer. For instance, emphasis can be placed either on the process leading to a policy outcome or on its underlying conditions. A process can be either observable or unobservable, and the underlying relevant conditions can be understood as single factors or complex configurations. Either samples, populations, or single cases can be invoked as the proper empirical ground for grasping them. Evidence can be arranged to either claim relevance or irrelevance. These differences reflect as many distinct assumptions about the shape of causation and build as many research strategies.

Causality in Policy Studies equips researchers to meet two related challenges in the field. First, algorithms for data analysis embed selected assumptions about causation that often remain unspoken. Knowing these assumptions is crucial to understanding how algorithms can be appropriately employed and eventually combined to compensate for their blind spots and weaknesses. Second, policy research is carried out within various disciplines (such as political science, sociology, economics, management, and administration), each often married to particular traditions. The book addresses the technical drive of such differentiation. In doing so, it provides the opportunity for researchers of any stripe to familiarize themselves with the strategies on which other streams build their claims.

In short, the book shows how to learn from different causal techniques, apply them consciously, and possibly make them speak to each other to get a better sense of findings. For this purpose, it structures the journey into causal knowledge in three stages. First, it introduces the foundational issues of causation (Chaps. 1 and 2). Then, it exposes the inner working of selected techniques for causal analysis (Chaps. 3, 4, 5, 6, 7, 8 and 9). Last, it considers some incompatibilities and complementarities among techniques to improve causal knowledge (Chaps. 10 and 11).

The red thread connecting all chapters is a reasonable realist stance. All share the tenets that causation is factual and entails generative and transfer processes unfolding at different levels of reality. Moreover, the chapters agree that causation can be known. Hypothetical statements about its manifestations, direction, and conditions can be given a testable shape. They also agree that causal statements should be

believed when logically and empirically compelling. The book's commitment to methodological pluralism follows from these tenets. The complexity of causal phenomena is such that no single technique can grasp its entirety. Still, each technique can illuminate particular facets in response to a precise research question. Indeed, asking whether a factor can yield one outcome differs from asking how it happens or under which conditions it obtains, and each response calls for adequate analytic tools. When pieced together, these responses can offer a better account of the phenomena of interest.

Methodological pluralism can deliver on the promise of better knowledge if the strengths and weaknesses of each technique are understood and tackled. To this end, each substantive chapter clarifies the research question a technique can answer, the research design and data treatment the technique requires for credible results, and the domain of validity of its findings. Wherever possible, a replicable example illustrates the deployment of the analysis as the sequence of operations and actual decisions. Of course, this selection of techniques is far from exhaustive of the methodological variety of policy studies. Nevertheless, this suite provides sharp insight into the different strategies to establish the tenability of a causal statement. As such, it can offer guidance beyond the boundaries of this book.

The edited format of the book aims at providing highly usable and solid knowledge for policy assessment and evaluation to MA students, PhD students, scholars, and practitioners in policy-related fields. Thus, each chapter is authored by a recognized scholar from different backgrounds, generations, and perspectives. Such a diverse yet "close-knit" team is essential to the volume. A single author could hardly have covered such a range of techniques with comparable expertise.

Public policies are tools and governance systems to tackle collective problems. Good policies call for a generation of open-minded scholars and practitioners willing to understand and learn from research conducted in different fields and capable of handling the techniques in their toolbox consciously and carefully. We hope you will have a good time going through the chapters. Enjoy your journey!

MILANO, Milano, Italy

Alessia Damonte
Fedra Negri

Acknowledgments

Every book is a collective enterprise and some more than others. Our first thanks go to the many students who have compelled this project and shaped it during our courses. They have been the real drivers of this effort, and we are more than grateful for how they kept our motivation high over the many months of writing and revising. Heartfelt thanks also go to Licia Papavero, Francesco Zucchini, and the board of the Ph.D. in Political Studies of the University of Milan. Their constant support to the Summer School in "Research Strategies in Policy Studies" (ReSPoS) has proven vital to the maturation of this project, which consolidates an experience dating back to 2013—now, a seemingly distant past. The School and the project, in turn, would not have been possible without the financial contributions of the Compagnia di San Paolo (Turin, Italy) through the Network for the Advancement of Social and Political Sciences (NASP) directed by Maurizio Ferrera. To them go our sincere gratitude. We also are greatly indebted to Springer's Senior Editor for Economics, Political Science, and Public Administration, Lorraine Klimowich, and the Editor of the 'Textbook on Political Analysis' series, Justin Esarey. Their precious suggestions and faultless encouragement have been fundamental to finalizing a project intentionally positioned at the crossroad of many disciplines and research standards. Our further debt of gratitude is owed to Luigi Curini and the Standing Groups "MetRiSP—Research Methods for Political Science" and "Political Science & Public Policy" of the SISP—Italian Society of Political Science. We have treasured their feedbacks on earlier versions and their backing the main idea. Finally, we gratefully acknowledge the financial support of Protego—an advanced project funded by the European Research Council—Grant agreement n°694632.

University of Milan
Milan, Italy

Alessia Damonte
Fedra Negri

Contents

Chapter 1
Introduction: The Elephant of Causation and the Blind Sages

Alessia Damonte and Fedra Negri

It was six men of Indostan, To learning much inclined, Who went to see the Elephant (Though all of them were blind), That each by observation Might satisfy his mind. John G. Saxe (1816–1887).

Abstract What does a policy outcome hinge on? The response is vital to policy-making and calls for the best of our knowledge from a variety of disciplines—from economics to sociology and from political science to public administration and management. The response entails a stance about causation, however, and almost every discipline has its own. Researchers are like the blind sages who had never come across the elephant of causation before and who develop their idea of the elephant by "touching" a different part of it. Which part of the elephant will you happen to touch? Will you be able to listen to and understand what the other sages will tell you?

1.1 Policy Decisions and Causal Theories

The common wisdom about public policy understands them as governments' decisions to tackle a collective problem. These decisions deploy rules, information, taxes, and expenditures to get "people to do things that they might not otherwise do" or "do things that they might not have done otherwise" (Schneider & Ingram, 1990: 513). By inducing a change in people's willingness and capacity to "do things," policy-makers expect the problem to disappear or, at least, take a more bearable shape.

A. Damonte (✉)
University of Milan, Milan, Italy
e-mail: alessia.damonte@unimi.it

F. Negri
University of Milan - Bicocca, Milan, Italy
e-mail: fedra.negri@unimib.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_1

Thus, the kernel of policy decisions is the causal theory that they encapsulate: first, of the behavior at the root of the collective problem; second and relatedly, of the capacity that certain tools have to make such behavior change for the better. The theory connects outcomes to behavior and then identifies the "carrots, sticks, and sermons" (Vedung, 2010) best suited to put or keep such behavior on a desirable track. For example, in their fight against cancer, governments can address smoking as a proven causal factor and assume people smoke if they have the wrong information or are shortsighted about the consequences of their behavior—else, they would reasonably quit. Governments can fund education campaigns to convey the right information, require tobacco products to carry warning labels, or disallow tobacco advertising and sponsorship. Moreover, to compensate for people's shortsightedness, they can levy "sin taxes" upon tobacco products to make prices a better signal of the hidden costs of smoking or enforce smoke bans that protect non-smokers. Whether a government applies none, one, or a mix of these tools, in turn, depends on policy-makers; whether their decisions reach the addressees properly, instead, is an administrative and a governance matter (e.g., McConnell, 2010). Regardless of the point of attack, the issue of policy success and failure inevitably appeals to causal theories on endowments, concerns, constraints, and incentives accounting for behavior (e.g., Ostrom, 2005).

Policy studies offer exemplary illustrations of the twofold stake of causal theories. First, these theories allow us to make sense of the world. Our bewilderment at some diversity in performance dissolves when we are offered satisfying accounts of relevant behaviors. Second, these theories have straightforward practical implications for individual and collective strategies. If we know which factors compel an event and suppress it, we can change the event's odds by controlling these factors. Then, the driving question remains: how can we get to know these factors well enough to build decisions on them?

1.2 The Elephant of Causation

Across the philosophy of science and social sciences, the responses to this question invite analogies with the blind sages in Saxe's poem (1872), who "prate about an Elephant that / Not one of them has seen."[1] Indeed, actual causation is the complex local production of an outcome and it is hard to identify before it unfolds. The usable knowledge of a causal process pinpoints the key factors of its unfolding that allow us to see it coming in the next instance and, eventually, change its odds (e.g., Craver and Kaplan, 2020). Such knowledge requires criteria to identify the key

[1] The poem tells the story of a group of blind sages who have never come across an elephant before and who learn what the elephant is like by touching it. Each blind sage feels a different part of the elephant's body, but only one part. They then describe the elephant based on their limited experience and "Though each was partly in the right, /And all were in the wrong!" (Saxe, 1872).

causal factors beyond the single case and credibly so. Historically, guidelines for identifying the key causal factors developed along two lines.

1.2.1 Elephants by the Principle

The most enduring guideline for determining the key causal factors before a process unfolds has come from the Aristotelian philosophy of science. There, causation was tracked back to four kinds of principles, known as "material," "formal," "efficient," and "final." The first two principles capture the structural features of a causal process, namely, its constituent elements and the shape of their arrangement. The latter two refer to agency and locate the key factors in outer stimuli or the drive from inner purposes (e.g., Moravcsik, 1974). The original "doctrine" maintained that adequate responses to any why-question appealed to all the four principles together.

Indeed, convincing accounts still locate actual causation in the interplay of structure and agency, as influential mechanistic perspectives make clear (e.g., Little, 2011; Craver, 2006). More often, current research streams specialize in single principles. For example, the causal role of "material" ascriptive features is a driving concern of gender and minority studies. The generative power of formal arrangements is the core tenet of, for instance, game theories. Studies on expected utility, values, habits, and emotions take heed of the final goals and motivations, providing fundamental assumptions for neo-institutionalist and behavioral approaches of various stripes. Efficient factors are any stimulus, intervention, or treatment that can elicit a response; thus, they are central to theories of policy instruments, regimes, or political communication, among many others.

With some exceptions (e.g., Bache et al., 2012; Kurki, 2006), current theories seldom claim an explicit legacy with the original canon. The doctrine has fallen into disrepute as improperly scientific, because it invoked a metaphysical reason to justify the causal standing of its four principles. The tenet that individuals with similar features, in a similar situation, with similar motivations, under equivalent stimuli did and will behave in similar ways was justified by the belief that all embodied the same metaphysical essence. As Aristotle argued in a seminal fragment, planets do not twinkle because planets are near things, and not twinkling was intrinsic to near things. Thus, the next planet will not twinkle, too, in force of its "near-thingness."

This line of reasoning easily lends itself to circular arguments that restate general assumptions instead of probing them. As late as 1673, Molière still had reasons to satirize it. In his comedy *The Hypochondriac*, a "docto doctore" explains in dog Latin that opium makes people sleepy because it embodies a "dormitive virtue." However, the ultimate criticism came from the British Empiricists, who saw in the appeal to essences a mode for preserving beliefs against evidence and a fundamental obstacle to progress and learning.

1.2.2 Elephants by the Rules

The rejection of metaphysical warrants has called for a different ground for causal inference. Whether a reliable connection exists between being a near thing and not twinkling across cases, so the argument goes, it can only be decided empirically.

Yet, causal evidence does not come to us with labels and numbers attached. Assumptions are still needed about the empirical traces that distinguish between relevant and irrelevant causal factors. In Hume's much-quoted words, causally relevant is:

> an object followed by another and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed. (Hume, 1748, Section VII, Part II, §60).

In short, a factor is relevant to an outcome in the single case under two warrants: the association of the two conforms to a *regular* pattern, and it supports *counterfactual* reasoning.

1.2.2.1 Regularity

The regularity warrant—"where all the objects, similar to the first, are followed by objects similar to the second"—renders the empirical footprint of Aristotelian essences without assuming them and builds on the repeated observation of similar occurrences.

All objects sharing the same feature are similar and constitute a distinct class. Regularity, then, is established between objects in different classes—for instance, in the class of "swan" and in the class of "white." It requires that any observation of the first class entails one in the second. When the regularity holds, causal knowledge can be circulated through handy formulae such as "if a swan, then white."

To apply to the next instance, these formulae have to prove faultless, which is hardly the case: classes and gauges are human constructs and can prove too strict or liberal to capture actual causation in the next instance. Hence, regularity holds provisionally only until we meet the black swan that forces a revision of the scope of our regularity tenets.

Regularity may also seem perfect just because we measured two consequences of the same process. These relationships are useful for prediction; however, they do not qualify as causal as they do not grant control over the events' odds as desired in public policy. Indeed, a barometric reading can be relied upon to prepare for extreme weather conditions but does not license the belief that the coming storm can be tamed by forcing the barometer's pointer. Thus, regularity can be a necessary trait of usable knowledge but insufficient to declare the causal standing of a relationship.

1.2.2.2 Counterfactual

The counterfactual—"where, if the first object had not been, the second never had existed"—enters the picture as the additional warrant to establish causal relevance and ideally applies to the factor in the single case independent of regularity. The warrant borrows from the classical rules of argumentation and the indirect proofs in geometric demonstrations; however, it displays an empirical edge. Counterfactuals link causal relevance to evidence that we could compel a change in the second object by manipulating the first.

From the Humean definition, manipulation is usually understood as suppression; more generally, it means switching the observed state of a feature into its opposite. Thus, counterfactual reasoning requires, first, that we imagine the first object with the switched feature and, then, that we can only draw impossible or contradictory conclusions from it (e.g., Levi, 2007). An exemplary illustration comes directly from Hume. Despite his deep skepticism toward the human mind's ability to fully understand causation, he conceded that our intuitions must be somehow right. To justify his claim, he reasoned that had our mind always got causation wrong (switching the feature), then humankind would have long gone extinct (drawing a conclusion), which contrasts with us thriving as a species (showing the conclusion absurd). Such counterfactual criterion improves on the regularity test, as regular non-causal features fail it: as a broken barometer cannot stop a storm, it cannot be recognized as having any causal standing.

However, counterfactuals have their limits, too. First, they cannot be established unless all the plausible alternative causes of the same outcome are ruled out. Hume's argument does not exclude that humankind's evolutionary success instead depends on, for instance, sheer luck—and the unaccounted alternative undermines the cogency of its conclusion. The second and related issue is serious to the point of earning the title of "fundamental problem of causal inference" in some quarters (e.g., Holland, 1988). Unless we cast the same causal process in the same unit with and without the feature of interest, we cannot establish whether switching the feature can change the outcome.

1.3 The Blind Sages' Portrayals as the Book's Blueprint

The criteria to establish causation by regularity and counterfactual evidence seem as straightforward as impossible to meet. Nevertheless, techniques have been developed as strategies to circumvent the Humean paradoxes and provide empirical warrants to the claim of causal relevance. As Little shows in Chap. 2, technical specialization has undermined the dialogue among techniques and their findings. The appeal to regularity, counterfactual, or mechanistic principles has turned into as many ultimate understandings of causation: "laws" and counterfactuals offered a

rival ground for experimental practices; mechanisms took distances from both and licensed causal analysis in actual cases only, under consideration that any conclusion about aggregates necessarily entails an unfaithful reduction—in the end, all models are wrong.

However, the possibility of integration remains when techniques commit to three considerations and are consistent with a reasonable scientific realism. First, causation is real, but our best knowledge of it remains a useful approximation. Second, regularity and counterfactuals are epistemic criteria to establish whether portrayals qualify as valid causal accounts; mechanisms are ontological assumptions about single actual elephants instead. Third, the difference between mechanistic description, models, and laws is not of kind but degree: when they address a common slice of the world, they provide a map of it with different details, abstraction, and scope. Under these commitments, techniques can be understood as devices to respond to special questions about the elephant.

1.3.1 Can this Single Factor Make Any Difference?

The family of experimental and quasi-experimental techniques offers the most renowned, successful, and contentious example at once due to the diffusion of randomized controlled trials as the "gold standard" of scientific knowledge production (e.g., Kabeer, 2020; Deaton & Cartwright, 2018; Dawid, 2000). This family shares the consideration that although we cannot observe a counterfactual directly, we can construe credible "twin worlds" and "treat" one so that the feature of interest provides the only difference to which the difference in responses can be ascribed.

As Battistin and Bertoni show in Chap. 3, this strategy keeps the role of causal assumptions to the minimum required by a stimulus-response model: the treatment is a supposedly efficient cause and connected to performance by a function of a specific shape—often, linear—without further details. Unsurprisingly, these techniques are a cornerstone of usable public policy knowledge: they can establish the capacity of a change in taxation, expenditure, information, and regulation to elicit some effect of interest, apparently without the need for further knowledge.

The credibility of this strategy's conclusions, however, rests heavily on the research design: findings are sound if the twin worlds are construed as statistically identical and independent aggregates, the treatment is forced evenly onto all the units of one world only, and the difference in responses is not affected by the treating procedure or unrelated endogenous dynamics. The threats arise as the statistical aggregates with identical parameters can hide a remarkable inner heterogeneity that may bias both groups' responses in unknown directions. As elaborated by Negri in Chap. 4 and Ornstein in Chap. 5, within the family, this heterogeneity is addressed as the result of selection biases that can be reduced by accounting for observed imbalances and crafting "populations of twins." The solution, however, leaves the issue open of the bending effects from unobservable factors.

The (quasi-)experimental family, in short, can provide reliable measures of the net effect of a treatment, but necessarily at the cost of disregarding the reasons for the diversity in the responses of the treated.

1.3.2 Through Which Structures?

The diversity in responses is instead the driving concern of the second group of techniques. They address it by flipping the experimental balance of model and design and committing themselves to additional assumptions. They conceive of the generative process as patterns of dependence and assign causal relevance to the bundle of factors that fit them.

The reliance on models sidelines the issue of unit selection as, ideally, any unit carries usable information about the tenability of the causal structure of interest. The structure, moreover, provides the fixed points that still make counterfactuals observable. However, models require criteria to select meaningful variables, and structural assumptions provide partial guidance to it. The main decisions can only be made in light of substantive theories about the generation of the outcome—hence, of some previous local knowledge. Within this framework, each technique relies on different languages and pursues different goals.

Path analysis develops within a Bayesian mindset and understands causation as ordered dependencies fitting a few known shapes: chains, colliders, and forks. As Röth clarifies in Chap. 6, these shapes explain because they elaborate on the connection between an alleged causal condition and the dependent by displaying the intermediate causal link, the common factor, or the equivalent alternative factors that support the hypothesis about the unfolding of the causal process before the outcome. The technique supports a neater identification of the mechanism linking a factor of interest and its outcome, affords counterfactual analysis, and provides specific suggestions about the "scope conditions" ensuring the mechanisms. Röth contends that these features qualify path analysis as the natural companion of experimental studies for its capacity to establish the contextual requirements that enhance and refine the validity of their findings.

Qualitative comparative analysis (QCA) instead builds on sets and Boolean algebra and understands causal structures as teams of individually necessary and jointly sufficient factors to an outcome. In Chap. 7, Damonte makes three points about the explanatory import of the technique. First, its assumptions about the shape of causation support complex causal theories about the interactions of triggering, enabling, or shielding conditions of some underlying causal process. Second, its parameters of fit allow diagnosing the underspecification of the theory to the cases at hand, while the algorithm provides a pruning counterfactual device that takes care of its overspecification. Last, sets remap qualities onto quantities, which warrant meaningful and sound solutions. Thus, QCA can formalize and test theories about the teams of conditions beneath policy success and failure across given cases beyond

special processes. As such, the technique especially suits the purpose of systematic *ex-post* evaluation of policy designs.

1.3.3 Through Which Process?

The knowledge of the dynamics of a causal situation is the missing piece of knowledge and the core concern of two further strategies, aiming to open up the black box of causation. Both share the direct interest in the actors and their interplay as the ultimate ground of causation, although their point of attack within the causal stream of actions is different.

Bayesian process tracing addresses causation within its local context. In Chap. 8, Bennett shows how analysts can rely on this technique to make causal sense of the chain of events to policy success or failure retrospectively. The strategy understands hypotheses as plausible Bayesian beliefs that we can entertain about the causal process and that evidence can confirm or disconfirm. The weight of evidence rests on the assumption that each hypothesis corresponds to a specific sequence of actions and events that leave empirical traces. When the connection between a piece of evidence and a hypothesis is unique, certain, or both, the actual retrieval of certain traces in a case contributes to ranking hypotheses by their relative likelihood and eventually licenses the ascription of the case to the hypothesis with the best standing.

Last but not the least, agent-based models make it possible to test hypotheses about causal processes as emergent phenomena in silico. As Squazzoni and Bianchi illustrate in Chap. 9, the technique relies on simulation to verify whether a certain alignment of assumptions about actors and their constraints, when translated into conditional rules of individual behavior and recursively played, returns performance values close to the empirical responses of actual systems. The strategy requires regularity and counterfactual assumptions about the options available to each agent, rendered as alternative states, and about the consequence of choosing a state conditional on the states of the relevant neighbors. These models shed light on the tenability of different understandings of the mechanism that alternative policy constraints or endowments activate in the field.

1.3.4 Considerations and Extensions

The order of the chapters, as Beach and Siewert reason in their Chap. 10, chimes with the common prescription in mixed method research that a better causal knowledge follows from a succession of techniques zooming into individual cases, where causation unfolds as actual processes and explanations can find their ultimate validation. However, they consider the downward path of mixed methods lays knowledge open to heterogeneity threats. The actual heterogeneity is always equal to the number of instances under analysis; cross-case knowledge, however, requires that

we dismiss some heterogeneity as irrelevant to afford comparisons and causal inferences. The move to local contexts implies a twofold shift—from a low to a high number of factors in the analysis and from coarse types to fine-grained tokens of evidence—that seldom support cross-case findings. Hence, they contend that a more fruitful and conventional strategy follows the upward path from local processes over structures to the causal capacity of single triggers. This path allows more conscious decisions about heterogeneity that can improve models and gauges.

In Chap. 11, Damonte and Negri conclude the journey. The chapter recognizes the fragmented image of causation that the previous contributions convey and asks whether such fragmentation is an undesirable state of affairs, as claimed by a long-honored narrative from the history of science, or an eventually valuable situation, as argued in the pluralist quarters of the philosophy of science. The point of contention concerns the inability to yield dovetailing knowledge that would affect strategies built on alternative tenets. The chapter revises these tenets and contends that, whereas ontology offers complementary angles of attack to the causal elephant and epistemology licenses interpretations that can estrange research communities from one another, methodological reasoning about models and designs reconciles the analyses when it emphasizes that causation corresponds to a few recognized shapes. These shapes, the chapter concludes, offer a rough yet common map of the elephant that strategies of any stripe can detail and enrich while pursuing their special research interests—thus contributing to better policy knowledge.

References

Bache, I., Bulmer, S., & Gunay, D. (2012). Europeanization: A critical realist perspective. In T. Exadaktylos & C. M. Radaell (Eds.), *Research design in European studies* (pp. 64–84). Springer.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355–376. https://doi.org/10.1007/s11229-006-9097-x

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95*(450), 407–424. https://doi.org/10.1080/01621459.2000.10474210

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.

Holland, P. W. (1988). Causal inference path analysis and recursive structural equations models. *ETS Research Report Series, 1988*(1), i–50. https://doi.org/10.1002/j.2330-8516.1988.tb00270.x

Hume, D. (1748). *An enquiry concerning human understanding*. Section VII.

Kabeer, N. (2020). Women's empowerment and economic development: A feminist critique of storytelling practices in 'Randomista' economics. *Feminist Economics, 26*(2), 1–26. https://doi.org/10.1080/13545701.2020.1743338

Kurki, M. (2006). Causes of a divided discipline: Rethinking the concept of cause in international relations theory. *Review of International Studies, 32*(2), 189–216. https://doi.org/10.1017/s026021050600698x

Levi, I. (2007). *For the sake of the argument: Ramsey test conditionals, inductive inference and nonmonotonic reasoning*. Cambridge University Press.

Little, D. (2011). Causal mechanisms in the social realm. In P. M. K. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 273–295). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199574131.003.0013

McConnell, A. (2010). Policy success, policy failure and gray areas in-between. *Journal of Public Policy, 30*(3), 345–362. https://doi.org/10.1017/S0143814X10000152

Moravcsik, J. M. E. (1974). Aristotle on adequate explanations. *Synthese, 28*, 3–17. https://doi.org/10.1007/BF00869493

Ostrom, E. (2005). *Understanding institutional diversity*. Princeton University Press.

Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics, 52*(2), 510–529. https://doi.org/10.2307/2131904

Vedung, E. (2010). Policy instruments: Typologies and theories. In M.-L. Bemelmans-Videc, R. C. Rist, & E. Vedung (Eds.), *Carrots, sticks and sermons: Policy instruments and their evaluation* (pp. 21–58). Transaction. https://doi.org/10.4324/9781315081748

Chapter 2
Causation in the Social Realm

Daniel Little

Abstract Explanation is at the center of scientific research, and explanation almost always involves the discovery of causal relations among factors, conditions, or events. This is true in the social sciences no less than in the natural sciences. But social causes look quite a bit different from causes of natural phenomena. They result from the choices and actions of numerous individuals rather than fixed natural laws, and the causal pathways that link antecedents to consequents are less exact than those linking gas leaks to explosions. It is, therefore, a crucial challenge for the philosophy of social science to give a compelling account of causal reasoning about social phenomena that does justice to the research problems faced by social scientists.

Learning Objectives
By studying this chapter, you will:

- Gain exposure to philosophical theories of causal explanation.
- Learn how "ontology" is important in social research.
- Learn about the theory of causal mechanisms.
- Become acquainted with how several causal research methodologies relate to social ontology.
- Become acquainted with scientific realism as an approach to social research.

2.1 Why Discuss the Ontology of Causation?

Ontology precedes methodology. We cannot design good methodologies for scientific research without having reasonably well-developed ideas about the nature of the phenomena that we intend to investigate (Little, 2020). This point is especially important in approaching the idea of social causation. Only when we have a reasonably clear understanding of the logic and implications of the scientific idea of

D. Little (✉)
University of Michigan-Dearborn, Dearborn, MI, USA
e-mail: delittle@umich.edu

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_2

causality can we design appropriate methods of inquiry for searching out causal relations. And only then can we give a philosophically adequate justification of existing methods—that is, an account of how the research method in question corresponds to a sophisticated understanding of the nature of the social world.

Here I will work within the framework of an "actor-centered" view of social ontology (Little, 2006, 2014, 2016). On this view, the social realm is constituted by individual actors who themselves have been cultivated and developed within ongoing social relations and who conduct their lives and actions according to their understandings and purposes. Social structures, social institutions, organizations, normative systems, cultures, and technical practices all derive their characteristics and causal powers from the socially constituted and situated individuals who make them up (Little, 2006).

This fact about social entities and processes suggests a high degree of contingency in the social world. Unlike chemistry, the social world is not a system of law-governed processes; it is instead a mix of different sorts of institutions, forms of human behavior, natural and environmental constraints, and contingent events. The entities that make up the social world at a given time and place have no essential ontological stability; they do not fall into "natural kinds"; and there is no reason to expect deep similarity across a number of ostensibly similar institutions—states, for example, or labor unions. The "things" that we find in the social world are heterogeneous and contingent. And the metaphysics associated with classical thinking about the natural world—laws of nature; common, unchanging structures; and fully predictable processes of change—do not provide appropriate building blocks for our understandings and expectations of the social world nor do they suggest the right kinds of social science theories and constructs.

Instead of naturalism, this actor-centered approach to social ontology leads to an approach to social science theorizing that emphasizes agency, contingency, and plasticity in the makeup of social facts. It recognizes that there is a degree of pattern in social life, but emphasizes that these patterns fall far short of the regularities associated with laws of nature. It emphasizes contingency of social processes and outcomes. It insists upon the importance and legitimacy of eclectic use of multiple social theories: social processes and entities are heterogeneous, and therefore, it is appropriate to appeal to different types of social theories as we explain various parts of the social world. It emphasizes the importance of path dependence in social outcomes.

Box 2.1 Definitions

Agency: The fact that social change and causation derives from the purposive actions of individual social actors.

Contingency: Social outcomes depend upon conjunctions of occurrences that need not have taken place, so the outcome itself need not have taken place. Closely related to "path dependency."

(continued)

Box 2.1 (continued)

Path dependency: The feature of social processes according to which minor and underdetermined events in an early stage of a process make later changes more probable. For example, the QWERTY arrangement of the typewriter keyboard was selected in order to prevent typists from jamming the mechanism by typing too rapidly. Fifty years later, after widespread adoption, it proved impossible to adopt a more efficient arrangement of the keys to permit more rapid typing.

Plasticity: A feature of an entity or group of entities according to which the properties of the entity can change over time. Biological species demonstrate plasticity through evolution, and social entities demonstrate plasticity through the piecemeal changes introduced into them by a variety of actors and participants.

How does this ontological perspective fit with current work in policy studies? There are several current fields of social research that illustrate this approach particularly well. One is the field of the "new institutionalism." Researchers in this tradition examine the specific rules and incentives that constitute a given institutional setting. They examine the patterns of behavior that these rules and incentives give rise to in the participants in the institution, and they consider as well the opportunities and incentives that exist for various powerful actors to either maintain the existing institutional arrangements or modify them. Kathleen Thelen's (2004) study of different institutions of skill formation in Germany, Great Britain, the United States, and Japan is a case in point. This approach postulates the causal reality of institutions and the specific ensembles of rules, incentives, and practices that make them up; it emphasizes that differences across institutions lead to substantial differences in behavior; and it provides a basis for explanations of various social outcomes. The rules of liability governing the predations of cattle in East Africa or Shasta County, California, create very different patterns of behavior in cattle owners and other landowners in the various settings (Ellickson, 1991). It is characteristic of the new institutionalism that researchers in this tradition generally avoid reifying large social institutions and look instead at the more proximate and variable sets of rules, incentives, and practices within which people live and act.

2.2 Scientific Realism About the Social World and Social Causation

We are best prepared for the task of discovering causal relationships in the social world when we adopt a realist approach to the social world and to social causation. We provide an explanation of an event or pattern when we succeed in identifying the real causal conditions and events that brought it about. The central tenet of causal

realism is a thesis about causal mechanisms and causal powers. Causal realism holds that we can only assert that there is a causal relationship between X and Y if we can offer a credible hypothesis of the sort of underlying mechanism that connects X to the occurrence of Y. The sociologist Mats Ekström puts the view this way: "the essence of causal analysis is … the elucidation of the processes that generate the objects, events, and actions we seek to explain" (Ekström, 1992: 115). Authors who have urged the centrality of causal mechanisms for explanatory purposes include Roy Bhaskar (1975), Nancy Cartwright (1989), Jon Elster (1989), Rom Harré and Madden (1975), Wesley Salmon (1984), and Peter Hedström (2005).

Scientific realism about social causes comes down to several simple ideas.

First, there is such a thing as social causation. Causal realism is a defensible position when it comes to the social world: there are real causal relations among social factors (structures, institutions, groups, norms, and salient social characteristics like race or gender). We can give a rigorous interpretation to claims like "racial discrimination causes health disparities in the United States" or "rail networks cause changes in patterns of habitation."

Second, causal relations among factors or events depend on the existence of real social-causal mechanisms linking cause to effect. Discovery of correlations among factors does not constitute the whole meaning of a causal statement. Rather, it is necessary to have a hypothesis about the mechanisms and processes that give rise to the correlation. Hypotheses about the causal mechanisms that exist among factors of interest permit the researcher to exclude spurious correlation (cases where variations in both factors are the result of some third factor) and to establish the direction of causal influence (cases where it is unclear whether the correlation between A and B results from A causing B or B causing A). So mechanisms are more fundamental than regularities.

Third, the discovery of social mechanisms in policy studies often requires the formulation of mid-level theories and models of these mechanisms and processes—for example, the theory of free-riders. For example, an urban policy researcher may observe that racially mixed high-poverty neighborhoods have higher levels of racial health disparities than racially mixed low-poverty neighborhoods. This is an observation of correlation. Researchers like Robert Sampson (2010) would like to know how "neighborhood effects" work in transmitting racial health disparities. What are the mechanisms by which a neighborhood influences the health status of an individual household? In order to attempt to answer this question, Sampson turns to mid-level hypotheses in urban sociology that contribute to a theory of the mechanisms involved in this apparent causal relationship. By mid-level theory, I mean essentially the same thing that Robert Merton (1963) conveyed when he introduced the term: an account of the real social processes that take place above the level of isolated individual action but below the level of full theories of whole social systems. Marx's theory of capitalism illustrates the latter; Jevons's theory of the individual consumer as a utility maximizer illustrates the former. Coase's theory of transaction costs (Coase, 1988) is a good example of a mid-level theory: general enough to apply across a wide range of institutional settings, but modest enough in

its claim of comprehensiveness to admit of careful empirical investigation. Significantly, the theory of transaction costs has spawned major new developments in the new institutionalism in sociology (Brinton & Nee, 1998).

And finally, it is important to recognize and welcome the variety of forms of social scientific reasoning that can be utilized to discover and validate the existence of causal relations in the social world. Properly understood, there is no contradiction between the effort to use quantitative tools to chart the empirical outlines of a complex social reality, and the use of theory, comparison, case studies, process tracing, and other research approaches aimed at uncovering the salient social mechanisms that hold this empirical reality together.

2.2.1 Critical Realism

Critical realism is a specific tradition within the late-twentieth-century analytic philosophy that derives from the work of Rom Harré and Roy Bhaskar (Harré & Madden, 1975; Bhaskar, 1975; Archer et al., 2016). In brief, the view holds that the ontological stance of realism is required for a coherent conception of scientific knowledge itself. Unqualified skepticism about "unobservable entities" makes scientific research and experimentation philosophically incoherent. We are forced to take the view that the entities postulated by our best theories of the world are "real"—whether electrons, viruses, or social structures. For Bhaskar, this ontological premise has much the status of Kant's transcendental arguments for causation and space and time: we cannot make sense of experience without postulating causation and locations in space and time (Bhaskar, 1975).

Concretely in the social sciences, this is taken to mean that we can be confident in asserting that social entities exist if these concepts play genuine roles in well-developed and empirically supported theories of the social world: for example, organizations, markets, institutions, social classes, normative systems, rules, ideologies, and social networks. Further, we can be confident in attributing causal powers and effects to the various social entities that we have identified—always to be supported by empirical evidence of various kinds.

2.3 What Is Causation?

Let us turn now to a more specific analysis of causation. What do we mean by a cause of something? Generally speaking, a cause is a circumstance that serves to bring about (or renders more probable) its effect, in a given environment of background conditions. Causes *produce* their effects (in appropriate background conditions). A current fruitful approach is to understand causal linkages in terms of the specific *causal mechanisms* that link cause to effect.

We can provide a preliminary definition of causation along these lines:

- A causes B in the presence of $C_i =_{def}$. A suffices to bring about B in the presence of conditions C_i (sufficiency).
- A causes B in the presence of $C_i =_{def}$. If C_i were present but A had not occurred, then B would not have occurred (necessity).

That is, A is necessary and sufficient in conditions C_i for the production of B. This definition can be understood in either a deterministic version or a probabilistic version. The deterministic version asserts that A in the presence of C_i always brings about B; the probabilistic version asserts that the occurrence of A in the presence of C_i increases the likelihood of the occurrence of B.

There is a fundamental choice to be made when we consider the topic of causation. Are causes real, or are causal statements just summaries of experimental and observational results and the statistical findings that can be generated using these sets of data? The first approach is the position described above as causal realism, while the second can be called causal instrumentalism. If we choose causal realism, we are endorsing the idea that there is such a thing as a *real* causal linkage between A and B; that A has the power to produce B; and that there is such a thing as causal necessity. If we choose causal instrumentalism, we are agnostic about the underlying realities of the situation, and we restrict our claims to observable patterns and regularities. The philosopher David Hume (2007) endorsed the second view; whereas many philosophers of science since the 1970s have endorsed the former view.

Most of the contributors to the current volume engage with the premises of causal realism. They believe that social causation is real; there are real social relations among social factors (structures, institutions, groups, norms, and salient social characteristics like race or gender), and there are real underlying causal mechanisms and powers that constitute those causal relations. According to scientific realists, a key task of science is to discover the causal mechanisms and powers that underlie the observable phenomena that we study.

Causal realists acknowledge a key intellectual obligation that goes along with postulating real social mechanisms: to provide an account of the ontological *substrate* within which these mechanisms operate**.** In the social realm, the substrate is the system of social actors whose mental frameworks, actions, and relationships constitute the social world. This is what is meant by an "actor-centered" ontology of the social world. On this view, every social mechanism derives from facts about individual actors, the institutional context, the features of the social construction and development of individuals, and the factors governing purposive agency in specific sorts of settings. Different research programs in the social sciences target different aspects of this nexus.

This view of the underlying reality of social causation justifies a conception of causal necessity in the social realm. Do causes make their effects "necessary" in any useful sense? This is the claim that Hume rejected—the notion that there is any "necessary" connection between cause and effect. By contrast, the notion of *natural necessity* is sometimes invoked to capture this idea:

- A causes B: given the natural properties of A and given the laws of nature and given the antecedent conditions, B necessarily occurs.

This can be paraphrased as follows:

- Given A, B occurs as a result of natural necessity.

So the sense of necessity of the occurrence of the effect in this case is this: given A and given the natural properties and powers of the entities involved, B had to occur. Or in terms of possible worlds and counterfactuals (Lewis, 1973), we can say:

- In any possible world in which the laws of nature obtain, when A occurs, B invariably occurs as well.

Applied to social causation within the context of an ontology of actor-centered social facts, here is what causal necessity looks like:

- Given the beliefs, intentions, values, and goals of various participants and given the constraints, opportunities, and incentives created by the social context, whenever A occurs, the outcome B necessarily occurs [financial crisis, ethnic violence, rapid spread of infectious disease …].

This conception aligns with Wesley Salmon's idea of the "causal structure of the world," applied to the social world (1984). And this in turn indicates why causal mechanisms are such an important contribution to the analysis of causation. A causal mechanism is a constituent of this "stream of events" leading from A to B.

Probabilistic causal relations involve replacing exceptionless connections among events with probabilistic connections among events. A has a probabilistic causal relationship to B just in case the occurrence of A increases (or decreases) the likelihood of the occurrence of B. This is the substance of Wesley Salmon's (1984) criterion of causal relevance. Here is Salmon's idea of causal relevance:

- A is causally relevant to B *if and only if* the conditional probability of B given A is different from the absolute probability of B (Salmon, 1984, adapted notation).

For a causal realist, the definition is extended by a hypothesis about an underlying causal mechanism. For example, smoking is causally relevant to the occurrence of lung cancer [working through physiological mechanisms X, Y, Z]. And cell physiologists are expected to provide the mechanisms that connect exposure to tobacco smoke to increased risk of malignant cell reproduction.

It is important to emphasize that we can be causal realists about probabilistic causes just as we can about deterministic causes. A causal power or capacity is expressed as a tendency to produce an outcome; but this tendency generally requires facilitating conditions in order to be operative. The causal power is appropriately regarded as being real, whether or not it is ever stimulated by appropriate events and circumstances. A given cube of sugar is soluble, whether or not it is ever immersed in water at room temperature.

These definitions have logical implications that suggest different avenues of research and inquiry in the social sciences. First, both the deterministic and the probabilistic versions imply the truth of a *counterfactual* statement: If A had not

occurred in these circumstances, B would not have occurred. (Or if A had not occurred in these circumstances, the probability of B would not have increased.) The counterfactual associated with a causal assertion suggests an experimental approach to causal inquiry. We can arrange a set of circumstances involving C_i and remove the occurrence of A and then observe whether B occurs (or observe the conditional probability of the occurrence of B).

Another important implication of a causal assertion is the idea of a set of necessary and sufficient conditions for the occurrence of E, the circumstance of explanatory interest. With deterministic causation, the assertion of a causal relationship between A and B implies that A is sufficient for the occurrence of B (in the presence of C_i) and often the assertion implies that A is a necessary condition as well. (If A had not occurred, then B would not have occurred.) On these assumptions, a valid research strategy involves identifying an appropriate set of cases in which A, C_i, and B occur, and then observe whether the appropriate covariances occur or not. J. L. Mackie (1974) provided a more detailed analysis of the logic of necessary and sufficient conditions in complex conjunctural causation with his concept of an INUS condition: "*insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" (62). Significantly, Mackie's formulation provides a basis for a Boolean approach to discovering causal relations among multiple factors.

These definitions and logical implications give scope to a number of different strategies for investigating causal relationships among various conditions. For probabilistic causal relationships, we can evaluate various sets of conditional probabilities corresponding to the presence or absence of conditions of interest. For deterministic causal relationships, we can exploit the features of necessary and sufficient conditions by designing a "truth table" or Boolean test of the co-occurrence of various conditions (Ragin, 1987). This is the logic of Mill's methods of similarity and difference (Mill, 1988; Little, 1995). For both deterministic and probabilistic causal relationships, we can attempt to discover and trace the workings of the causal mechanisms that link the occurrence of A to the occurrence of B.

2.3.1 Causal Mechanisms

As noted above, the central tenet of causal realism is a thesis about the real existence of causal mechanisms and causal powers. The fundamental causal concept is that of a mechanism through which A brings about or produces B (Little 2011). According to this approach, we can only assert that there is a causal relationship between A and B if we can offer a credible hypothesis of the sort of underlying mechanism that connects A to the occurrence of B. This is central to our understanding of causation from single-case studies to large statistical studies suggesting causal relationships between two or more variables. Peter Hedström and other exponents of analytical sociology are recent voices for this approach for the social sciences (Hedström, 2005; Hedström & Ylikoski, 2010). An important paper by Machamer et al. (2000) sets the terms of current technical discussions of causal mechanisms, and James

Mahoney (2001) surveyed the various theories of causal mechanisms and called for a greater specificity.

What is a causal mechanism? Consider this formulation: a causal mechanism is a sequence of events, conditions, and processes leading from the explanans to the explanandum (Little, 1991: 15, 2016: 190–192). A causal relation exists between A and B if and only if there is a set of causal mechanisms that lead from A to B. This is an ontological premise, asserting that causal mechanisms are real and are the legitimate object of scientific investigation.

The theory has received substantial development in the biological sciences. Glennan et al. (2021) put the mechanisms theory in the form of six brief theses:

(1) The most fruitful way to define mechanisms is that a mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon.
(2) Scientists can only discover, describe, and explain mechanisms through the construction of models, and these models are inevitably partial, abstract, idealized and plural.
(3) Mechanistic explanations are ubiquitous across the empirical sciences.
(4) Emphasizing that mechanistic explanations are ubiquitous in all scientific disciplines does not entail that all scientific explanations are mechanistic.
(5) The diversity of kinds of mechanisms requires and explains the diversity of tools, strategies and heuristics for mechanism discovery.
(6) The mechanisms literature is a rich source of insights that can be used to address challenging reasoning problems in science, technology and evidence-based policy.

This definition is developed for explanations in biology, but it works well with typical examples of social mechanisms.

The idea that there are real mechanisms embodied in a given domain of phenomena provides a way of presenting causal relations that serves as a powerful alternative to the pure regularity view associated with Hume and purely quantitative approaches to causation. Significantly, this is the thrust of Judea Pearl's development of structural equation modeling (discussed below): in order to get a basis for causal inference out of a statistical analysis of a large dataset, it is necessary to provide a theory of the causal mechanisms and relations that are at work in this domain (Pearl, 2021).

Mechanisms bring about specific effects. For example, "over-grazing of the commons" is a mechanism of resource depletion. Whenever the conditions of the mechanism are satisfied, the result ensues. Moreover, we can reconstruct why this would be true for purposive actors in the presence of a public good (Hardin, 1968). Or consider another example from the social sciences: "the mechanism of stereotype threat causes poor performance on standardized tests by specific groups" (Steele, 2011). This mechanism is a hypothesized process within the cognitive–emotional system of the subjects of the test, leading from exposure to the stereotype threat through a specified cognitive–emotional mechanism to impaired performance on the test. So we can properly understand a claim for social causation along these lines: "C causes E" rests upon the hypothesis that "there is a set of causal mechanisms that convey circumstances including C to circumstances including E." In the social realm, we can be more specific. "C causes E" implies the belief that "there is a set of opportunities, incentives, rules, and norms in virtue of which actors in the presence of C bring about E through their actions."

Are there any social mechanisms? There are many examples from every area of social research. For example: "Collective action problems often cause strikes to fail." "Increasing demand for a good causes prices to rise for the good in a competitive market." "Transportation systems cause shifts of social activity and habitation." "Recognition of mutual interdependence leads to medium-term social cooperation in rural settings." In each case, we have a causal claim that depends on a hypothesis about an underlying behavioral, cognitive, or institutional mechanism producing a pattern of collective behavior.

The discovery of social mechanisms often requires the formulation of mid-level theories and models of these mechanisms and processes—for example, the theory of free-riders or the theory of grievance escalation in contentious politics. Mid-level theories in the social sciences can be viewed as discrete components of a toolbox for explanation. Discoveries about specific features of the workings of institutions, individual-collective paradoxes, failures of individual rationality like those studied in behavioral economics—all of these mid-level theories of social mechanisms can be incorporated into an account of the workings of specific social ensembles. The response of a university to a sudden global pandemic may be seen as an aggregation of a handful of well-known institutional dysfunctions, behavioral patterns, and cognitive shortcomings on the part of the various actors.

Aage Sørensen summarizes a causal realist position for the social and policy sciences in these terms: "Sociological ideas are best reintroduced into quantitative sociological research by focusing on specifying the mechanisms by which change is brought about in social processes" (Sørensen, 1998: 264). Sørensen argues that social explanation requires better integration of theory and evidence. Central to an adequate explanatory theory, however, is the specification of the mechanisms that are hypothesized to underlie a given set of observations. "Developing theoretical ideas about social processes is to specify some concept of what brings about a certain outcome—a change in political regimes, a new job, an increase in corporate performance, … The development of the conceptualization of change amounts to proposing a mechanism for a social process" (Sørensen, 1998: 239–240). If an educational policy researcher finds that there is an empirical correlation between schools that have high turnover of teaching staff and high dropout rates, it is very important to investigate whether there is a mechanism that leads from teacher turnover to student dropout. Otherwise, both characteristics may be the joint result of a third factor (inadequate school funding, for example). Sørensen makes the critical point that one cannot select a statistical model for analysis of a set of data without first asking the question, "What in the nature of the mechanisms do we wish to postulate to link the influences of some variables with others?" Rather, it is necessary to have a hypothesis of the mechanisms that link the variables before we can arrive at a justified estimate of the relative importance of the causal variables in bringing about the outcome.

Emphasis on causal mechanisms for adequate social explanation has several favorable benefits for policy research. Policy research is always concerned about

causation: what interventions can be made that would bring about different outcomes? When policy researchers look carefully for the social mechanisms that underlie the processes that they study, they are in a much better position to diagnose the reasons for poor outcomes and to recommend interventions that will bring about better outcomes. Emphasis on the need for analysis of underlying causal mechanisms takes us away from uncritical reliance on uncritical statistical models.

2.3.2 Causal Powers

Some philosophers of science have argued that substantive theories of causal powers and properties are crucial to scientific explanation. Leading exponents of this view include Rom Harré (Harré & Madden 1975), Nancy Cartwright (1989), and Stephen Mumford (2009). Nancy Cartwright places real causal powers and capacities at the center of her account of scientific knowledge (1989). As she and John Dupré put the point, "things and events have causal capacities: in virtue of the properties they possess, they have the power to bring about other events or states" (Dupré & Cartwright, 1988). Cartwright argues, for the natural sciences, that the concept of a real causal connection among a set of events is more fundamental than the concept of a law of nature. And most fundamentally, she argues that identifying causal relations requires substantive theories of the causal powers ("capacities", in her language) that govern the entities in question. Causal relations cannot be directly inferred from facts about association among variables. As she puts the point, "No reduction of generic causation to regularities is possible" (1989: 90). The importance of this idea for sociological research is profound; it confirms the notion shared by many researchers that attribution of social causation depends inherently on the formulation of good, middle-level theories about the real causal properties of various social forces and entities.

Cartwright's philosophy of causation points to the idea of a causal power—a set of propensities associated with a given entity that actively bring about the effect. The causal powers theory rests on the claim that causation is conveyed from cause to effect through the active *powers and capacities* that inhere in the entities making up the cause.

The idea of an ontology of causal powers is that certain kinds of things (metals, gases, military bureaucracies) have internal characteristics that lead them to interact causally with the world in specific and knowable ways. This means that we can sometimes identify dispositional properties that attach to kinds of things. Metals conduct electricity; gases expand when heated; military bureaucracies centralize command functions (Harré & Madden, 1975). Stephen Mumford and Rani Lill Anjum explore the philosophical implications of a powers theory of causation (2011).

The language of causal powers allows us to incorporate a number of typical causal assertions in the social sciences: "Organizations of type X produce lower rates of industrial accidents"; "paramilitary organizations promote fascist mobilization"; "tenure systems in research universities promote higher levels of faculty research productivity." In each case, we are asserting that a certain kind of social organization possesses, in light of the specifics of its rules and functioning, a disposition to stimulate certain kinds of participant behavior and certain kinds of aggregate outcomes. This is to attribute a specific causal power to species of organizations and institutions.

Sociologist James Coleman offered the view that we should distinguish carefully between macro-level social factors and micro-level individual action (Coleman, 1990). He held that all social causation proceeded through three distinct paths: social factors that influence individual behavior, individuals who interact with each other and create new social facts, and the creation of new macro-level social factors that are the aggregate result of individual actions and interactions at the micro-level. Coleman did not believe that there were direct causal influences from one macro-level social fact to another macro-level social fact. Coleman offered a diagram of this view, which came to be known as "Coleman's boat" (Fig. 2.1). On this view, when we say that a certain social entity, structure, or institution has a certain power or capacity, we mean something reasonably specific: given its configuration, it creates an environment in which individuals commonly perform a certain kind of action. This is the downward strut in the Coleman's boat diagram, labeled 1 in Fig. 2.1. This approach has two important consequences. First, social powers are not "irreducible"—rather, we can explain how they work by analyzing the specific environment of formation and choice they create. And second, they cannot be regarded as deriving from the "essential" properties of the entity. Change the institution even slightly and we may find that it has very different causal powers and capacities. Change the rules of liability for open-range grazing and you get different patterns of behavior by ranchers and farmers (Ellickson, 1991).

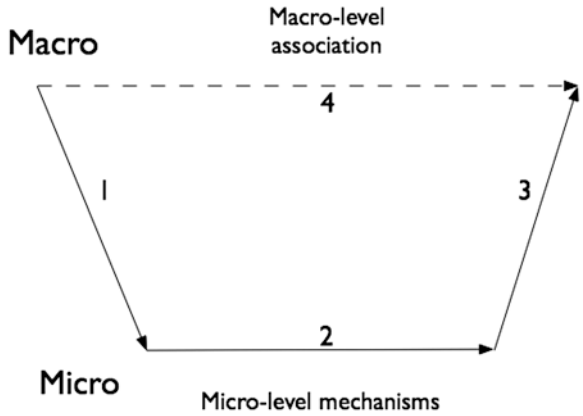

Fig. 2.1 Coleman's boat. (Author's diagram after Coleman, 1990)

2.3.3 Manipulability and Invariance

Several other aspects of the causal structure of the world have been important in recent discussions of causality in the social sciences. Jim Woodward is a leading exponent of the manipulability (or interventionist) account. He develops his views in detail in his recent book, *Making Things Happen: A Theory of Causal Explanation* (2003). The view is an intuitively plausible one: causal claims have to do with judgments about how the world would be if we altered certain circumstances. If we observe that the concentration of sulfuric acid is increasing in the atmosphere leading to acid rain in certain regions, we might consider the increasing volume of H_2SO_4 released by coal power plants from 1960 to 1990. And we might hypothesize that there is a causal connection between these facts. A counterfactual causal statement holds that if X (increasing emissions) had not occurred, then Y (increasing acid rain) would not have occurred. The manipulability theory adds this point: if we could remove X from the sequence, then we would alter the value of Y. And this, in turn, makes good sense of the ways in which we design controlled experiments and policy interventions.

Woodward extends this analysis to develop the idea of a relationship that is "invariant under intervention." This idea follows the notion of experimental testing of a causal hypothesis. We are interested in the belief that "X causes Y." We look for interventions that change the state of Y. If we find that the only interventions that change Y, do so through their ability to change X, then the X–Y relation is said to be invariant under intervention, and X is said to cause Y (Woodward, 2003: 369–370). Woodward now applies this idea to causal mechanisms. A mechanism consists of separate components that have intervention–invariant relations to separate sets of outcomes. These components are modular: they exercise their influence independently. And, like keys on a piano, they can be separately activated with discrete results. This amounts to a precise and novel specification of the meaning of "causal mechanism": "So far I have been arguing that components of mechanisms should behave in accord with regularities that are invariant under interventions and support counterfactuals about what would happen in hypothetical experiments" (374).

A related line of thought on causal analysis is the idea of *difference-making*. This approach to causation focuses on the explanations we are looking for when we ask about the cause of some outcome. Here philosophers note that there are vastly many conditions that are causally necessary for an event but do not count as being explanatory. Lee Harvey Oswald was alive when he fired his rifle in Dallas; but this does not play an explanatory role in the assassination of Kennedy. Crudely speaking, we want to know which causal factors were *salient* and which factors made a difference in the outcome. Michael Strevens (2008) provides an innovative explication of this set of intuitions through the idea of "Kairetic" explanation, a formal way of identifying salient causal factors out of a haystack of causally involved factors in the occurrence of an event guided by generality, cohesion, and accuracy. "To this end, I formulate a recipe that extracts from any detailed description of a causal process a higher level, abstract description that specifies only difference-making properties of the process" (Strevens 2008: xiii).

2.4 Pluralism About Causal Inquiry

This volume is concerned with the problem of causal inquiry and methods for the discovery of causal relations among factors. How can social researchers identify causal relations among social events and structures? The problem of causal inference is fundamental to methodology in the social and policy sciences. A well-informed and balanced handbook of political science methodology is provided by Box-Steffensmeier et al. (2008). Here I will provide a brief discussion of several approaches to causal inferences in the social sciences that follows the typology offered there. Especially relevant is Henry Brady's contribution to the volume (Brady, 2008).

In their introduction to the volume, Box-Steffensmeier, Brady, and Collier propose that there are three important kinds of questions to answer when we are investigating the idea of causal relations in the social world. First is semantic: what do we mean by statements such as "A causes B"? Second is ontological: what are the features of the world that we intend to identify when we assert a causal relationship between A and B? And third is epistemological: through what kinds of investigations and processes of inference can we establish the likelihood of a causal assertion about the relationship that exists among two or more features of the social world? The last question brings us to scientific methodology and a variety of techniques of causal inquiry and inference. However, Box-Steffensmeier, Brady, and Collier are correct in asserting the prior importance of the other two families of questions. We cannot design a methodology of inquiry without having a reasonably well-developed idea of what it is that we are searching for, and that means we must provide reasonable answers to the semantic and ontological questions about causation first. The editors also make a point that is central to the current chapter as well, in favor of a pluralism of approaches to the task of causal inquiry in the social sciences (2008: 29). There is no uniquely best approach to causal inquiry in the social and policy sciences. The editors refer explicitly to a range of approaches that can be used to investigate causation in the social world: qualitative and quantitative investigation, small-n or large-n studies, experimental data, detailed historical narratives, and other approaches.

Henry Brady (2008) provides a useful typology of several families of methods of inquiry and inference that have developed within the social sciences and that find a clear place within the semantic and ontological framework of causation that is developed in this chapter. Brady distinguishes among "neo-humean regularity" approaches, counterfactual approaches, manipulation approaches, and mechanism approaches. And he shows how a wide range of common research methods in the social sciences fall within one or the other of these rubrics. Each of these families of approaches derives from a crucial feature of what we mean by a causal relationship: the fact that causes commonly produce their effects, giving rise to observable regularities; the fact that causes act as sufficient and necessary conditions for their effects, giving rise to the possibility of making inferences about counterfactual scenarios; the fact that causes produce or inhibit other events, giving rise to the

possibility of intervening or manipulating a sequence of events; and the fact that causal relations are real and are conveyed by specific (unobservable) sequences of mechanisms leading from cause to effect, giving rise to the importance of attempting to discover the operative mechanisms.

Brady's typology suggests a variety of avenues of causal inquiry that are possible in the social sciences, given the foregoing analysis of social causes. The ideas sketched in previous sections about the ontology of social causation support multiple avenues for discovering causation. Causes produce their effects, causes work through mechanisms, causal relationships should be expected to result in strong associations among events, and causal necessity supports counterfactual reasoning. We can thus design methods of inquiry that take advantage of the various of ontological characteristics of social causation.

First, the primacy of "real underlying causal mechanisms" suggests that direct research aimed at discovery of the social pathways through which a given outcome is produced by the actions of individual actors within given institutional and normative circumstances is likely to be fruitful. Theory formation about the "institutional logics" created by a given institutional setting can be supplemented by direct study of cases to attempt to identify the pathways hypothesized (Thornton et al., 2012). These insights into the ontology of causation provide encouragement for case-based methods of inquiry, including process tracing, comparative studies, and testing of middle-level social theories of mechanisms. This is a set of methodological ideas supporting causal inquiry developed in detail by George and Bennett (2005), Steinmetz (2004, 2007), and Ermakoff (2019).

Second, the logic of necessary and sufficient conditions associated with the concept of a cause implies methods of research based on experimentation and observation. If we hypothesize that X is a necessary condition for the occurrence of Y, we can design a research study that searches for cases in which Y occurs but X does not. Ragin (1987), Mill (1988), and Tarrow (2010) describe the logic of such cases. The logic of necessary and sufficient conditions also supports research designs based on experimental and quasi-experimental methods—research studies in which the researcher attempts to isolate the phenomenon of interest and observes the outcomes with and without the presence of the hypothetical causal factor. Woodward (2003) illustrates the underlying logic of the experimental approach.

John Stuart Mill's methods of similarity and difference (1988) derive from this feature of the logic of causation. If we believe that A_1 & A_2 are jointly sufficient to produce B, we can evaluate this hypothesis by finding a number of cases in which A_1, A_2, and B occur and examine whether there are any cases where A_1 & A_2 are present but B is absent. If there is such a case, then we can conclude that A_1 & A_2 are not sufficient for B. Likewise, if we believe that A_3 is necessary for the occurrence of B, we can collect a number of cases and determine whether there are any instances where B occurs but A_3 is absent. If so, we can conclude that W is not necessary for the occurrence of B.

2.4.1 Case Studies and Process Tracing

Alexander George and Andrew Bennett (2005) argue for the value of a case study method of social research. The core idea is that investigators can learn about the causation of particular events and sequences by examining the events of the case in detail and in comparison with carefully selected alternative examples. Here is how George and Bennett describe the case study method:

> The method and logic of structured, focused comparison is simple and straightforward. The method is "structured" in that the researcher writes general questions that reflect the research objective and that these questions are asked of each case under study to guide and standardize data collection, thereby making systematic comparison and cumulation of the findings of the cases possible. The method is "focused" in that it deals only with certain aspects of the historical cases examined. The requirements for structure and focus apply equally to individual cases since they may later be joined by additional cases. (George & Bennett, 2005: 67)

The case study method is designed to identify causal connections within a domain of social phenomena. How is that to be accomplished? The most important tool that George and Bennett describe is the method of process tracing. "The process-tracing method attempts to identify the intervening causal process—the causal chain and causal mechanism—between an independent variable (or variables) and the outcome of the dependent variable" (206). Process tracing requires the researcher to examine linkages within the details of the case they are studying and then to assess specific hypotheses about how these links might be causally mediated.

2.4.2 Quantitative Research Based on Observational Data

Quantitative studies of large populations are supported by this theory of causation, if properly embedded within a set of hypotheses about causal relations among the data. In his presentation of the logic of "structural equation modeling" (SEM) and causal inference, Judea Pearl (2000, 2021) is entirely explicit in stating that pure statistical analysis of covariation cannot establish causal relationships. In particular, Pearl argues that a causal SEM requires:

> A set A of qualitative causal assumptions, which the investigator is prepared to defend on scientific grounds, and a model MA that encodes these assumptions. (Typically, MA takes the form of a path diagram or a set of structural equations with free parameters. A typical assumption is that certain omitted factors, represented by error terms, are uncorrelated with some variables or among themselves, or that no direct effect exists between a pair of variables.) (Pearl, 2021: 71)

Aage Sørensen takes a similar view and describes the underlying methodological premise of valid quantitative causal research in these terms:

> Understanding the association between observed variables is what most of us believe research is about. However, we rarely worry about the functional form of the relationship.

> The main reason is that we rarely worry about how we get from our ideas about how change is brought about, or the mechanisms of social processes, to empirical observation. In other words, sociologists rarely model mechanisms explicitly. In the few cases where they do model mechanisms, they are labeled mathematical sociologists, not a very large or important specialty in sociology. (Sørensen, 2009: 370)

Purely quantitative studies do not establish causation on their own; but when provided with accompanying hypotheses about the mechanisms through which the putative causal influences obtain, quantitative study can substantially increase our confidence in inferences about causal relationships among factors. Quantitative methods for research on causation advanced significantly through the development of structural equation models (SEMs) and the structural causal model methodology described by Judea Pearl and others (Pearl, 2000; Pearl, 2009, 2021). This approach explicitly endorses the notion that quantitative methods require background assumptions about causal mechanisms: "one cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable" (Pearl, 2009: 99).

2.4.3 Randomized Controlled Trials and Quasi-experimental Research

The method of randomized controlled trials (RCT) is sometimes thought to be the best possible way of establishing causation, whether in biology or medicine or social science. An experiment based on random controlled trials can be described simply. It is hypothesized that:

(H) A causes B in a population of units P.

An experiment testing H is designed by randomly selecting a set of individuals from P into G_{test} (the test group) and randomly assigning a different set of individuals from P into $G_{control}$ (the control group). G_{test} and $G_{control}$ are exposed to A (the treatment) under carefully controlled conditions designed to ensure that the ambient conditions surrounding both tests are approximately the same. The status of each group is then measured with regard to B, and the difference in the value of B between the two groups is said to be the "average treatment effect" (ATE). If the average treatment effect is greater than zero, there is prima facie reason to accept H.

This research methodology is often thought to capture the logical core of experimentation and is sometimes thought to constitute the strongest evidence possible for establishing or refuting a causal relationship between A and B. It is thought to represent a purely observational way of establishing causal relations among factors. This is so because of the random assignment of individuals to the two groups (so potentially causally relevant individual differences are averaged out in each group) and because of the strong efforts to isolate the administration of the test so that each group is exposed to the same unknown factors that may themselves influence the outcome to be measured. As Handley et al. (2018) put the point: "Random

allocation minimizes selection bias and maximizes the likelihood that measured and unmeasured confounding variables are distributed equally, enabling any differences in outcomes between the intervention and control arms to be attributed to the intervention under study" (Handley et al., 2018: 6). The social and policy sciences are often interested in discovering and measuring the causal effects of large social conditions and interventions—"treatments", as they are often called in medicine and policy studies. It might seem plausible, then, that empirical social science should make use of random controlled trials whenever possible, in efforts to discover or validate causal connections.

However, this supposed "gold standard" status of random controlled trials has been seriously challenged in the last several years. Serious methodological and inferential criticisms have been raised of common uses of RCT experiments in the social and behavioral sciences, and philosopher of science Nancy Cartwright has played a key role in advancing these criticisms. Cartwright and Hardie (2012) provided a strong critique of common uses of RCT methodology in areas of public policy, and Cartwright and others have offered convincing arguments to show that inferences about causation based on RCT experiments are substantially more limited and conditional than generally believed.

A pivotal debate among experts in a handful of fields about RCT methodology took place in a special issue of *Social Science and Medicine* in 2018. This volume is an essential reading for anyone interested in causal reasoning. Especially important is Deaton and Cartwright (2018). The essence of their critique is summed up in the abstract: "We argue that the lay public, and sometimes researchers, put too much trust in RCTs over other methods of investigation. Contrary to frequent claims in the applied literature, randomization does not equalize everything other than the treatment in the treatment and control groups, it does not automatically deliver a precise estimate of the average treatment effect (ATE), and it does not relieve us of the need to think about (observed or unobserved) covariates" (Deaton & Cartwright, 2018). Deaton and Cartwright provide an interpretation of RCT methodology that places it within a range of comparably reliable strategies of empirical and theoretical investigation, and they argue that researchers need to choose methods that are suitable to the problems that they study.

One of the key concerns they express has to do with extrapolating and generalizing from RCT studies (Deaton & Cartwright, 2018: 3). A given RCT study is carried out in a specific and limited set of cases, and the question arises whether the effects documented for the intervention in this study can be extrapolated to a broader population. Do the results of a drug study, a policy study, or a behavioral study give a basis for believing that these results will obtain in the larger population? Their general answer is that extrapolation must be done very carefully. "We strongly contest the often-expressed idea that the ATE calculated from an RCT is automatically reliable, that randomization automatically controls for unobservables, or worst of all, that the calculated ATE is true [of the whole population]" (Deaton & Cartwright, 2018: 10).

The general perspective from which Deaton and Cartwright proceed is that empirical research about causal relationships—including

experimentation—requires a broad swath of knowledge about the processes, mechanisms, and causal powers at work in the given domain. Here their view converges philosophically with that offered by Pearl above. This background knowledge is needed in order to interpret the results of empirical research and to assess the degree to which the findings of a specific study can plausibly be extrapolated to other populations.

These methodological and logical concerns about the design and interpretation of experiments based on randomized controlled trials make it clear that it is crucial for social scientists to treat RCT methodology carefully and critically. Deaton and Cartwright agree that RCT experimentation is a valuable component of the toolkit of sociological investigation. But they insist that it is crucial to keep several philosophical points in mind. First, there is no "gold standard" method for research in any field; rather, it is necessary to adapt methods to the nature of the data and causal patterns in a given field. Second, Cartwright (like most philosophers of science) is insistent that empirical research, whether experimental, observational, statistical, or Millian, always requires theoretical inquiry into the underlying mechanisms that can be hypothesized to be at work in the field. Only in the context of a range of theoretical knowledge is it possible to arrive at reasonable interpretations of (and generalizations from) a set of empirical findings.

Many issues of causation in the social and policy sciences cannot be addressed in a controlled laboratory environment. In particular, in many instances, it is impossible to satisfy the condition of random assignment of individuals to control and treatment groups. Much data available for social science and policy research is gathered from government databases (Medicaid, Department of Education, Internal Revenue Service) and was assembled for statistical and descriptive purposes. Hypotheses about the causes of failing schools, ineffective prison reforms, or faulty regulatory systems are not amenable to the strict requirements of randomized controlled trials. However, social and policy scientists have developed practical methods for probing causation in complex social settings using natural experiments, field experiments, and quasi-experiments.

Quasi-experiments, field experiments, and natural experiments are sometimes defined as "randomized controlled trials carried out in a real-world setting" (Teele, 2014: 3). This definition is misleading, because the crucial feature of RCTs is absent in a quasi-experiment: the random assignment of units to control and treatment groups. What quasi-experiments have in common is an effort to replace random assignments of units to control and treatment groups with some other way of stratifying available data that would permit inference about cause and effect. Quasi-experiments involve making use of observational data about similar populations that have been exposed to different and potentially causally relevant circumstances. The researcher then attempts to discover treatment effects based on statistical properties of the two groups. In this volume, Battistin and Bertoni (Chap. 3) describe an ingenious set of constructs to uncover the effects of cheating on educational performance examination scores in Italy, based on what they refer to as "instrumental variables" and "regression discontinuity design." The former is a component of the composition of the control group that can be demonstrated to be random. The

authors show how this randomness can be exploited to discover the magnitude of effects of the non-random components in the composition of the control group. The latter term takes advantage of the fact that some data sets (class size in Italy, for example) are "saw-toothed" with respect to a known variable. The example they use is the government policy in Italy that regulates class size. School populations increase linearly, but government policy establishes the thresholds at which a school is required to create a new class. So class size increases from the minimum to the maximum, then declines sharply, and continues. This fact can be exploited to examine school performance in classes currently near the minimum versus classes currently near the maximum. This approach removes school population size from the selection and therefore succeeds in removing a confounding causal influence, which is exactly what randomization was intended to do.

The reasoning illustrated in Battistin and Bertoni (Chap. 3) is admirable in the authors' effort to squeeze meaningful causal inferences out of a data set that is awash with non-random elements. However, as Battistin and Bertoni plainly demonstrate, it is necessary to be rigorously critical in developing and evaluating these kinds of research designs and inferences. Stanley Lieberson's *Making It Count* (1985) formulates a series of difficult challenges for the logic of quasi-experimental design that continues to serve as a cautionary tale for quantitative social and policy research. Lieberson believes that there are almost always unrecognized forms of selection bias in the makeup of quasi-experimental research designs that potentially invalidates any possible finding. Cartwright and Hardie (2012) extend these critical points by underlining the limitations on generalizability (external validity) that are endemic to experimental reasoning. So selection bias is still a possibility that can interfere with valid causal reasoning in the design of a quasi-experiment.

What conclusions should we draw about experiments and quasi-experiments? What is the status of randomized controlled trials as a way of isolating causal relationships, whether in sociology, medicine, or public policy? The answer is clear: RCT methodology is a legitimate and important tool for sociological research, but it is not fundamentally superior to the many other methods of empirical investigation and inference in use in the social sciences. Methodologies supporting the design and interpretation of quasi-experiments are also subject to important methodological cautions in the social science and policy studies. It is necessary to remain critical and reflective in assessing the assumptions that underlie any social science research design, including randomized controlled trials and sophisticated quasi-experiments.

2.4.4 Generative Models and Simulation Methods

Advances in computational power and software have made simulations of social situations substantially more realistic than in previous decades. An early advance took place in general equilibrium theory, leading to a set of models referred to as "computable general equilibrium models." Instead of using a three-sector model to

illustrate the dynamics of a general equilibrium model of a market economy, it is now feasible to embody assumptions for one hundred or more industries and work out the equilibrium dynamics of this substantially more realistic representation of an economic system using a computable model (Taylor, 1990). Of special interest for political scientists and policy scholars is the increasing sophistication of agent-based models (de Marchi and Page, 2008). Kollman et al. (2003) provide a highly informative overview of the current state of the field in their *Computational Models in Political Economy*. They describe the chief characteristics of an agent-based model in these terms:

The models typically have four characteristics, or methodological primitives: agents are diverse, agents interact with each other in a decentralized manner, agents are boundedly rational and adaptive, and the resulting patterns of outcomes comes often do not settle into equilibria…. The purpose of using computer programs in this second role is to study the aggregate patterns that emerge from the "bottom up" (Kollman et al. 2003: 3).

An often-cited early application of agent-based models was Thomas Schelling's segregation model. Schelling demonstrated that residential segregation was likely to emerge from a landscape in which two populations had tolerant but finite requirements for the ethnic composition of their neighborhoods (Schelling, 1978). A random landscape populated with a mix of the two populations almost always develops into a segregated landscape of the populations after a number of iterations. Agent-based models can be devised to provide convincing "generative" explanations of a range of collective phenomena; and when developed empirically by calibrating the assumptions of the model to current empirical data, their results can result in reasonable predictions about the near-term future of a given social phenomenon (Epstein, 2006).

We can look at ABM simulation techniques as a form of "mechanisms" theory. A given agent-based model is an attempt to work out the dynamics of individual-level actions at the meso- and macro-level; and this kind of result can be interpreted as an empirically grounded account of the mechanisms that give rise to a given kind of social phenomenon. This feature of agent-based model methodology gives researchers yet another tool through which to probe the social world for causal relations among social features.

2.5 Realism and Methodological Pluralism

Let us draw to a close. Here are some chief features of social science research that proceeds in ways consistent with this realist view of causation in the social world:

- Productive social science research makes use of eclectic multiple theories and do not expect a unified social theory that explains everything.
- Realist social scientists are modest in their expectations about social generalizations.

- They look for causal mechanisms as a basis for social explanation.
- They anticipate heterogeneity and plasticity of social entities.
- They are prepared to use eclectic methodologies—quantitative, comparative, case study, ethnographic—to discover the mechanisms and mentalities that underlie social change.
- Causal reasoning requires background theories about causal relationships in the domain under study. These theories are corrigible, but some set of assumptions about "the causal structure of the world" is unavoidable.

Central in these ideas is the value of *methodological pluralism*. The ultimate goal of research in the social and policy sciences is to discover causal relationships and causal mechanisms. We want to know how the social world works and how we might intervene to change outcomes that are socially undesirable. There are a wide range of methods of inquiry and validation that are used in the social sciences: ethnographic methods (interviews and participant observation), case study analysis, comparative case study research, models and simulations of social arrangements of interest, and large-scale statistical studies. The philosophical position of methodological pluralism is the idea that there is a place in social and policy research for all of these tools and more besides. What holds them together is the fact that in each case, our ultimate concern is to discover the causal relationships that appear to hold in the social world and the mechanisms that underlie these relationships.

The central conclusion to be drawn here is that multiple methods of empirical investigation are available, and our research efforts will be most productive when we are able to connect empirical findings with hypotheses about social-causal mechanisms that are both theoretically and observationally supported. And equally importantly, it is crucial for researchers from different methodological traditions to interact with each other so that their underlying assumptions about causation and causal inference can be refined and validated.

Review Questions

1. What is an "actor-centered" approach to social explanation and policy research?
2. What is a social mechanism? Can you give an example or two?
3. Why is the assumption of random assignment of subjects to control and treatment groups so important for the design of an experiment?
4. What is an agent-based model? Why is it useful in trying to discover causes in the social world?
5. What is the difference between "ontology" and "methodology" in the social sciences?

References

Archer, M. S., Decoteau, C., Gorski, P., Little, D., Porpora, D., Rutzou, T., Smith, C., Steinmetz, G., & Vandenberghe, F. (2016). What is critical realism? *Perspectives, 38*(2), 4–9. http://www.asatheory.org/current-newsletter-online/what-is-critical-realism

Bhaskar, R. (1975). *A realist theory of science*. Leeds Books.

Box-Steffensmeier, J. M., Brady, H. E., & Collier, D. (2008). *The Oxford handbook of political methodology. The Oxford handbooks of political science*. Oxford University Press.
Brady, H. (2008). Causation and explanation in social science. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 217–270). Oxford University Press.
Brinton, M. C., & Nee, V. (Eds.). (1998). *New institutionalism in sociology*. Russell Sage Foundation.
Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press.
Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.
Coase, R. H. (1988). *The firm, the market, and the law*. University of Chicago Press.
Coleman, J. S. (1990). *Foundations of social theory*. Harvard University Press.
De Marchi, S., & Page, S. (2008). Agent-based modeling. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 71–94). Oxford University Press.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.
Dupré, J., & Cartwright, N. (1988). Probability and causality: Why Hume and indeterminism Don't mix. *Nous, 22*, 521–536.
Ekström, M. (1992). Causal explanation of social action: The contribution of Max Weber and of critical realism to a generative view of causal explanation in the social sciences. *Acta Sociologica, 35*(2), 107–123.
Ellickson, R. C. (1991). *Order without law: How neighbors settle disputes*. Harvard University Press.
Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press.
Epstein, J. M. (2006). *Generative social science: studies in agent-based computational modeling. Princeton studies in complexity*. Princeton University Press.
Ermakoff, I. (2019). Causality and history: Modes of causal investigation in historical social sciences. *Annual Reviews, 45*, 581–606.
George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. BCSIA Studies in International Security. MIT Press.
Glennan, S., Illari, P., & Weber, E. (2021). Six theses on mechanisms and mechanistic science. *Journal for General Philosophy of Science*. https://doi.org/10.1007/s10838-021-09587-x
Handley, M. A., Lyles, C. R., McCulloch, C., & Cattamanchi, A. (2018). Selecting and improving quasi-experimental designs in effectiveness and implementation research. *Annual Review of Public Health, 39*, 5–25.
Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.
Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity*. Basil Blackwell.
Hedström, P. (2005). *Dissecting the social: On the principles of analytical sociology*. Cambridge University Press.
Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*, 49–67.
Hume, D. (2007). *A treatise of human nature*. Edited by David Fate. Clarendon Press.
Kollman, K., Miller, J. H., & Page, S. E. (2003). *Computational models in political economy*. MIT Press.
Lewis, D. K. (1973). *Counterfactuals*. Harvard University Press.
Lieberson, S. (1985). *Making it count: The improvement of social research and theory*. University of California Press.
Little, D. (1991). *Varieties of social explanation: An introduction to the philosophy of social science*. Westview Press.
Little, D. (1995). Causal explanation in the social sciences. *Southern Journal of Philosophy Supplement, 1995*.
Little, D. (2006). Levels of the social. In S. Turner & M. Risjord (Eds.), *Handbook for philosophy of anthropology and sociology* (Vol. 15, pp. 343–371). Elsevier Publishing.

Little, D. (2011). Causal mechanisms in the social realm. In P. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences*. Oxford University Press.
Little, D. (2014). Actor-centered sociology and the new pragmatism. In J. Zahle & F. Collin (Eds.), *Individualism, holism, explanation and emergence*. Springer.
Little, D. (2016). *New directions in the philosophy of social science*. Rowman & Littlefield Publishers.
Little, D. (2020). Social ontology De-dramatized. *Philosophy of the Social Sciences., 51*(1), 92–105.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Mackie, J. L. (1974). *The cement of the universe; a study of causation*. Clarendon Press.
Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum, 16*(3), 575–593.
Merton, R. K. (1963). On sociological theories of the middle range. In R. K. Merton (Ed.), *Social theory and social structure*. Free Press.
Mill, J. S. (1988). *The logic of the moral sciences*. Open Court.
Mumford, S.. (2009). Causal powers and capacities. In H. Beebee (Ed.), *The Oxford handbook of causation* (Christopher Hitchcock and Peter Charles Menzies). Oxford University Press.
Mumford, S., & Anjum, R. L. (2011). *Getting causes from powers*. Oxford University Press.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146.
Pearl, J. (2021). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). Guilford Press.
Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Sampson, R. (2010). Neighborhood effects, causal mechanisms and the social structure of the city. In P. Demeulenaere (Ed.), *Analytical sociology and social mechanisms*. Cambridge University Press.
Schelling, T. C. (1978). *Micromotives and macrobehavior*. Norton.
Sørensen, A. B. (1998). Theoretical mechanisms and the empirical study of social processes. In P. Hedström & R. Swedberg (Eds.), *Social mechanisms: An analytical approach to social theory*. Cambridge University Press.
Sørensen, A. B. (2009). Statistical models and mechanisms of social processes. In Peter Hedström & Bjorn Wittrock (Eds.), *Frontiers of sociology* (Annals of the international institute of sociology) (vol. 11). Brill.
Steele, C. M. (2011). *Whistling Vivaldi: How stereotypes affect us and what we can do*. W. W. Norton.
Steinmetz, G. (2004). Odious comparisons: Incommensurability, the case study, and 'small N's' in sociology. *Sociological Theory, 22*(3).
Steinmetz, G. (2007). *The Devil's handwriting: Precoloniality and the German colonial state in Qingdao, Samoa, and Southwest Africa*. University of Chicago Press.
Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.
Tarrow, S. (2010). The strategy of paired comparison: Toward a theory of practice. *Comparative Political Studies, 43*(2), 230–259.
Taylor, L. (1990). *Socially relevant policy analysis: Structuralist computable general equilibrium models for the developing world*. Cambridge, Mass.: MIT Press.
Teele, D. (Ed.). (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. New Haven.
Thelen, K. A. (2004). *How institutions evolve: The political economy of skills in Germany, Britain, the United States, and Japan*. Cambridge University Press.

Thornton, P. H., Ocasio, W., & Lounsbury, M. (2012). *The institutional logics perspective: A new approach to culture, structure, and process*. Oxford University Press.
Woodward, J. (2003). *Making things happen: A theory of causal explanation* (Oxford studies in philosophy of science). Oxford University Press.

Suggested Readings

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.
Lieberson, S. (1985). *Making it count: The improvement of social research and theory*. University of California Press.
Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum, 16*(3), 575–593.

Chapter 3
Counterfactuals with Experimental and Quasi-Experimental Variation

Erich Battistin and Marco Bertoni

Abstract Inference about the causal effects of a policy intervention requires knowledge of what would have happened to the outcome of the units affected had the policy not taken place. Since this counterfactual quantity is never observed, the empirical investigation of causal effects must deal with a missing data problem. Random variation in the assignment to the policy offers a solution, under some assumptions. We discuss identification of policy effects when participation to the policy is determined by a lottery (randomized designs), when participation is only partially influenced by a lottery (instrumental variation), and when participation depends on eligibility criteria making a subset of participant and non-participant units as good as randomly assigned to the policy (regression discontinuity designs). We offer guidelines for empirical analysis in each of these settings and provide some applications of the methods proposed to the evaluation of education policies.

Learning Objectives
By studying this chapter, you will:

- Learn to speak the language of potential outcomes and counterfactual impact evaluation.
- Grasp different concepts of validity of a research design.
- Understand why randomization helps to detect causal effects.
- Discover how to exploit natural experiments and discontinuities to learn about causality when proper experiments are not feasible.
- Discuss the credibility of the assumption underlying different empirical strategies.

E. Battistin
University of Maryland, College Park, MD, USA
e-mail: ebattist@umd.edu

M. Bertoni (✉)
University of Padova, Padova, Italy
e-mail: marco.bertoni@unipd.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_3

3.1 Introduction

Do smaller classes yield better school outcomes? To answer this and many similar questions, one needs to compare the outcome in the *status quo* (a large class) to the outcome that would have been observed if the input of interest was set to a different level (a small class). The comparison of students enrolled in small and large classes is always a tempting avenue to answer this causal question. As this comparison involves different students, its validity rests on the assumption that students currently enrolled in small and large classes would have presented the same outcome, on average, had they been exposed to the same number of classmates. This remains an untestable assumption that must be discussed on a case-by-case basis.

The chapter discusses ways to combine policy designs and data to corroborate the validity of this assumption. Sections 3.2 and 3.3 introduce the counterfactual causal analysis talk. They describe the concepts of treatments, potential outcomes and causal effects, and the attributes characterizing the validity of a research design. Section 3.4 is about the beauty and limitations of randomized assignment to "treatment" (e.g., a small class) and paves the way for the discussion in the following sections. Specifically, these sections deal with methods for causal reasoning when randomization is not feasible. Section 3.5 provides an example of instrumental variation in treatment assignment arising from a natural experiment. Section 3.6 is devoted to the closest cousin to randomization, the regression discontinuity design. Section 3.7 offers some concluding remarks.

Our discussion of empirical methods for causal reasoning is far from exhaustive. For example, we do not discuss research designs that exploit longitudinal data and rely on assumptions on pre-treatment outcome trends (e.g., difference-in-differences and synthetic control methods). Similarly, we do not cover matching methods (see Chap. 4 of this volume). In addition, our presentation will mostly focus on the reasoning underlying design-based identification and will only barely touch issues related with estimation. The interested reader can refer to the book by Angrist and Pischke (2008) for a discussion of these topics.

3.2 Causation and Counterfactual Impact Evaluation: The Jargon

It is useful to start by clarifying what we mean by "causes" and "treatment effects." We consider a population of units indexed by i, with $i = 1, \ldots, N$. Although our narrative will often consider individuals as the units of analysis, the same setting extends to other statistical units such as households, villages, schools, or municipalities.

3.2.1 Causes as Manipulable Treatments

In the population we study, some units are exposed to a cause, which is a treatment or intervention that manipulates factors that may affect a certain outcome. For instance, we might be interested in studying whether class size at primary school affects student performance. Class size here is the treatment and performance is the outcome, which is typically measured using standardized tests. In many countries, class size formation depends on grade enrollment so that, across cohorts, the number of students in the class may change because enrollment changes or because a specific policy affects the regulation. We will use the words "cause", "treatment", or "intervention" interchangeably.

The avenue we take here has some limitations, as not all causes worth considering are manipulable in practice (consider, for example, gender, ethnicity, or genetic traits). Moreover, the design-based approach we describe below may be coarse at times and aimed at shedding light on one particular aspect of a more articulated model. For example, empirical evidence on the causal effects of class size on achievement bundles up the possible contribution of multiple channels that may lead to a better learning environment in small classes. The investigation of channels and mechanisms behind the uncovered effects calls for theories and structural models. The most relevant question to consider turns on the quality of the design-based strategy and on our faith to prop up a more elaborate theoretical framework.

We focus only on binary treatments, that is, we assume that treatment status is described by a binary random variable D_i taking value one if unit i is exposed to treatment ("treated" or "participant") and zero otherwise ("untreated", "non-participant", or "control"). In the class size example, this amounts to considering a setting in which students can be enrolled in small or large classes. The extension to the case of multi-valued or continuous treatment (for example, the number of classmates) is logically identical but requires a more cumbersome notation. More in general, the binary case is always worth of consideration even in a more general context as it helps understand the main challenges in the quest for detecting causal effects. A related issue concerns public policies that are designed as "bundles" of multiple components. In those cases, policy-makers are often interested in disentangling the effect of every component of the policy. We abstract from this problem in our discussion, but emphasize here that the ability to address this question will depend, in general, on the exposure of subjects to different components.

We must take a stand on the reasons why different units end up having a value of D_i equal to one or zero. This is the so-called "assignment rule" and is at the core of any evaluation study. Assignment to treatment can be totally random. In our class size example, this happens when students are randomized to a small or a large class with equal probability and independently of socio-economic background or past performance. When randomization is not at work, participation to treatment is most likely the result of choices made by the units themselves, administrators of the program, or policy makers. For example, parents can choose to enroll their children in schools with smaller classes in the hope of a better learning environment. Finally,

participation to treatment may depend on admission rules that units must comply with. The case of class size formation based on total enrollment is a good example, as the chance of being enrolled in a small class depends on a school's yearly total recruitment. As we shall see, our ability to assess causal effects grows with knowledge of the assignment rule.

3.2.2 *Effects as Differences Between Factual and Counterfactual Outcomes*

It is essential to set the stage for a transparent definition of the treatment effect. To do so, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes experienced if unit i is treated ($D_i = 1$) or untreated ($D_i = 0$), respectively. The unit-level treatment effect of D_i on Y_i is the difference between $Y_i(1)$ and $Y_i(0)$: $\Delta_i = Y_i(1) - Y_i(0)$. Decades of empirical studies using micro-data analyses have taught us that treatment effects most likely vary across units or groups of units with very similar demographics. The notation employed here accommodates for this possibility (the manuals by Angrist & Pischke, 2008, and Imbens & Rubin, 2015, use the same approach).

The definition of Δ_i unveils the fundamental problem that we face when we want to estimate this quantity from the data. While the two potential outcomes can be logically defined for each unit, they can never be observed simultaneously for the same unit. This is true regardless of the assignment rule and the richness or sample size of data we will ever work with. Specifically, the data can reveal only $Y_i(1)$ for units with $D_i = 1$ and $Y_i(0)$ for units with $D_i = 0$. We can, therefore, express the observed outcome Y_i as follows: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) = Y_i(0) + D_i(Y_i(1) - Y_i(0))$. As simple as this can be, lack of observability of both potential outcomes implies lack of observability of the unit-level effect Δ_i. We can think of the unit-level causal effect as the difference between an observed (factual) and an unobserved (counterfactual) potential outcome. Factual quantities are those that can be computed from the data. Counterfactual quantities can be logically defined but can never be computed from data. For treated units, we observe $Y_i = Y_i(1)$ and $Y_i(0)$ is the counterfactual. The opposite is true for control units, for whom we observe $Y_i = Y_i(0)$ and $Y_i(1)$ is the counterfactual.

One way to get around this limitation is to settle for less than unit-level effects. We might be interested in considering average treatment effects for the population or only for some sub-groups. For instance, we define the average treatment effect (ATE) as the average of the individual-level treatment effect in the whole population: $ATE = E(Y_i(1) - Y_i(0))$. This parameter reflects our expectation of what would happen if we were to expose to treatment a randomly chosen unit from the population. Alternatively, we can consider the average treatment effect for the treated (ATT), which describes our expectation for units who have been exposed to treatment: $ATT = E(Y_i(1) - Y_i(0) | D_i = 1)$. Analogously, the average treatment effect for

the non-treated (ATNT) is informative about what would have happened to the untreated if they had been exposed to the intervention:

$$ATNT = E\left(Y_i\left(1\right) - Y_i\left(0\right) \mid D_i = 0\right).$$

Whether any of the above causal parameters can be retrieved from the data will have to be discussed on a case-by-case basis our understanding of the assignment rule plays a key role in this discussion.

3.2.3 What the Data Tell (And When)

Our journey to learn about treatment effects begins by comparing features of the observed outcome Y_i for treated and control units. For instance, the data reveal the average outcomes for treated units, $E(Y_i|D_i = 1)$, and control units, $E(Y_i|D_i = 0)$. Recalling the definition of potential outcomes, the naïve comparison of average outcomes by treatment group, $E(Y_i|D_i = 1) - E(Y_i|\ D_i = 0) = E(Y_i(1)|D_i = 1) - E(Y_i(0)|\ D_i = 0)$, conveys the correlation between the treatment, D_i, and the outcome, Y_i.

The causal interpretation of such naïve comparison is controversial in most cases. To see why, we can add and subtract from the right-hand side of the previous equation the quantity $E(Y_i(0)|D_i = 1)$. This is a counterfactual quantity, as the outcome $Y_i(0)$ cannot be observed for treated units, and represents what would have happened to treated units had they not participated to treatment. We can arrange the terms and write:

$$E\left(Y_i \mid D_i = 1\right) - E\left(Y_i \mid D_i = 0\right) = E\left(Y_i\left(1\right) - Y_i\left(0\right) \mid D_i = 1\right) + E\left(Y_i\left(0\right) \mid D_i = 1\right) - E\left(Y_i\left(0\right) \mid D_i = 0\right). \tag{3.1}$$

It follows that the naïve comparison on the left-hand side of Eq. 3.1 is equal to the sum of the ATT and the term $E(Y_i(0)|\ D_i = 1) - E(Y_i(0)|\ D_i = 0)$, which is often called "selection bias". It is worth noting that this representation does not hinge on any assumptions. It is the result of a simple algebraic trick and, as such, is always true.

Selection bias is an error in the causal reasoning. It is different from zero when, in the absence of treatment, the group with $D_i = 1$ would have performed differently from the group with $D_i = 0$. The same concept is conveyed by the "correlation is not causation" *motto*: correlation (the naïve treatment–control comparison) has no causal interpretation (that is, it does not coincide with the ATT) unless the selection bias is zero. This reframes the quest for causal effects as a discussion on the existence of selection bias. A non-zero bias follows from having groups defined by $D_i = 1$ and $D_i = 0$ that are not representative of the same population, in the sense that participation to treatment depends on non-random selection. At the end of the day, selection bias reflects compositional differences between treatment and control

units. Taking up our class size example, parents with a strong preference for smaller classes are most likely selected in terms of socio-economic background and demographics. If this selection translates into a better learning potential of their children, forming classes as a reflection of parental preference must create dis-homogenous groups of students. In this case, detecting a correlation between class size and achievement might just reveal dis-homogeneity across classes rather than a true causal effect of class size.

Importantly, for the time being, we are agnostic about whether this dis-homogeneity concerns characteristics of units that are observed in the data at hand or not. In fact, any strategy that can adjust for compositional differences between treated and control units also corrects for this bias. One leading example to consider here is randomization. When classes are formed by a coin toss, composition is the same. Even when it is because of sampling variability, differences in composition must be as good as random. We will formalize this idea in Sect. 3.4, below. Instead, Chapters 4 and 5 in this volume present methods to alleviate imbalances along observable dimensions and discuss the identifying assumptions that permit to reach causal conclusions once these differences are eliminated.

3.3 Shades of Validity

The assessment of a causal channel from treatment to the outcome depends on the properties of the research design. In short, this is the toolbox of empirical methods that allows one to distinguish between correlation and causality. Any strategy falling short on this minimum requirement is not a valid option to consider for a good researcher. On the other hand, a good research design must be able to detect precisely the causal relationship of interest. That is, you do not want your design to be underpowered for the size of the treatment effect. Finally, the ideal research design should be able to provide causal statements that apply to the largest share of units in the population and extend to other contexts and times. The concern here is one of generalizability, which is of fundamental importance for offering evidence-based policy recommendations. Causal talk makes use of these three ideas of validity in the development of a research design. This is what we will discuss briefly next. The seminal textbook by Cook and Campbell (1979) provides a deeper treatment of these topics.

3.3.1 Internal Validity: The Ability to Make a Causal Claim from a Pattern Documented in the Data

Internal validity concerns the ability of assessing whether the correlation between treatment and outcome depicts a causal relationship or if it could have been observed even in the absence of the treatment. Therefore, internal validity is solely concerned

with the presence of selection bias. It is achieved under a *ceteris paribus* comparison of units, when all else but the treatment is kept constant between treated and control units. As we discussed above, this calls for the same composition of treatment (small class) and control (large class) units. An internally valid conclusion is the one without selection bias. One of the main advantages of using randomization is that such *ceteris paribus* condition is met by design. Because of this, a properly conducted randomization yields internally valid causal estimates.

3.3.2 Statistical Validity: Measuring Precisely the Relationship Between Causes and Outcomes in the Data

Statistical validity refers to the appropriate use of statistical tools to assess the extent of correlation between treatment and outcomes. It is fundamentally concerned with standard errors and accuracy in assessing a statistical relationship. The main question addressed by statistical validity is whether the chosen data and techniques of statistical inference can produce precise estimates of very small treatment effects (a statistically precise zero) or if, instead, the research design will likely produce statistical zeros (a statistically insignificant effect). An insignificant effect that is statistically different from zero is a powerful oxymoron to summarize the idea underlying statistical validity.

3.3.3 External Validity: The Ability to Extend Conclusions to a Larger Population, over Time and Across Contexts

External validity is about the predictive value of a particular causal estimate for times, places, and units beyond those represented in the study that produced it. The concern posed by external validity is one of generalizability and out-of-sample prediction. For example, an internally valid estimate for a given sub-group of the population might not be informative about the treatment effect for other (potentially different and policy-relevant) sub-groups. Similarly, ATT is, in general, different from ATE. Replicability of the same results in other contexts and times is of fundamental interest for providing policy recommendations.

3.4 Random Assignment Strengthens Internal Validity

As Andrew Leigh puts it in his book "*Randomistas: How Radical Researchers Are Changing the World*," (Leigh, 2018) randomized controlled trials (RCTs) use "the power of chance" to assign the groups. Randomization can be achieved by flipping a coin, drawing the shorter straw, or using a computer to randomly assign statistical

units to groups. In any of these cases, the result would be the same: the treatment and the control group are random samples from the same population.

Random assignment ensures that treatment and control units are the same in every respect, including their expected $Y_i(0)$. It follows that, in RCTs, selection bias must be zero since $E(Y_i(0)|\ D_i = 1) = E(Y_i(0)|\ D_i = 0)$. In other words, what we observe for control units approximates what would have happened to treated units in the absence of treatment. It is worth noting that random assignment does not work by eliminating individual differences, but it rather ensures that the composition of units being compared is the same.

RCTs ensure a *ceteris paribus* (i.e., without confounds) comparison of treatment and control groups. Because of this, an RCT provides an internally valid research design for assessing causality. Evidence in support of this validity can be obtained using pre-intervention measurements. In fact, it is a good practice to collect this information and test the validity of the design by carrying out a battery of "balancing" tests. In a properly implemented randomization, there are no selective differences in the distribution of pre-intervention measurements between treated and control units. This statement does not rule out the possibility of between-group differences arising from sampling variability, which is a problem concerning the statistical validity (that is, the precision of point estimates) of RCTs.

Finally, under random assignment, the naïve comparison will provide internally valid conclusions about the average treatment effect on the treated (ATT), as we have that $E(Y_i|D_i = 1) - E(Y_i|\ D_i = 0) = E(Y_i(1) - Y_i(0)|D_i = 1)$. In addition, under randomization, the groups with $D_i = 1$ and $D_i = 0$ are representative of the same population so that $E(Y_i(1) - Y_i(0)|D_i = 1) = E(Y_i(1) - Y_i(0))$. This means that the causal conclusions hold for any unit randomly selected from the population.

Random assignment to treatment is not uncommon in numerous fields of the social sciences. One such example is the lottery-based allocation of pupils to schools that are oversubscribed. This alternative to the traditional priority criterion based on proximity should dampen school stratification caused by wealthy parents buying houses in the close vicinity of high-quality schools. As a result, among the pool of applicants to a school where oversubscription is resolved by a lottery, getting a seat or not is completely random. Some researchers (see Cullen et al., 2006, for an example) have exploited this to evaluate the educational effects of attending one's preferred school.

Another example is the Oregon Health Insurance Experiment (see Finkelstein et al., 2012). Medicaid is one of the landmark US public health insurance programs and provides care for millions of low-income families. In 2008, the state of Oregon extended coverage of Medicaid by selecting eligible individuals with a lottery. This gave researchers the unique opportunity to provide credible causal estimates of the effect of health insurance eligibility on health care utilization, medical expenditure, medical debt, health status, earnings, and employment.

Although RCTs are considered as the "gold standard" for providing internally valid estimates of causal effects, they are not without shortcomings (see the excellent surveys by Duflo et al., 2008 and Peters et al., 2018). External validity is often perceived as the main limitation and more so for small-scale experiments on very

specific subpopulations. Bates and Glennerster (2017) propose a framework to discuss generalizability based on four steps: identify the theory behind the program; check if local conditions hold for that theory to apply; evaluate the strength of the evidence for the required general behavioral change; evaluate whether the implementation process can be carried out well. External validity is granted if these four conditions apply in a context different from the one where the experiment was conducted. Statistical validity as well may challenge the significance of many small-scale experiments (see Young, 2019).

RCTs have other limitations. Many RCTs are carried out as small-scale pilots that shall be eventually scaled up to the entire population. Causal reasoning in this context must consider the general equilibrium effects arising from this change in scope. These effects are concerned with the possible externalities for non-participants when the policy is implemented on a larger scale and the implications for market equilibria. An additional concern about RCTs is that the sole fact of being "under evaluation" may generate some behavioral response that has nothing to do with a treatment effect.[1] Replicability of experiments also has been called into question in many fields of the social sciences (see Open Science Collaboration, 2015, for psychology and Camerer et al., 2016, for economics).

What happens when randomization is not a feasible option? This is the question to which we turn next.

3.5 Internally Valid Reasoning Without RCTs: Instrumental Variation

3.5.1 A Tale of Pervasive Manipulation

Randomizations obtained by design are not the only way to ensure ceteris paribus comparisons. Randomness in the assignment to treatment may arise indirectly from natural factors or events independently of the causal channel of interest. Under assumptions that we shall discuss, these factors can be used instrumentally to pin down a meaningful casual parameter. The most important takeaway message here is that we must use assumptions to make up for the lack of randomization. Because of this, much of the simplicity of the research design is lost, and internal validity must be addressed on a case-by-case basis. We will present an example of the toolbox for good empirical investigations using administrative data on student achievement and, further below, class size.

Our working example makes use of standardized tests from INVALSI (a government agency charged with educational assessment) for second and fifth graders in Italian schools for the years 2009–2011. Italy is an interesting case study as it is

[1] Such quirky responses are called "Hawthorne" effects for treated subjects and "John Henry" effects for controls.

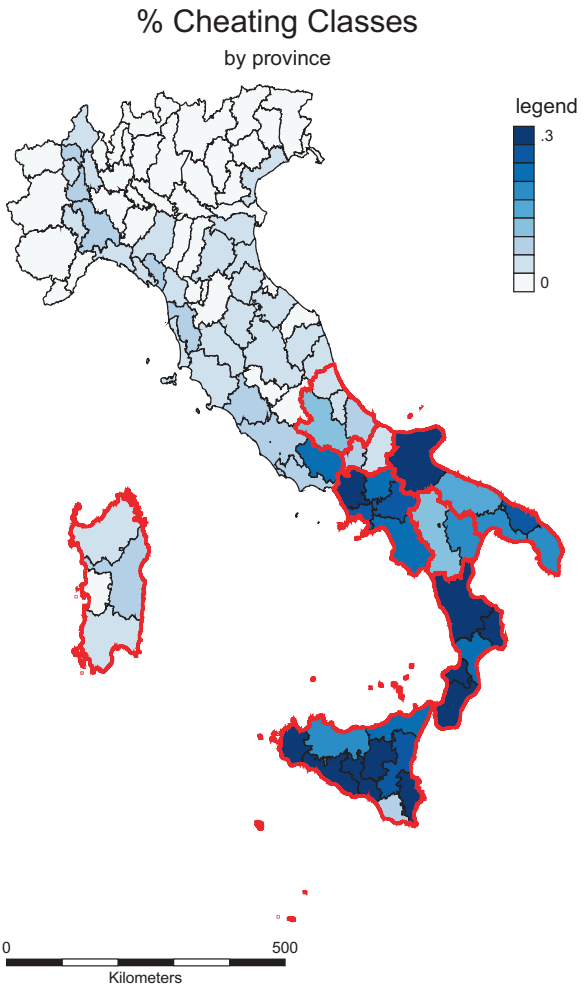

Fig. 3.1 Manipulation by province (Angrist et al., 2017). (Mezzogiorno regions are bordered with dashed lines)

characterized by a sharp North–South divide along many dimensions, among which school quality. This divide motivates public interventions to improve school inputs in the South. As testing regimes have proliferated in the country, so has the temptation to cut corners or cheat at the national exam.[2] As shown in Fig. 3.1, the South is distinguished by widespread manipulation on standardized tests. INVALSI tests are usually proctored and graded by teachers from the same school, and past work by Angrist et al. (2017) has shown that manipulation takes place during the grading process. Classes with manipulated scores are those where teachers did not grade exams honestly.

Consider the causal effect of manipulation on test scores. As scores are inflated, the sign of this effect is obvious. However, the size of the causal effect (that is, by

[2] Cheating or manipulation is not unique to Italy, as discussed in Battistin (2016).

how much scores are inflated) is difficult to measure because manipulation is not the result of random factors. The incentive to manipulate likely decreases as true scores increase so that the distribution of students' true scores is not the same across classes with teachers grading honestly or dishonestly. Again, this is a problem about the composition of the two groups, as treatment classes (with manipulated scores) and control classes (with honest scores) need not be representative of the same population.

When empirical work is carried out using observational data, as it is the case here, it is always illuminating to start from the thought experiment. This is the hypothetical experiment that would be used to measure the causal effect of interest if we had the possibility to randomize units. With observational data, the identification strategy consists of the assumptions that we must make to replicate the experimental ideal. The thought experiment in the case of INVALSI data corresponds to distributing manipulation (the treatment) across classes at random. The identification strategy here amounts to the set of assumptions needed to mimic the very same experimental ideal *even if* manipulation is not random. How can this be possible?

Econometrics combined with the institutional context come to the rescue. It turns out that about 20% of primary schools in Italy are randomly assigned to external monitors, who supervise test administration and the grading of exams from local teachers in selected classes within the school (see Bertoni et al., 2013, and Angrist et al., 2017, for details on the institutional context). Table 3.1 shows that monitors are indeed assigned to schools using a lottery. Schools with monitors are statistically indistinguishable from the others along several dimensions, including average class size and grade enrollment. For example, the table shows that the average class size in unmonitored classes of the country is 19.812 students. The difference between treated and control classes is as small as 0.035 students and statistically indistinguishable from zero. Additional evidence on the lack of imbalance between schools with and without monitors is in Angrist et al. (2017). In the next section, we discuss how to use the monitoring randomization to learn about the effects of manipulation on scores.

3.5.2 General Formulation of the Problem

In our example, the class is the statistical unit of analysis and the treatment is manipulation ($D_i = 1$ if class scores are manipulated and $D_i = 0$ if they are honestly reported). INVALSI has developed a procedure to reveal D_i, so treatment status is observed in the data. Scores (standardized by grade, year, and subject) are the class-level outcome, Y_i. The presence of external monitors is described by a binary random variable Z_i, with $Z_i = 1$ for classes in schools with monitors and $Z_i = 0$ otherwise. In the applied econometrics parlance, variables like Z_i—which is randomly assigned and can influence treatment status—are called "instruments."

The ordinary least squares (OLS) regression of Y_i on D_i summarizes the correlation between manipulation and reported scores. Estimation results obtained from

Table 3.1 Covariate balance in the monitoring experiment (Angrist et al., 2017)

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	19.812	0.0348	20.031	0.0179	19.456	0.0623
	[3.574]	(0.0303)	[3.511]	(0.0374)	[3.646]	(0.0515)
Grade enrollment at school	53.119	−0.4011	49.804	−0.5477	58.483	−0.1410
	[30.663]	(0.3289)	[27.562]	(0.3913)	[34.437]	(0.5909)
% in class sitting the test	0.939	0.0001	0.934	0.0006	0.947	−0.0007
	[0.065]	(0.0005)	[0.066]	(0.0006)	[0.062]	(0.0008)
% in school sitting the test	0.938	−0.0001	0.933	0.0005	0.946	−0.0010
	[0.054]	(0.0005)	[0.055]	(0.0006)	[0.051]	(0.0008)
% in institution sitting the test	0.937	−0.0001	0.932	0.0005	0.945	−0.0010
	[0.045]	(0.0004)	[0.043]	(0.0005)	[0.045]	(0.0007)
N	140,010		87,498		52,512	

Columns 1, 3, and 5 show means and standard deviations for variables listed at the left. Other columns report coefficients from regressions of each variable on a treatment dummy (indicating classroom monitoring), grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Standard deviations for the control group are in square brackets; robust standard errors are in parentheses

[a]$p<0.01$, [b]$p<0.05$, [c]$p<0.1$

OLS are reported in Table 3.2, and a positive correlation between cheating and test score is revealed in all columns. For instance, the value of the coefficient reported in Column (1) of Panel A implies that when we consider data for the whole of Italy, the average math score in classes with manipulated scores is 1.414 standard deviations higher than in classes where teachers did not manipulate scores.[3] However, as discussed above, this result cannot be given any causal interpretation, as the samples with $D_i = 0$ and $D_i = 1$ are non-randomly selected.

Unlike D_i, the status Z_i is randomly assigned. So, it is can be instructive to consider the regression of Y_i on Z_i, summarizing the correlation between manipulation and monitoring. As Z_i is randomly assigned, the latter regression yields the causal effect of monitoring on scores (orthodox empiricists often call this regression the "reduced form equation"). Results in Columns (1)–(3) of Table 3.3 show a negative effect of monitoring on test scores in all columns (see Bertoni et al., 2013). For example, from Column (1) of Panel A, we learn that the average math score in schools with external monitors is 0.112 standard deviations lower than in schools without monitors. Arguably, the negative effect of monitoring on scores passes through a reduction of manipulation.

We need to enrich our causal inference vocabulary to consider potential outcomes based on the 2x2 scenarios that result from the cross-tabulation of D_i and Z_i: $Y_i(D_i, Z_i)$. Similarly, we need to adjust the notation to express the idea that Z_i

[3] Here and in what follows, INVALSI scores are standardized to have zero mean and unit variance by subject and year.

Table 3.2 Correlation between score manipulation and test scored

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Score manipulation	1.414[a]	1.404[a]	1.413[a]
	(0.006)	(0.009)	(0.007)
Means	0.007	−0.074	0.141
(sd)	(0.637)	(0.502)	(0.796)
N	139,996	87,491	52,505
	B. Language		
Score manipulation	1.179[a]	1.085[a]	1.213[a]
	(0.005)	(0.007)	(0.006)
Means	0.01	−0.005	0.035
(sd)	(0.523)	(0.428)	(0.649)
N	140,003	87,493	52,510

All models control for a quadratic polynomial in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and the proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). [a]p<0.01, [b]p<0.05, [c]p<0.1

Table 3.3 Monitoring effects on test scores and score manipulation (Angrist et al., 2017)

	Test scores			Score manipulation		
	Italy	North/Center	South	Italy	North/Center	South
	(1)	(2)	(3)	(4)	(5)	(6)
	A. Math					
Monitor at institution (M_{igkt})	−0.112[a]	−0.075[a]	−0.180[a]	−0.029[a]	−0.010[a]	−0.062[a]
	(0.006)	(0.005)	(0.012)	(0.002)	(0.001)	(0.004)
Means	0.007	−0.074	0.141	0.064	0.02	0.139
(sd)	(0.637)	(0.502)	(0.796)	(0.246)	(0.139)	(0.346)
N	140,010	87,498	52,512	139,996	87,491	52,505
	B. Language					
Monitor at institution (M_{igkt})	−0.081[a]	−0.054[a]	−0.131[a]	−0.025[a]	−0.012[a]	−0.047[a]
	(0.004)	(0.004)	(0.009)	(0.002)	(0.001)	(0.004)
Means	0.01	−0.005	0.035	0.055	0.023	0.11
(sd)	(0.523)	(0.428)	(0.649)	(0.229)	(0.149)	(0.313)
N	140,010	87,498	52,512	140,003	87,493	52,510

Columns 1–3 report the reduced form effects of having a monitor at the institution on test scores. Columns 4–6 show the first-stage estimates of the effect of having a monitor at the institution on score manipulation. All models control for a quadratic polynomial in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). [a]p<0.01, [b]p<0.05, [c]p<0.1

affects D_i. We define potential treatments $D_i(0)$ and $D_i(1)$ as the treatment status that individual i has when exposed to $Z_i = 0$ and $Z_i = 1$, respectively. In our running example, the realized score Y_i corresponds to the potential score realized for the observed combination $\{D_i = d, Z_i = z\}$, while the realized manipulation D_i coincides with the potential manipulation realized for the observed value of $Z_i = z$. For example, $Y_i(1, 1)$ represents the score that would be recorded for class i if teacher grading was dishonest ($D_i = 1$) and the school had an INVALSI monitor ($Z_i = 1$). Recall that, since only selected classes within the school are monitored, dishonest behavior from teachers in unmonitored classes within the school is always possible (see Bertoni et al., 2013).

Depending on the values taken by $D_i(0)$ and $D_i(1)$, we can divide classes into four groups depending on the behavior of teachers grading the exams (see Battistin et al., 2017, for a similar approach):

- Complying dishonest teachers (C), who grade dishonestly without monitors and grade honestly with monitors: $D_i(0)=1$ and $D_i(1) = 0$.
- Always dishonest teachers (A), who always grade dishonestly regardless of the presence of monitors: $D_i(0)=1$ and $D_i(1) = 1$.
- Never dishonest teachers (N), who always grade honestly regardless of the presence of monitors: $D_i(0)=0$ and $D_i(1) = 0$.
- Non-complying dishonest teachers (D), who grade honestly without monitors and grade dishonestly with monitors: $D_i(0)=0$ and $D_i(1) = 1$.

This classification does not hinge on any assumptions and represents the taxonomy of all possible behavioral responses from teachers arising from the monitoring status of the school. The fact that both D_i and Z_i are binary limits to four the number of such responses.

3.5.3 Assumptions

The identification strategy for the analysis of natural experiment builds on four assumptions. We now discuss each of them with reference to our specific running example on the effect of manipulation on test scores. We refer the reader to Angrist and Pischke (2008) for a more general discussion.

3.5.3.1 The "Monotonicity" Assumption

We begin our investigation by assuming lack of non-complying dishonest teachers (D-teachers) in the data. This is a rather innocuous assumption in our context. A violation would represent a quirky behavioral response to the presence of monitors. This assumption is also known as monotonicity condition. It is a restriction on the behavior of units stating that when we move the instrument Z_i from z' to z'', all agents respond by changing their D_i *in the same direction* or by leaving it unaltered. In our

case, this assumption implies that (a) honest teachers without monitors at school would be honest teachers even with a monitor and (b) dishonest teachers without monitors at school might grade honestly under the threat of a monitor at school. In the former case, the value of D_i is unchanged by monitoring and remains zero; in the latter case, the value of D_i may remain one or turn to zero with monitoring. The events (a) and (b) imply that the distribution of the variable D_i must move toward zero in the presence of school monitoring. Ruling out the presence of *D*-teachers implies that monitors cannot change the variable D_i in the opposite direction, from zero to one. This exemplifies why the variable Z_i must induce a monotone (towards zero) behavior for all teachers.

Monotonicity plays a crucial role in natural experiments: under this assumption, we are left with three compliance types—*C*, *A*, and *N*—whose shares in the populations can be represented by π_C, π_A, π_N, respectively. Manipulators are a mixture of always dishonest teachers (*A*-teachers) and complying dishonest teachers (*C*-teachers) without monitors. Honest teachers are composed of never dishonest teachers (*N*-teachers) and complying dishonest teachers (*C*-teachers) with monitors.

3.5.3.2 The "As Good as Random" Assumption

A second key relationship among the variables involved arises because schools are randomly assigned to either $Z_i = 1$ or $Z_i = 0$. Because of the monitoring experiment, the two groups of schools must have the same composition with respect to any variable, including potential outcomes and potential treatment statuses. It, therefore, follows that $\{Y_i(1, 1), Y_i(0, 1), Y_i(1, 0), Y_i(0, 0), D_i(0), D_i(1)\} \perp Z_i$. In our case, this "as good as random" assumption holds by design, because monitors have been explicitly assigned at random to schools.

3.5.3.3 The "Exclusion Restriction"

The causal reasoning builds upon an exclusion restriction. This formalizes the causal construct that the effect of Z_i on Y_i shall be solely because of the effect of Z_i on D_i. In the example considered here, this restriction can be put across considering the following equations:

$$Y_i(0,1) = Y_i(0,0),$$
$$Y_i(1,1) = Y_i(1,0).$$

Therefore, the exclusion restriction implies that there are only two potential outcomes, indexed against D_i: $Y_i(D_i)$. For example, the first equation implies that scores under honest grading ($D_i = 0$) would be the same irrespective of the presence of monitors. Similarly, the second equation implies that dishonest grading ($D_i = 1$) would yield the same score independently of school monitoring. The latter

condition would be violated if, for example, always dishonest teachers cheated differently in the presence of external monitors at school. This possibility is discussed in Battistin et al. (2017) and is ruled out in the case of INVALSI data by results in Angrist et al. (2017).

3.5.3.4 The "First-Stage" Requirement

The assumed causal link from D_i to Z_i can be verified in the data by running an OLS regression of D_i on Z_i. In fact, it is a good practice to verify the size and statistical strength of this "first-stage" regression in any study based on quasi-experimental variation, as the causal chain we have in mind originates from the effect of Z_i on D_i. Should we observe any effect of Z_i on Y_i but no effect of Z_i on D_i, it would be hard to justify that the random variation in Z_i affected Y_i via the ability of Z_i to move D_i. Estimates of the "first-stage" relationship between exposure to monitors and manipulation are reported in Columns (4)–(6) of Table 3.3. As expected, score manipulation is less likely in schools where monitors are present. For example, Column (4) of Panel A indicates that the probability of score manipulation is 2.9 percentage points lower in schools of the country with monitors. This is equivalent to a 36% decrease in the probability of manipulation with respect to the mean in non-monitored schools (equal to 6.4%). As demonstrated by the estimates in Columns (5) and (6) of Table 3.3, this decrease is stronger in Southern Italy than in the North and Center of the country and strongly statistically significant.

3.5.4 Better LATE than Never

To nail down the causal effect of manipulation on scores, we proceed by comparing the expected value of the product Y_iD_i for schools with and without monitors. This product is equal to Y_i for units with $D_i = 1$ and to 0 for units with $D_i = 0$. Given all the assumptions made so far, we have that:

$$E\left(Y_iD_i \mid Z_i = 1\right) = \pi_A * E\left(Y_i\left(1\right) \mid A\right),$$

$$E\left(Y_iD_i \mid Z_i = 0\right) = \pi_c * E\left(Y_i\left(1\right) \mid C\right) + \pi_A * E\left(Y_i\left(1\right) \mid A\right).$$

In the first equation, neither *C*-teachers or *N*-teachers show up, because for them $D_i = 0$ when $Z_i = 1$ so that $Y_iD_i = 0$.[4] Because of the monotonicity assumptions, there

[4] A consequence of random assignment of Z_i and of the exclusion restriction is that conditional on the compliance types defined above, potential outcomes are independent of Z_i, that is, $\{Y_i(1), Y_i(0)\} \perp Z_i \mid \{D_i(0), D_i(1)\}$. In fact, conditional on a given compliance type, there is a one-to-one mapping between Z_i and D_i, and therefore, knowledge of Z_i implies knowledge of D_i.

are no *D*-type teachers either. Therefore, the only group left is that of *A*-teachers, whose fraction in the population is π_A and for whom we always observe $Y_i(1)$. In a similar fashion, we do not see *N*-teachers in the second line, as for them, $D_i = 0$ when $Z_i = 0$. Consequently, after ruling out the presence of *D*-teachers by monotonicity, only *C*- and *A*-teachers show up in this equation. *C*-teachers account for a fraction π_C of the population, and for them, we observe $Y_i(1)$ as in this case $Z_i = 0$, and therefore, $D_i = 1$.

For these very same reasons, if we compare the share of manipulators in schools with and without external monitors, we obtain:

$$E\left(D_i \mid Z_i = 1\right) = \pi_A,$$

$$E\left(D_i \mid Z_i = 0\right) = \pi_c + \pi_A.$$

The former expression suggests that only *A*-teachers have $D_i = 1$ when $Z_i = 1$; the latter that are both *C*- and *A*-teachers have $D_i = 1$ when $Z_i = 0$. Analogous expressions can be derived for $E(Y_i(0)|\, C)$, $E(Y_i(0)|\, N)$ and for π_N if one substitutes D_i with $(1 - D_i)$ in the above. We have that:

$$E\left(Y_i\left(1-D_i\right) \mid Z_i = 1\right) = \pi_C * E\left(Y_i\left(0\right) \mid C\right) + \pi_N * E\left(Y_i\left(0\right) \mid N\right),$$

$$E\left(Y_i\left(1-D_i\right) \mid Z_i = 0\right) = \pi_N * E\left(Y_i\left(0\right) \mid N\right),$$

$$E\left(\left(1-D_i\right) \mid Z_i = 1\right) = \pi_C + \pi_N,$$

$$E\left(\left(1-D_i\right) \mid Z_i = 0\right) = \pi_N.$$

In the first and third equation, *A*-teachers do not show up because they always have $D_i = 1$ so that $Y_i(1 - D_i) = 0$ and $(1 - D_i) = 0$.[5] Because of the monotonicity assumptions, there are no *D*-type teachers either. Therefore, only *C*- and *N*-teachers are left. *C*-teachers account for a fraction π_C of the population. Since in this case $Z_i = 1$, for them, we observe $D_i = 0$ and, therefore, $Y_i(0)$. *N*-teachers are a share π_N of the population, as for them, D_i is always equal to 0, and we observe $Y_i(0)$.

Similarly, in the second and fourth line, we do not see *A*- and *C*-teachers, as for them $D_i = 1$ when $Z_i = 0$. Consequently, after ruling out the presence of *D*-teachers by monotonicity, only *N*-teachers are left.

[5] A consequence of random assignment of Z_i and of the exclusion restriction is that conditional on the compliance types defined above, potential outcomes are independent of Z_i, that is, $\{Y_i(1), Y_i(0)\} \perp Z_i \mid \{D_i(0), D_i(1)\}$. In fact, conditional on a given compliance type, there is a one-to-one mapping between Z_i and D_i, and therefore, knowledge of Z_i implies knowledge of D_i.

By rearranging the equations above, it is easy to obtain:

$$E\left(Y_i\left(1\right)\mid C\right)=\frac{E\left(Y_iD_i\mid Z_i=1\right)-E\left(Y_iD_i\mid Z_i=0\right)}{E\left(D_i\mid Z_i=1\right)-E\left(D_i\mid Z_i=0\right)}, \tag{3.2}$$

and

$$E\left(Y_i\left(0\right)\mid C\right)=\frac{E\left(Y_i\left(1-D_i\right)\mid Z_i=1\right)-E\left(Y_i\left(1-D_i\right)\mid Z_i=0\right)}{E\left(\left(1-D_i\right)\mid Z_i=1\right)-E\left(\left(1-D_i\right)\mid Z_i=0\right)}. \tag{3.3}$$

The difference between the last two expressions yields:

$$E\left(Y_i\left(1\right)-Y_i\left(0\right)\mid C\right)=\frac{E\left(Y_i\mid Z_i=1\right)-E\left(Y_i\mid Z_i=0\right)}{E\left(D_i\mid Z_i=1\right)-E\left(D_i\mid Z_i=0\right)}, \tag{3.4}$$

which represents the average causal effect of manipulation for classes with teachers who graded honestly because of school monitoring (that is, classes with *C*-teachers). Intuitively, this happens because—in the absence of *D*-teachers—this is the only group of teachers for whom the presence/absence of monitors generates variation in manipulation. Borrowing the definition by Angrist and Imbens (1994), the parameter on the left-hand side of (3.4) is the *local average treatment effect* (LATE). The word "local" here is motivated by causal conclusions only licensed for a subset of classes in the population.

Importantly, the expression on the right-hand side of Eq. 3.4 involves only the variables observed so that the causal parameter can be identified from the data. Standard econometric results imply that LATE is estimated by the coefficient on D_i in a two-stage least squares (TSLS) regression of Y_i on D_i, using Z_i to instrument for D_i.[6] Table 3.4 reports the estimates of the LATE parameter in our running example and reveals that manipulation causally increased scores of students assigned to complying dishonest teachers. For example, Column (1) of Panel (A) tells us that score manipulation increases math results in classes with *C*-teachers by 3.827 standard deviations. This causal effect is much larger than the naïve comparison of scores by treatment status reported in Column (1) of Panel A in Table 3.2. Why is it the case? As illustrated in Sect. 3.2.3, the naïve comparison is equal to a causal effect plus selection bias. In this case, selection bias corresponds with the difference in average score of manipulators and non-manipulators if manipulation was not possible at all. As we have argued, manipulation is less likely to occur in classes with higher average true scores. So, selection bias is likely to be negative, that is, $E(Y_i(0)|\ D_i = 1) < E(Y_i(0)|\ D_i = 0)$.

[6] A similar result applies to the expressions in (3.2) and (3.3) when TSLS regressions of Y_iD_i on D_i and of $Y_i(1 - D_i)$ on $(1 - D_i)$, respectively, are considered.

Table 3.4 Local average treatment effect of score manipulation on test scores

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Score manipulation (D_{igkt})	3.827[a]	7.393[a]	2.886[a]
	(0.188)	(0.804)	(0.158)
Means	0.007	−0.074	0.141
(sd)	(0.637)	(0.502)	(0.796)
N	139,996	87,491	52,505
	B. Language		
Score manipulation (D_{igkt})	3.279[a]	4.523[a]	2.786[a]
	(0.180)	(0.456)	(0.178)
Means	0.01	−0.005	0.035
(sd)	(0.523)	(0.428)	(0.649)
N	140,003	87,493	52,510

All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). [a]p<0.01, [b]p<0.05, [c]p<0.1

3.5.5 *External Validity of Causal Conclusions*

Causal conclusions can be drawn for classes with exams graded by *C*-teachers, and TSLS yield internally valid estimates of $E(Y_i(1) - Y_i(0)|\, C)$. However, we have that $E(Y_i(1) - Y_i(0)|\, C) \neq E(Y_i(1) - Y_i(0))$ in general. It follows that that the ability to extend causal conclusions to all classes—that is, the external validity of $E(Y_i(1) - Y_i(0)|\, C)$—is precluded in general. Using the expressions derived in the previous section, we can write:

$$\pi_C = E\left(D_i \mid Z_i = 0\right) - E\left(D_i \mid Z_i = 1\right), \tag{3.5}$$

so that the data is informative about the size of the population for whom this design can provide evidence about a causal effect. This is already a starting point to understand the extent of the external validity problem of causal estimates obtained by LATE. In the case of INVALSI data, the value of π_C is equal to 2.9% for math and 2.5% for language. This can be seen from Column (4) of Table 3.3, which reports the coefficient of Z_i in the first-stage regression of $D_i$on Z_i using data for all classes

in the country. This is equal to the opposite of π_C.[7] In the South, the share of *C*-teachers grows to 6.2% for math and 4.7% for language, as can be seen from Column (6) of the same table.

In our example, the size of the compliant subpopulation is relatively small. How could one extend the conclusions drawn for a possibly small share of complying dishonest teachers to the remaining classes in the population? We follow Angrist (2004) and note that the data provide information about $E(Y_i(1)|\, A)$ and $E(Y_i(0)|\, \mathrm{N})$ as well. These quantities can be obtained using expressions like those we derived above (see Battistin et al., 2017, for details). For example, we have that:

$$E\left(Y_i\left(1\right)|\, A\right) = E\left(Y_i\,|\, D_i = 1, Z_i = 1\right),$$

$$E\left(Y_i\left(0\right)|\, N\right) = E\left(Y_i\,|\, D_i = 0, Z_i = 0\right).$$

The first equality holds because—in the absence of *D*-teachers—only *A*-teachers manipulate scores in the presence of monitors. Similarly, only *N*-teachers report honestly without monitors.

If potential outcomes were homogeneous across types in the population, then we would have that $E(Y_i(1)|\, A) = E(Y_i(1)|\, C)$ and $E(Y_i(0)|\, N) = E(Y_i(0)|\, C)$. If these two equalities cannot be rejected from the data, we would feel more confident about extending the results obtained for classes with complying dishonest teachers to other classes in the population.[8]

In Table 3.5, we report the comparison of $E(Y_i(1)|\, C)$ vis-à-vis $E(Y_i(1)|\, A)$ and $E(Y_i(0)|\, C)$ vis-à-vis $E(Y_i(0)|\, N)$ for Southern Italy, where the problem of manipulation is more pervasive. While the data does not reject that $E(Y_i(1)|\, C)$ is equal to $E(Y_i(1)|\, A)$, the empirical evidence suggests that $E(Y_i(0)|\, C)$ is much smaller than $E(Y_i(0)|\, N)$. For instance, as reported in Panel A of Table 3.5, for math, we have that $E(Y_i(1)|\, C)$ and $E(Y_i(1)|\, A)$ are very similar and, respectively, equal to 1.426 and 1.236 standard deviations. On the other hand, while $E(Y_i(0)|\, C)$ is equal to −1.662 standard deviations, $E(Y_i(0)|\, N)$ is much higher and equal to −0.655 standard deviations. Therefore, in this case, the data advise against the generalization of the LATE of manipulation on scores outside of the population of complying dishonest teachers.

[7] The number reported in the table is the estimate of π_C with its sign flipped. This is because the expression for share of C − teachers π_C is in (5).The coefficient on Z_i in the regression of D_i on Z_i identifies instead $E(D_i|\, Z_i = 1) - E(D_i|\, Z_i = 0)$, that is, the opposite of π_C.

[8] Needless to say, full homogeneity of potential outcomes across types requires also that $E(Y_i(1)|\, N) = E(Y_i(1)|\, C)$ and $E(Y_i(0)|\, A) = E(Y_i(0)|\, C)$. However, the data will never reveal $E(Y_i(1)|\, \mathrm{N})$ and $E(Y_i(0)|\, A)$, as we never get to observe $D_i = 1$ for N-teachers and $D_i = 0$ for A-teachers. Hence, the latter two conditions cannot be tested empirically.

Table 3.5 Average potential outcomes by type: South of Italy

	Test scores		
	Complying dishonest (C)	Always dishonest (A)	Never dishonest (N)
	(1)	(2)	(3)
	A. Math		
$E(Y_i(1))$	1.426[a]	1.236[a]	
	(0.020)	(0.119)	
$E(Y_i(0))$	−1.453[a]		−0.527[a]
	(0.157)		(0.104)
N	52,505	52,505	52,505
	B. Language		
$E(Y_i(1))$	1.147[b]	1.029[a]	
	(0.018)	(0.103)	
$E(Y_i(0))$	-1.662[a]		−0.655[a]
	(0.176)		(0.084)
N	52,510	52,510	52,510

$E(Y_i(1)|\ C)$ and $E(Y_i(0)|\ C)$ are obtained from 2SLS regressions as detailed in the text. $E(Y_i(1)|\ A)$ and $E(Y_i(0)|\ N)$ are computed from OLS regressions that estimate $E(Y_i|\ D_i = 1, Z_i = 1\ )$ and $E(Y_i|\ D_i = 0, Z_i = 0)$, respectively. All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). [a]p<0.01, [b]p<0.05, [c]p<0.1

3.6 Causal Reasoning with Administrative Rules: The Case of Regression Discontinuity Designs

3.6.1 Larger Classes, Worse Outcomes?

The benefits of reducing student–teacher ratios on learning, educational achievement, and eventually long-term labor market outcomes have been of long-standing concern to parents, teachers, and policy-makers. Observational studies often show a negative relationship between class size and student achievement. Yet the conclusions of such studies might be subject to the problem of self-sorting of students into smaller classes.

In many countries, class size formation depends on grade enrollment using a deterministic rule, and Italy is no exception. As discussed in Angrist et al. (2017), until 2008, class size in primary schools in Italy must be between 10 and 25. A reform in 2009 modified these limits to 15 and 27, respectively. Class formation is regulated by law, and grade enrollment above multiples of the cap to maximum size leads to the formation of a new class. To see this, consider the cap at 25 students in place until 2008. Schools enrolling up to 25 students must form one class. One additional student enrolled after 25 would force principals to form one additional

class, with an average class size of 13 students. The same idea extends to any multiple of 25 students. For example, crossing the 50-student limit is enough to form three classes instead of two and so forth. Because of the regulation in place, class size decreases sharply when enrollment moves from just below to just above multiples of 25. Angrist and Lavy (1999) called this relationship "Maimonides' rule" after the medieval scholar and sage Moses Maimonides who commented on a similar rule in the Talmud.[9] Exceptions to the rule in Italy are allowed in some cases. For example, a 10% deviation from the maximum (3 students) in either direction is possible at the discretion of school principals and upon the approval from the Ministry of Education. The presence of students with disabilities or special education needs is often advocated to justify non-compliance with the law. Moreover, principals can form classes smaller than 10 students in the most remote areas of the country.

By allowing actual class size to deviate from the class size mandated by law, these exceptions generate fuzziness in the relationship between actual and predicted class size. This can be seen in Fig. 3.2, where we report the average class size in the country by grade enrollment at school for second graders before 2008.[10] The sawtooth-shaped solid line reports predicted class size as a function of enrollment, the Maimonides' rule, while the dots report average actual class size by enrollment. The law predicts class size to be a non-linear and discontinuous function of enrollment. Actual class size follows predicted class size closely and more so for schools enrolling less than 75 students (which is the majority of schools in the country). In addition, discontinuities in the actual class size/enrollment relationship show up at multiples of 25 enrolled students. Given the soft nature of the rule, however, they are weaker than the sharp ones observed for predicted class size.

3.6.2 Visual Interpretation

Figure 3.3 offers a visual representation of the size of these discontinuities and is constructed using classes at schools with enrollment that falls in a [−12,12] window around the first four cutoffs shown in Fig. 3.2. Enrollment values in each window are centered to be zero at the relevant cutoff. The y-axis shows average class size conditional on the centered enrollment value shown on the x-axis. The figure also plots fitted values generated by *locally linear regression* (LLR) fits to class-level

[9] More precisely, let f_{igkt} be the predicted class size of class i in grade g at school k in year t. We have that $f_{igkt} = \frac{r_{gkt}}{\left[int\left(\left(r_{gkt}-1\right)/c_{gt}\right)+1\right]}$, where r_{gkt} is beginning-of-the-year grade enrollment at school k, $c_{gt}$is the relevant cap (25 or 27) for grade g, and $int(x)$ is the largest integer smaller than or equal to x.

[10] Similar patterns hold also for the period after the 2008 reform and for fifth graders, as shown by Angrist et al. (2017).

Fig. 3.2 Class size by enrollment among second-grade students in pre-reform years (Angrist et al., 2017). (It shows actual class size and class size as predicted by the Maimonides' rule in pre-reform years for second-grade students)

data, as described in Angrist et al. (2017). This representation is convenient in that one can think that small classes are those in schools with grade enrollment to the right of zero. The figure shows a clear drop at this value. Class size is minimized at about 3–4 students to the right of this value, as we would expect were Maimonides' rule to be tightly enforced.

How can we use these discontinuities in class size to assess a causal effects of class size? School enrollment may be positively correlated with test scores, for example, because larger schools are typically in urban areas, and this relationship need not be linear. However, we would be tempted to infer a causal effect of class size on test score if we observed a discontinuous change in test scores at the *exact* values of enrollment that are multiples of the maximum class size caps, where class size also discontinuously changes. This is the idea underlying the evaluation design that goes by the name of regression discontinuity (RD).

Figure 3.4 exemplifies this idea. It reports the change in average test scores as normalized enrollment moves from below to above the recentered enrollment cut-offs, separately for North and Central Italy and for the South. There is evidence of a positive discontinuity in scores as we move from below to above the cutoff in Southern Italy. Evidence of jumps for the rest of the country is instead much more limited, suggesting the possibility of causal effects of class size on learning mostly for schools in the South.

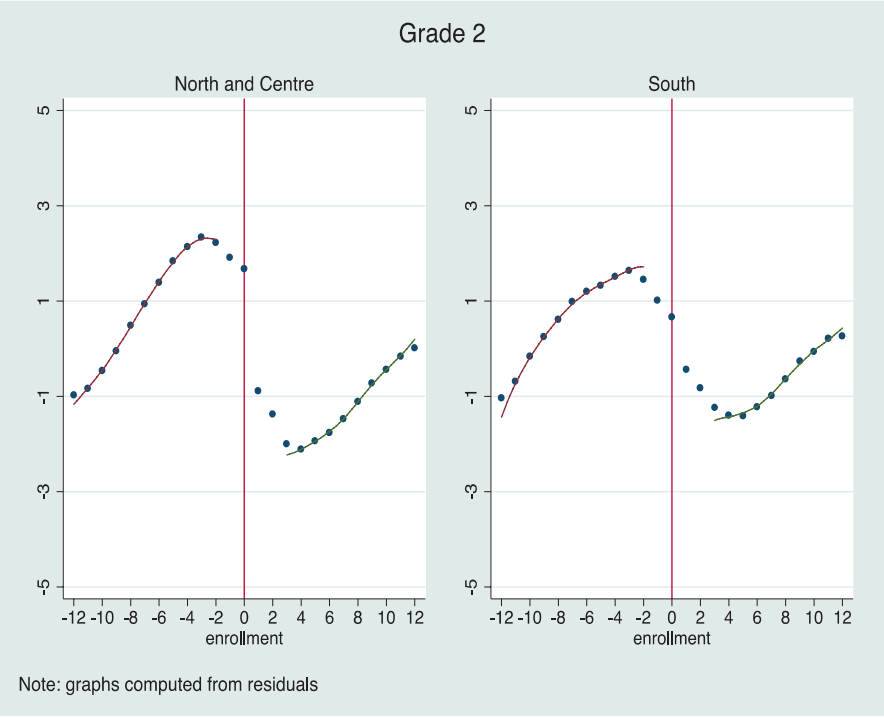

Fig. 3.3 Class size by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017). (Graphs plot residuals from a regression of class size on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit.)

The idea underlying the RD design is that the comparison of scores of classes just above and just below the enrollment cutoffs identified by the Maimonides' rule is informative of effects of class size. Still, not all classes above the cutoffs are small and not all classes below are large, because of discretion in the application of the rule. Intuitively, if compliance with the rule was perfect, then the graphical analysis would already reveal the causal effect. If compliance is not perfect, we may want to use the rule as an instrument for class size formation. Intuitively, the crucial assumption here is that the Maimonides' rule must affect performance at school only because it affects class size formation. A juxtaposition with the identification results discussed in Sect. 3.5 reveals that, in this case, the causal effect of class size on learning is identified only for schools that would form smaller classes because of compliance with the rule. We will come back to this point later in this section.

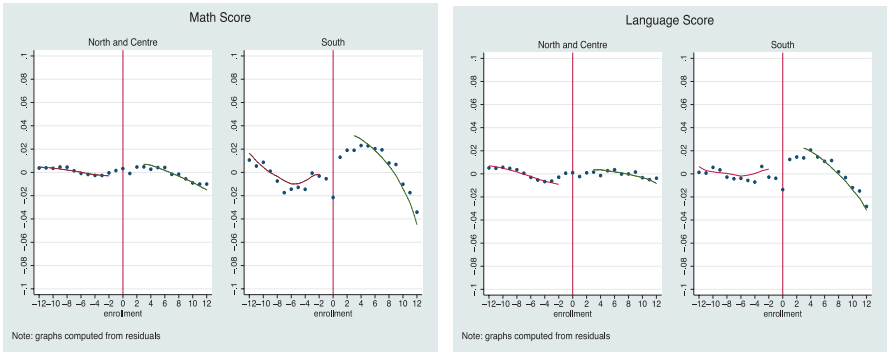

Fig. 3.4 Test scores by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017), (Graphs plot residuals from a regression of test scores on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit.)

3.6.3 General Formulation of the Problem

Following our running example, the class is the statistical unit of analysis and the treatment is class size.[11] To ease the narrative, we distinguish between small and large classes and move to the background the possibility of a "continuous" treatment (number of students in class). Small classes will have $D_i = 1$ and large classes $D_i = 0$. In our narrative, the Maimonides' rule predicts small classes to the right of the recentered cutoffs in Fig. 3.2. Similarly, a large class is predicted for grade enrollment at or below the cutoffs in the same figure. Potential outcomes $Y_i(1)$ and $Y_i(0)$ are the average test score that class i would get if it was small or large. Grade enrollment at school of class i is r_i. Without loss of generality and consistent with Fig. 3.3, we recentered grade enrollment at zero using a [−12,12] window around cutoffs.

3.6.3.1 The Sharp RD Design

We start our discussion by assuming full compliance of school principals with the Maimonides' rule. In other words, we pretend that all classes with r_i at or above zero are small and that all classes with r_i below zero are large. This is equivalent to

[11] We will drop all indexes other than i in what follows. The data contains additional dimensions, but we ignore them for expositional simplicity. One dimension is grade and year. However, scores are standardized by grade and year, so we can ignore them. As a result of this normalization, we end up having repeated measurements over time for classes at the same school. Another dimension is the reform regime. We recenter enrollment to the right cutoff depending on the regulation in place, and we, therefor, abstract from this dimension.

assuming a deterministic relationship between r_i and class size, which we express using the following notation: $D_i = 1(r_i \geq 0)$. We use this *sharp* setting to write the comparison of outcomes for classes in schools with grade enrollment in a neighborhood of the Maimonides' cutoff. The notion of cutoff proximity will be exemplified by using limits from below and above zero. Accordingly, the notation $r_i^+ = 0$ in what follows should read "just above the Maimonides' cutoff"; the notation $r_i^- = 0$ is instead "just below the Maimonides' cutoff."

We have that:

$$\lim_{r\to 0^-} E\left(Y_i \mid r_i = r\right) = \lim_{r\to 0^-} E\left(Y_i\left(0\right) \mid r_i = r\right) + \lim_{r\to 0^-} E\left(D_i\left(Y_i\left(1\right) - Y_i\left(0\right)\right) \mid r_i = r\right)$$
$$= \lim_{r\to 0^-} E(Y_i\left(0\right) \mid r_i = r),$$

because in classes to the left of the Maimonides' cutoff D_i is zero so that the second term vanishes. For classes with r_i above zero, we have:

$$\lim_{r\to 0^+} E\left(Y_i \mid r_i = r\right) = \lim_{r\to 0^+} E\left(Y_i\left(0\right) \mid r_i = r\right) + \lim_{r\to 0^+} E\left(D_i\left(Y_i\left(1\right) - Y_i\left(0\right)\right) \mid r_i = r\right),$$
$$= \lim_{r\to 0^+} E\left(Y_i\left(0\right) \mid r_i = r\right) + \lim_{r\to 0^+} E(Y_i\left(1\right) - Y_i\left(0\right) \mid r_i = r),$$

because D_i is one deterministically. It follows that the outcome difference between small and large classes at the cutoff can be written as:

$$\lim_{r\to 0^+} E\left(Y_i \mid r_i = r\right) - \lim_{r\to 0^-} E\left(Y_i \mid r_i = r\right) = \lim_{r\to 0^+} E\left(Y_i\left(1\right) - Y_i\left(0\right) \mid r_i = r\right)$$
$$+ \lim_{r\to 0^+} E\left(Y_i\left(0\right) \mid r_i = r\right) - \lim_{r\to 0^-} E\left(Y_i\left(0\right) \mid r_i = r\right).$$

The parallel with the naïve comparison discussed in Eq. 3.1 is striking: the comparison of outcomes for small ($r_i^+ = 0$) and large ($r_i^- = 0$) classes is equal to a causal effect for units just to the right of $r_i = 0$:

$$\lim_{r\to 0^+} E(Y_i\left(1\right) - Y_i\left(0\right) \mid r_i = r),$$

plus a selection bias term:

$$\lim_{r\to 0^+} E\left(Y_i\left(0\right) \mid r_i = r\right) - \lim_{r\to 0^-} E(Y_i\left(0\right) \mid r_i = r),$$

measuring differences in a local neighborhood of $r_i = 0$ that would have occurred even without treatment (i.e., if class size could be only large). What conditions are needed to ensure that the latter term is zero? A closer look at the two terms in the last expression reveals an idea of *continuity*. The condition:

$$\lim_{r\to 0^+} E\left(Y_i\left(0\right)\mid r_i = r\right) = \lim_{r\to 0^-} E(Y_i\left(0\right)\mid r_i = r), \tag{3.6}$$

is sufficient to eliminate selection bias and is equivalent to assuming that the relationship between the outcome $Y_i(0)$ and grade enrollment is continuous at $r_i = 0$. This is a mild regularity condition, which most likely holds in most applications, and has a very simple interpretation: our hopes to give any causal interpretation to discontinuities in school performance observed around Maimonides' cutoffs must rest on the assumption that there would have been no discontinuity in performance crossing from $r_i^- = 0$ over to $r_i^+ = 0$ had the Maimonides' rule been irrelevant for forming a small class. Assumption (3.6) combined with its counterpart for the $Y_i(1)$ outcome:

$$\lim_{r\to 0^+} E\left(Y_i\left(1\right)\mid r_i = r\right) = \lim_{r\to 0^-} E(Y_i\left(1\right)\mid r_i = r), \tag{3.7}$$

ensures:

$$\lim_{r\to 0^+} E\left(Y_i\mid r_i = r\right) - \lim_{r\to 0^-} E\left(Y_i\mid r_i = r\right) = E\left(Y_i\left(1\right) - Y_i\left(0\right)\mid r_i = 0\right). \tag{3.8}$$

Assumption (3.7) brings to the problem the same regularity condition in (3.6), with a similar interpretation.

The notion of continuity of potential outcomes around Maimonides' cutoffs is evocative of the properties of a full randomization of students to small and large classes in schools with grade enrollment near $r_i = 0$. For example, assumption (3.6) can be interpreted as an independence condition between $Y_i(0)$ and D_i *locally* with respect to the Maimonides' cutoff. This is the same sort of condition that we discussed in Sect. 3.4 above. It follows that the internal validity of RD estimates obtained from (3.8) hinges upon the assumption that students in schools with values of r_i near zero are as good as randomly assigned to small and large classes, as in a local randomized experiment. In Sect. 3.6.4 below, we discuss how potential violations of such condition may arise in practice and propose some tests to assess the plausibility of this assumption.

Compared to a standard randomized experiment, we pay a price in terms of external validity, as RD estimates are internally valid only around Maimonides' cutoffs. The extrapolation of this effect away from the cutoff requires further assumptions about the global shape of the potential outcome functions, that must be discussed on a case-by-case basis. We refer the interested reader to the work by Battistin and Rettore (2008), Angrist and Rokkanen (2015), Dong and Lewbel (2015), and Bertanha and Imbens (2020).

RD estimates of causal effects are obtained from the sample analogue of the expression in (3.8).[12] The simplest way to proceed is by comparing the mean sample outcomes for small and large classes within a fixed distance from the Maimonides' cutoff $r_i = 0$. The simplicity of this estimator is very appealing, but we may

[12] Lee and Lemieux (2010) provide a thorough discussion of estimation issues in RD designs. We refer the interested reader to their survey for additional details.

encounter statistical validity issues if the data are "sparse" around the Maimonides' cutoff. In fact, we face a trade-off. On the one hand, to enhance statistical validity, we would be tempted to enlarge the width of the neighborhood around the Maimonides' cutoff considered for estimation. On the other hand, by so doing, we would end up using also data points far away from the cutoff. If the relationship between Y_i and r_i was not flat, this could endanger the internal validity of the design.

To minimize this trade-off, researchers often rely on semi-parametric estimators. Kernel-weighted local regressions of the outcome on a low-order (linear or quadratic) polynomial in r_i estimated separately for classes to the left and to the right of r_i are the most common option (as in Fig. 3.4). By giving a larger weight to data point that are closer to the cutoff and allowing for a non-flat relationship between test scores and enrollment, this estimator permits to enlarge sample size while maintaining internal validity. A flexible parametric regression of Y_i and r_i that uses all the available data could also be an option when sample size is small, but this may raise additional issues if high-order polynomials are adopted (see Gelman & Imbens, 2019).

3.6.3.2 The Fuzzy RD Design

When compliance with the Maimonides' rule is far from perfect, as in Italian primary schools, the sharp setting described in the previous section no longer applies. The fuzziness introduced by non-compliance can be dealt with using the class size predicted from the Maimonides' rule as an instrumental variable for the actual class size. The key assumption underlying this approach is that the regulation on class size formation must influence standardized tests only because the regulation affects how classes are eventually formed. This is, once again, an exclusion restriction of the form discussed in Sect. 3.5.3.3, above.

A few refinements of this idea are needed in this setting because the Maimonides' rule yields experimental-like variation only near $r_i = 0$, implying that the "as good as random" condition in Sect. 3.5.3.2 must hold only *locally* with respect to this point. Complying classes here are those turning small because of compliance with the class size regulation when grade enrollment crosses from $r_i^- = 0$ over to $r_i^+ = 0$ (see Sect. 3.5.3.1). Moreover, the first-stage condition, which ensures that the Maimonides' rule shapes—at least in part—the way classes in Italy are eventually formed stems from the following contrast:

$$\lim_{r \to 0^+} E\left(D_i \mid r_i = r\right) - \lim_{r \to 0^-} E\left(D_i \mid r_i = r\right). \tag{3.9}$$

Eq. 3.9 compares the share of small classes just above and just below the Maimonides' cutoff $r_i = 0$. Contrary to the case of a sharp RD, where this contrast is one because of full compliance, fuzziness arising from it makes this quantity lower than one depending on the number of complying classes. The more severe is the extent of non-compliance, the lower will be the external validity of the causal conclusions, as we discussed in Sect. 3.5.5.

The same argument used in Sect. 3.5 extends to the case considered here and can be used to write:

$$E\left[Y_i(1) - Y_i(0) \mid C,\ r = 0\right] = \frac{\lim\limits_{r \to 0^+} E\left(Y_i \mid r = 0\right) - \lim\limits_{r \to 0^-} E\left(Y_i \mid r = 0\right)}{\lim\limits_{r \to 0^+} E\left(D_i \mid r = 0\right) - \lim\limits_{r \to 0^-} E\left(D_i \mid r = 0\right)}. \tag{3.10}$$

The expression in Eq. 3.10 reveals that a causal effect is retrieved by the ratio of the discontinuities in the outcome and in the treatment probability at the Maimonides' cutoff. This expression bears strong similarities with Eq. 3.4 above, once we assign the role played by the instrumental variable to a dummy for being above the Maimonides' cutoff, $Z_i = 1(r_i \geq 0)$. In fact, Hahn et al. (2001) showed that non-compliance leads the fuzzy RD design to be informative about a local average treatment effect, strengthening this similarity. However, the parameter uncovered by the fuzzy RD is local in two senses. First, it refers only to complying classes. Second, it yields causal conclusions only about classes with a value of r_i close to 0, limiting external validity even further.

Following the analogy to the instrumental variable case, discussed in Sect. 3.5, estimation of fuzzy RD effects is usually carried out using two-stage least square (TSLS) methods. The general idea is to instrument the treatment dummy D_i with the dummy $Z_i = 1(r_i \geq 0)$. As in the sharp RD case, researchers can choose to model the relationship between test scores and enrollment using either parsimonious local regressions or flexible global polynomial regressions. In general, and unlike in the sharp RD case, a single TSLS regression is estimated using data on both sides of the cutoff but permitting the polynomial in r_i to have a different shape on each side of the cutoff. This is done by including interaction terms between the polynomial in r_i and D_i that are instrumented by interaction terms between the polynomial in r_i and Z_i.[13]

The estimated fuzzy RD effects of class size on test scores for our running example are reported in Table 3.6 and show a negative and significant effect of class size reduction for compliers at the relevant discontinuity cutoffs. For simplicity, these are obtained using continuous class size. For instance, according to the estimates reported in Column (1) of Panel A, when we consider data for the whole of Italy, we estimate that math scores would increase by an average of 0.06 standard deviations if we decreased class size by 1 unit. As revealed by Columns (2) and (3) and in accordance with Fig. 3.4, the magnitude of such effect is much larger in Southern Italy than in the rest of the country.

[13] Further details about estimation in the fuzzy RD design are discussed in Lee and Lemieux (2010a, b).

Table 3.6 Local average treatment effect of class size on test scores (Angrist et al., 2017)

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Class size	−0.0609[a]	−0.0417[a]	−0.1294[a]
	(0.0196)	(0.0171)	(0.0507)
N	140,010	87,498	52,512
	B. Language		
Class size	−0.0409[a]	−0.0215	−0.0937[b]
	(0.0155)	(0.0136)	(0.0403)
N	140,010	87,498	52,512

The table reports 2SLS estimates using class size cutoffs as an instrument. All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Class size coefficients show the effect of 10 students. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). [a]p<0.01, [b]p<0.05, [c]p<0.1

3.6.4 Validating the Internal Validity of the Design

An underlying assumption behind the approach discussed so far is that units cannot precisely manipulate their value of the running variable. For instance, suppose that parents of pupils with above-average ability could perfectly predict enrollment by school and choose to apply only for schools where enrollment is locally above the relevant cutoffs so that their pupils would systematically end up in smaller classes.[14] If this was the case, then the RD design would be invalid, as the ability composition of pupils in schools where enrollment is just above and just below the cutoff would be different.

In general, if units cannot precisely manipulate their value of the score, there should be no systematic differences between units with similar values of the score. Therefore, a test for the internal validity of an RD design is to verify whether there are discontinuities in these covariates at the cutoff. If predetermined variables that correlate with the outcome are discontinuous at the cutoff, then continuity of potential outcomes is unlikely to hold. These tests are akin to the "balancing" tests presented for the pure randomization case but are carried out locally, at the cutoff.

Table 3.7 reports results for these tests and shows precisely estimated zero effects of passing the RD cutoffs on some predetermined controls, such as the share of students present in class on the day of the test, supporting the validity of this RD design.

[14] For instance, Urquiola and Verhoogen (2009) show evidence of discontinuities between enrollment and household characteristics in Chilean private schools.

Table 3.7 Covariate balance for class size discontinuities (Angrist et al., 2017)

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
% in class sitting the test	0.9392	0.0000	0.9345	0.0001	0.9471	0.0000
	[0.0643]	(0.0001)	[0.0657]	(0.0001)	[0.061]	(0.0001)
% in school sitting the test	0.9386	0.0001	0.9339	0.0001	0.9464	0.0001
	[0.0534]	(0.0001)	[0.0548]	(0.0001)	[0.05]	(0.0001)
% in institution sitting the test	0.9374	−0.0001	0.9327	−0.0001	0.9451	−0.0000
	[0.0436]	(0.0001)	[0.0426]	(0.0001)	[0.0441]	(0.0001)
N	140,010		87,498		52,512	

Columns 1, 3, and 5 show means and standard deviations for variables listed at the left. Other columns report coefficients from regressions of each variable on predicted class size, a quadratic in grade enrollment, segment dummies and their interactions, grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Standard deviations for the control group are in square brackets; robust standard errors are in parentheses. [a]p<0.01, [b]p<0.05, [c]p<0.1

3.7 Conclusion

This chapter has discussed a selected number of approaches among the most popular in the toolbox of good empiricists interested in causal relationships. Randomization, instrumental variation, and discontinuity designs are very closely related members of the same family and, when properly implemented, are thought to yield the most credible estimates of the causal effects of public interventions.

The beauty of randomized assignment is that the composition of "treatment" and "control" groups is by design not driven by any form of selection. In this case, differences in the composition of groups due to sampling variation tend to vanish as sample size increases so that the main concern should be the one of statistical validity. External validity and general equilibrium effects may also be a concern, especially if the intervention has to be implemented in different contexts or scaled up to cover a whole country.

Instrumental variation is a good way to go when randomized assignment is not viable. It seeks sources of random variation that have affected indirectly the chance of receiving "treatment." Clearly, a good source of variability must affect only the treatment assignment and, through this, the outcome of interest. Sources of external random variation affecting at the same time both treatment allocation and the outcome will not allow to distinguish the effect of the instrument on the outcome from the effect of the treatment on the same outcome. As we have made clear, the price to pay for the lack of randomized assignment to treatment is external validity: estimates of causal effects obtained from instrumental variation are limited to the fraction of the population changing the treatment status because of the instrument. How large and comparable this fraction is with respect to the entire population is an

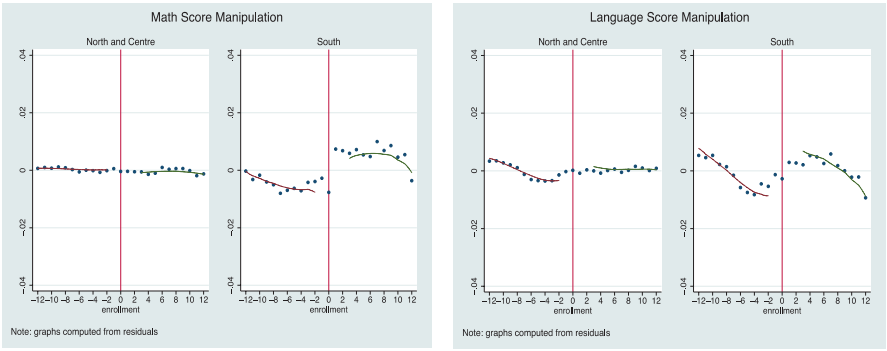

Fig. 3.5 Score manipulation by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017). (Graphs plot residuals from a regression of test scores on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit)

empirical matter, which should be discussed on a case-by-case basis. We have discussed some test for homogeneity of potential outcomes that allow to extend validity to the whole population of interest.

Finally, the idea of regression discontinuity is most easily put across by thinking of a properly conducted randomization only locally with respect to the discontinuity cutoff. The pros are clear-cut, and the cons concern the external validity of the estimates away from the relevant discontinuity.

What else could possibly go wrong? Books and chapters like this are always written to show a path forward for the implementation of methods. The day-to-day experience as a researcher is way more intricate. For example, Figure 3.5 taken from Angrist et al. (2017) casts doubt on the validity of the assumptions used in our discussion on the effects of class size. It shows that score manipulation also changes discontinuously at $r_i = 0$ in Southern Italy, suggesting that teachers in small classes are more likely to manipulate scores. As a result, the alleged causal effect of class size on test scores in Southern Italy discussed above does not reflect more learning in smaller classes, but increased manipulation of scores in smaller classes. As discussed by Angrist et al. (2017), these findings show how class size effects can be misleading even where internal validity is probably not an issue.

This example should prompt the reader to weigh methods with a grain of salt and a proactive attitude: the most credible approach to causal inference is often a combination of different identification strategies, and its credibility must stem from the institutional context under investigation rather than clueless statistical assumptions.

Review Questions

1. Why is the naïve comparison of mean outcomes for treated and control subjects not always informative of a causal effect?

2. What are the differences between internal, external, and statistical validity of a research design?
3. How does random assignment of the treatment help to achieve internal validity?
4. Under which assumptions do natural experiments and discontinuities provide a feasible avenue to estimate causal relationships?
5. What is the price to pay in terms of validity when pursuing these empirical strategies with respect to a proper randomization?

Replication Material

Access to data and codes is available from the American Economic Association website at: https://www.aeaweb.org/articles?id=10.1257/app.20160267

References

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *Economic Journal, 114*(494), C52–C83.

Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a small moment: class size and moral hazard in the Italian mezzogiorno. *American Economic Journal: Applied Economics, 9*(4), 216–249.

Angrist, J. D., & Imbens, G. W. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2), 467–475.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*(2), 533–575.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association, 110*(512), 1331–1344.

Bates, M. A., & Glennerster, R. (2017). The Generalizability Puzzle. *Stanford Social Innovation Review, Summer, 2017*, 50–54.

Battistin, E. (2016). *How manipulating test scores affects school accountability and student achievement*. *IZA World of Labor*.

Battistin, E., De Nadai, M., & Vuri, D. (2017). Counting rotten apples: Student achievement and score manipulation in Italian elementary schools. *Journal of Econometrics, 200*(2), 344–362.

Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics, 142*(2), 715–730.

Bertanha, M., & Imbens, G. W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics, 38*(3), 593–612.

Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics, 104*, 65–77.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Rand McNally.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica, 74*(5), 1191–1230.

Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics, 97*(5), 1081–1092.

Duflo, E., Glennerster, R., & Kremer, M. (2008a). Using randomization in development economics research: A toolkit. *Handbook of development economics, 4*, 3895–3962.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., & Oregon Health Study Group. (2012). The Oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics, 127*(3), 1057–1106.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics, 37*(3), 447–456.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201–209.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lee, D. S., & Lemieux, T. (2010a). Regression discontinuity designs in economics. *Journal of economic literature, 48*(2), 281–355.

Leigh, A. (2018). *Randomistas: How radical researchers changed our world*. Yale University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics–development policy, randomized controlled trials, and external validity. *The World Bank Research Observer, 33*(1), 34–64.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review, 99*(1), 179–215.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics, 134*(2), 557–598.

Suggested Readings

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.

Duflo, E., Glennerster, R., & Kremer, M. (2008b). Using randomization in development economics research: A toolkit. *Handbook of development economics, 4*, 3895–3962.

Lee, D. S., & Lemieux, T. (2010b). Regression discontinuity designs in economics. *Journal of economic literature, 48*(2), 281–355.

Chapter 4
Correlation Is Not Causation, Yet… Matching and Weighting for Better Counterfactuals

Fedra Negri

Abstract Anyone who has attended a statistics class has heard the old adage "correlation does not imply causation," usually followed by a series of hilarious graphs showing spurious correlations. Even if we strongly agree with it, this reminder has been taken a little too far: it is repeated like a mantra to criticize every observational study as being unable to detect causation behind statistical association. This chapter helps the reader go beyond the mantra, firstly, by explaining that "correlation does not imply causation" in observational studies because of selection bias (i.e. the composition of treatment and control groups follows a non-random selection) and parametric model dependence. Then, it introduces readers to weighting and matching techniques, smart statistical tools for reducing imbalance in the empirical distribution of pretreatment covariates between the treatment and control groups. Lastly, it provides an empirical illustration by focusing on two powerful algorithms: the entropy balancing (EB) and the coarsened exact matching (CEM). The chapter ends with caveats.

Learning Objectives
After studying this chapter, you should be able to:

- Understand under which assumptions correlation unveils causation in observational studies.
- Understand the inferential logic behind the commonest propensity score matching procedures and their key implementation steps.
- Understand the logical and computational problems related to the so-called "propensity score tautology".
- Grasp the theoretical and computational improvements introduced by entropy balancing and coarsened exact matching, respectively.

F. Negri (✉)
University of Milan, Milan, Italy

University of Milan - Bicocca, Milan, Italy
e-mail: fedra.negri@unimib.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_4

- Generate well-balanced samples on the statistical software Stata through the *ebalance* and the *cem* algorithms.
- Openly discuss the necessary conditions for their inferences on observational data to justify a causal interpretation.

4.1 Introduction

The very first notion almost everyone learns in their introductory statistics classes is that "correlation does not imply causation." Usually, students are presented with several examples of spurious correlations to stress that just because two variables move in *tandem*, this does not necessarily signal a causal relationship between them. A typical example is the negative and statistically significant correlation between final college grades and the amount of time students spend studying (Atkinson et al., 1996), and a number of funny graphs are available online (see: www.tylervigen.com).

Let us put it clearly: we strongly agree that "correlation does not imply causation." However, we also think that in the everyday practice of statistics and especially statistics teaching, the message this sentence carries has been taken a little too far and beyond its scope. In fact, it is repeated like a mantra, to criticize every observational study as being unable to detect causation behind statistical association. The warning "correlation does not imply causation" has made many social scientists feel so uncomfortable with causal inference that they even try to avoid causal language (King et al., 1994: 75–76). Terms such as "effect" or "impact" and verbs such as "to determine" or "to shape" are routinely avoided in scientific publications and replaced by the calculatedly ambiguous "association" and "link" and "to increase/to decrease" (Hernán, 2018).

Here, two related points should be stressed. First, while "correlation does not imply causation" for sure, "causation *does* imply correlation": if two variables are causally related, a change in one has to trigger a change in the other (Cook & Campbell, 1979; Miles & Shevlin, 2001: 113). Second, even when a statistical association, such as a regression coefficient, supports our preexisting views, theoretical claims, or a scenario we wish to be true (the so-called confirmation bias), uncertainty about causal inference will never be completely eliminated in observational studies. Thus, a statistical association is a non-sufficient, but still necessary, condition to make a causal claim. This means that we should not give up. Rather, we should provide the reader with the best and most honest estimate of the uncertainty of our causal claims (King et al., 1994: 75–76).

The chapter is structured as follows. Section 4.2 explains why "correlation does not imply causation" in observational studies, i.e. because of selection bias and model dependence. Section 4.3 introduces the reader to matching procedures, smart statistical tools that adjust for composition to correct for selection bias due to observable characteristics (Chap. 3, Sect. 3.2.5 and 3.2.6, provides a more general discussion on selection bias given by unobservable factors). In detail, this section

reviews and simplifies for the reader the latest contributions in the matching literature to emphasize both strengths and limitations of these techniques. Section 4.4 provides an application using the statistical software Stata by describing the algorithms developed by Heinmueller (2012), Iacus et al. (2009, 2011, 2012, 2019). Some *caveats* complete the chapter.

4.2 Not Just a Mantra: Correlation Is Not Causation Because…

4.2.1 Causal Inference Entails an Identification Problem

Causal inference (i.e. the process by which we make claims about causal relationships) can be thought of as an identification problem. Informally, a parameter is identified in a model if it is theoretically possible to learn its true value with an infinite number of observations (Matzkin, 2007: section 3.1). An identification problem arises when we do not have enough information to learn the true value of that parameter even if the sample is infinite (Manski, 1995).

The potential outcomes framework (Rubin, 1974; Holland, 1986) formalizes the causal inference identification problem and labels it as the "fundamental problem of causal inference." As discussed at length in Chap. 3 (see Sects. 3.2.2 and 3.2.3 for details), in the potential outcome framework, each unit i has two potential outcomes, $Y_i(1)$ if unit i is treated ($D_i = 1$) and $Y_i(0)$ if unit i is untreated ($D_i = 0$), but only one actual outcome, which depends on the actual treatment that unit i receives. Thus, the unit-level treatment effect, $\Delta_i = Y_i(1) - Y_i(0)$, is impossible to estimate because one of the two potential outcomes cannot be identified for each unit: for treated units, we observe $Y_i = Y_i(1)$ only; for control units, we observe $Y_i = Y_i(0)$ only.

Usually, we focus on the average treatment effect (ATE), which is the difference in the pair of potential outcomes averaged over the entire population of interest: $ATE = E(Y_i(1) - Y_i(0))$. Frequently, the ATE is defined for the subpopulation exposed to the treatment, the average treatment effect for the treated (ATT): $ATT = E(Y_i(1) - Y_i(0)| D_i = 1)$. Analogously, the average treatment effect for the non-treated (ATNT) is given by: $ATNT = E(Y_i(1) - Y_i(0)| D_i = 0)$.

However, moving from the unit-level treatment effect to the average treatment effects for the treated (ATT) or the non-treated (ATNT) does not solve our initial causal inference identification problem. Indeed, as regards the ATT, no additional amount of data will allow us to observe the average outcome under control for those units in the treatment condition, $E(\text{Yi}(0)|D_i = 1)$. Moving to the ATNT, no additional amount of data will allow us to observe the average outcome under treatment for those units in the control condition, $E(Y_i(1)| D_i = 0)$. The advanced reader may find a more formalized discussion in Keele (2015: 314–318).

Thus, the potential outcomes framework helps us in understanding that causal inference entails an unavoidable identification problem. Since no additional data can help us in solving this problem, we need to find a credible identification strategy.

4.2.2 Each Identification Strategy Entails a Set of Assumptions

An identification strategy is a research design and entails a set of assumptions, whose plausibility critically depends on the empirical context and should be discussed on a case-by-case basis (Angrist & Pischke, 2009; Morgan & Winship, 2014). The plausibility of some assumptions is testable. Think, for example, of the degree of compliance with the treatment assignment in a randomized experiment or to the first-stage requirement in a natural experiment with instrumental variation (see Chap. 3, Sect. 3.5.3.4, for details). Unfortunately, this is not always the case: untestable assumptions are unavoidable in causal inference. This is why reasoning about the plausibility of the assumptions entailed by the research design the researcher has chosen is a crucial preliminary step for social scientists aiming at detecting causal effects. This step precedes data collection and statistical analysis and often involves qualitative information about the institutional and empirical context (Keele, 2015: 323–324).

In what follows, we summarize the assumptions needed for statistical estimates to be given a causal interpretation under different research designs. Chapter 3 has already described three common research designs: randomized experiments, where treatment assignment is random, and quasi-experiments providing convincing substitutes to randomization, namely, instrumental variation and regression discontinuity designs (see Chap. 3, Sect. 3.5 and 3.6, for details).

Ideally, randomized experiments can achieve valid and relatively straightforward causal inferences if three requirements are met: (1) random selection of units to be observed from a given population, (2) random assignment of values of the treatment to each observed unit, and (3) large sample size. Random selection (1) avoids selection bias by guaranteeing that the probability of selection from a given population is related to the potential outcomes only by random chance. Combining random selection (1) with large sample size (3) guarantees that the chance that something will go wrong is extremely small. Random assignment (2) guarantees the absence of omitted variable bias even without any control variables included. Here, too, random assignment (2) plus large sample size (3) minimizes the chance of omitted variable bias (Ho et al., 2007: 205–206; see also Chap. 3, Sect. 3.4, for details).

However, social science research usually uses observational data that do not meet all of the three requirements. For example, survey research guarantees large sample size (3), but it is becoming more and more difficult to randomly select respondents due to increasing nonresponse rates (1), and it is almost impossible to fulfil random assignment requirement (2).

When dealing with observational data, a key further assumption is needed for statistical estimates to be given a causal interpretation: the so-called "selection on observables" assumption (Barnow et al., 1980; Heckman & Robb, 1985). Informally, the researcher has to assume that there is a set of covariates X_i such that treatment assignment D_i is random conditional on these covariates. This assumption is nonrefutable because it cannot be verified with observed data (Manski, 2007).

This assumption has a number of different names. In econometrics, it is also known as "no omitted variable bias," to emphasize that the model specification must include all the variables that are causally prior to the treatment assignment D_i, that are empirically related to D_i, and that affect the observed potential outcome Y_i, conditional on D_i (Goldberger, 1991; King et al., 1994: 76–82). Remember that only random assignment guarantees that D_i is independent of any X_i, whether measured or not, except by random chance (see Chap. 3, Sect. 3.4).

In statistics, the same assumption is known as "ignorability," to underline that the treatment assignment D_i and the unobserved potential outcomes are independent after conditioning on a set of covariates X_i and the observed potential outcomes so that there are no unobserved factors capable of biasing our estimates (Rubin, 1978). Alternative labels are the "absence of unmeasured confounding" or "conditional independence assumption."

Whatever the name, "selection on observables is a very strong assumption [...]. Generally, selection on observables needs to be combined with a number of different design elements before it becomes credible" (Keele, 2015: 322). Indeed, even admitting that the researcher has in mind the list of "correct" covariates to be incorporated in the model specification to meet this assumption, (1) additional data collection may be expensive and onerous, and (2) long model specifications increase the likelihood of incurring into over or bad control (Angrist & Pischke, 2009: 69). Problem (2) arises when we include in the model specification posttreatment covariates. In an experimental setting, it is quite easy to identify pretreatment and posttreatment covariates. With observational data, things get harder. Think, for example, about the items of a survey: if we exclude respondents' exogenous characteristics such as age, gender, citizenship, or parental level of education, it may be hard to state for sure that a covariate is "truly" pretreatment, and thus, it is not a consequence of D_i. Note that a further complication, known as the "M-bias" (Pearl, 2009a, b) will be discussed at length in Chap. 6.

This section aims to make it clear that there is no easy way-out and there is no magic. The identification problem cannot be solved by simply looking at data. Rather, we need to resort to identification strategies and each of them rests on a series of assumptions. When the data are observational, a very strong assumption is added to the list: the "selection on observables" one. This is the reason why "correlation [per se] does not imply causation." However, this is not the end of the story: selection on observables can be combined with statistical tools to boost its credibility (Keele, 2015).

4.2.3 Last but not Least: Model Dependence

Of course, any specific statistical tool we choose to boost the credibility of our identification strategy will make additional assumptions (Ho et al., 2007: 2010–2011).

Let us be honest: as social and political scientists, we usually spend a considerable amount of time in collecting, merging, cleaning, and recoding raw data. Then, we finally load our data set into our favorite statistical software and run several model specifications by using the parametric statistical technique that best fits our data (e.g., OLS, discrete choice models, duration models, etc.).

The main problem with this common procedure is that all parametric methods assume that we know the "right" model specification before looking at the estimates. A model is "right" if it is (a really good approximation to) the data-generating process. Otherwise, the model will miss important aspects of reality and inference will be systematically wrong or overly precise.

Instead, what happens in everyday research is that we start from a generic model specification suggested by our theoretical framework, previous works, or common sense, and then, we modify it by adding or removing control variables and interaction terms, changing the operationalization of some variables or the functional form, restricting the sample, etc.

Following this inductive procedure, we end up with several alternative estimates of the statistical relationship between our variable of interest and the dependent variable. However, to improve readability, we typically choose no more than ten model specifications to be included in our written work. This choice, made after looking at the estimates, entails methodological and ethical dilemmas. Moreover, it forces us to convince the readers (and the reviewers) that we have picked up the "right" specifications, not simply the ones that most supported our starting hypotheses.

Thus, even if rarely admitted, correlation also does not imply causation in observational studies because effect estimates may be model dependent, at least to some degree (Ho et al., 2007).

4.3 Preprocessing Data with Matching to Improve the Credibility of the Estimates

Imagine we want to estimate the effect of a policy in situations when controlled randomization is unfeasible, unethical, or politically sensitive and there are no convincing natural experiments providing a substitute for randomization such as the ones described in Chap. 3, Sects. 3.5 and 3.6 (i.e., instrumental variation and RDD). In these situations, matching may be a powerful non-parametric technique for boosting the credibility of the estimates. It is grounded on the idea that some serious statistical problems (i.e. model dependence, estimation error, and bias) can be downplayed by dropping heterogeneous observations from the raw data and thus limiting inferences to a carefully selected subsample.

4.3.1 No Magic: What Matching Can and Cannot Do

Before addressing any technicality, we want to stress a key point about matching. It is not a method of estimation of causal effects, it is "only" a non-parametric statistical tool for preprocessing raw data so that the treatment group becomes as similar as possible to the control group on a set of covariates chosen by the researcher (Arceneaux et al., 2006; Sekhon, 2009). Once treated units have been matched with control ones according to one among the available matching procedures, some method of estimation is needed to obtain an estimate of the causal effect. If the treatment and control groups are exactly balanced on the set of covariates chosen by the researcher (i.e. if the treatment and control covariate distributions are the same), then the method of estimation can credibly be a simple difference in means between the outcomes of the two groups. However, if the two groups are not exactly balanced (i.e. if there are still systematic differences between them, as usually happens), then the researcher has to further adjust the matched sample by using the parametric model they would have used anyway (e.g., Ho et al., 2007; Iacus et al., 2019). Thus, matching is just a convincing way to select the observations on which some methods of estimation should be later applied (with their own additional assumptions).

Exactly as when we interpret the coefficient of a multivariate regression model as a causal effect, matching procedures are grounded on the strong assumption of selection on observables. This means that it should be theoretically plausible that selection into treatment is completely determined by a set of covariates X_i that can be observed by the researcher such that conditioning on X_i, the assignment to treatment is as good as random. To put it differently, it should be theoretically plausible that there are not additional unobservable variables capable of pushing units into treatment.[1]

[1] Given that both matching and regression are based on the selection on observables assumption, the reader may wonder whether matching is really different from a regression with properly identified control variables. This question is the object of a heated debate among methodologists. Some maintained that both regression and matching are control strategies, and therefore, the differences between the two are unlikely to be of major empirical importance (Angrist & Pischke, 2009: section 3.3.1). Others have pointed out shortcomings of regression relative to matching: Dehejia and Wahba (1999), for example, found that propensity score matching procedures have more closely approximate results from a randomized experiment than regression alone. Further, some have underlined that regression is a parametric approach imposing a global linear relationship between Xs and Y and that it uses all the available observations, thereby involving a certain amount of extrapolation, while matching is a non-parametric approach that discards observations for which a reasonably close match cannot be found (Martini & Sisti, 2009: 221–225). Others have stated that matching involves several choices in its implementation, which could lead to subjectivity in the results. According to Imbens and Wooldridge, "the best practice is to combine linear regression with either propensity score or matching methods" (2008: 19–20) as in this way, the estimated effect will explicitly rely on local, rather than global, linear approximations to the regression function. Even though adjudicating between these views is beyond the scope of this chapter, the application discussed in Sect. 4.4 embraces this last suggestion and thus combines the CEM algorithm with OLS regression.

However, compared to regression, preprocessing raw data with matching eliminates, or at least reduces, the selection bias due to the set of covariates chosen by the researcher, which renders any subsequent parametric adjustment either irrelevant (if balance is fully achieved) or less important (if balance is partially achieved). To put it simply, given the plausibility of the selection on observables assumption, preprocessing data with matching makes causal effect estimates based on the subsequent parametric analyses far less dependent on modeling choices and specifications. Quoting Ho et al. (2007: 233): "Analysts using preprocessing have two chances to get their analyses right, in that if either the matching procedure or the subsequent parametric analysis is specified correctly (and even if one of the two is incorrectly specified), causal estimates will still be consistent" (on this, see also Robins & Rotnitzky, 2001). Moreover, it has been proved that when matching is applied carefully so that n is not much smaller in the matched sample than in the original sample, it leads to a reduction in both bias and variance of estimates from subsequent parametric analyses (Rubin & Thomas, 1996; Imai & van Dyk, 2004).

4.3.2 Useful Starting Point: Exact Matching

Let us formalize the selection on observables assumption. Remember that we aim to estimate the average treatment effect for the treated: $ATT = E(Y_i(1) - Y_i(0)| D_i = 1)$. Unfortunately, we do not observe the average outcome under control for those units in the treatment condition, $E(\text{Yi}(0)|D_i = 1)$. Instead, we observe the average outcome under control for those units in the control condition, $E(\text{Yi}(0)|D_i = 0)$. As discussed in Chap. 3, Sect. 3.2.3, a naive comparison of outcomes by treatment status provides a biased estimate of the ATT:

$$E\left(Y_i\left(1\right) \mid D_i = 1\right) - E\left(Y_i\left(0\right) \mid D_i = 0\right) =$$
$$E\left(Y_i\left(1\right) - Y_i\left(0\right) \mid D_i = 1\right) + \left[E\left(Y_i\left(0\right) \mid D_i = 1\right) - E\left(Y_i\left(0\right) \mid D_i = 0\right)\right]$$

The first term on the right-hand side of the equation is the ATT (the quantity we are interested in); the second term is the sample selection bias that accounts for the differences in outcome under control between treated and control units. We already know that only if the three requirements of an ideal RCT are met (i.e. (1) random selection, (2) random treatment assignment, and (3) large sample size), the sample selection bias is zero, and thus, the naive comparison of outcomes by treatment status provides an unbiased estimate of the ATT.

Now, let X_i be a set of pretreatment covariates. The selection of the set of covariates X_i by the researcher is a critical step. According to the usual rules for avoiding omitted variable bias, X_i should include all variables that affect both the treatment assignment D_i and, controlling for the treatment, the dependent variable Y_i (this does not mean that every available pretreatment variable should be included in X_i because it will reduce efficiency). However, to avoid a "posttreatment bias" (King & Zeng,

2007), variables that may be even remotely consequences of the treatment variable should never be included in X_i (Cox, 1958: section 4.2; Rosenbaum, 1984; Rosenbaum, 2002: 73–4).

According to the selection on observables assumption, once we condition on X_i, assignment to treatment D_i is independent from the unobserved potential outcomes $Y_i(0)$ and $Y_i(1)$:

$$Y_i(1), Y_i(0) \perp D_i \,|\, X_i$$

Under this assumption, conditioning on X_i, the average outcome under control for those units in the control condition is equal to the average outcome under control for those units in the treatment condition:

$$E\left(Y_i(0) \mid D_i = 0, X_i\right) = E\left(Y_i(0) \mid D_i = 1, X_i\right) = E\left(Y_i(0) \mid X_i\right)$$

Similarly, conditioning on X_i, the average outcome under treatment for those units in the control condition is equal to the average outcome under treatment for those units in the treatment condition:

$$E\left(Y_i(1) \mid D_i = 0, X_i\right) = E\left(Y_i(1) \mid D_i = 1, X_i\right) = E\left(Y_i(1) \mid X_i\right)$$

Thus, the expected value of Y_i is independent from D_i, given X_i. Using the Law of Iterated Expectations, the ATT is given by:

$$\begin{aligned} ATT &= E\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] = E\left[\, E\left[Y_i(1) - Y_i(0) \mid D_i = 1,\ X_i\right] \mid D_i = 1\right] \\ &= E\left[\, E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 1,\ X_i\right] \mid D_i = 1\right] \end{aligned}$$

The term $E\,[\,Y_i(0) \mid D_i = 1, X_i]$ is counterfactual, but under the selection on observables assumption, we have:

$$ATT = E\left[\, E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 0,\ X_i\right] \mid D_i = 1\right]$$

We can rewrite it as:

$$ATT = E\left[\, \delta_x \mid D_i = 1\right]$$

where δ_x is the difference in means by treatment status at each value of X_i.

$$\delta_x = E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 0,\ X_i\right]$$

This is the identification strategy employed by the so-called "exact matching." Informally, it suggests preprocessing the data so that each treated unit is matched

with all the available control units that have exactly the same covariates values (do not confuse the exact matching with the one-to-one exact matching, which is more limited because it uses only one control unit for each treated unit). If, after exact matching, a large number of treated units are exactly matched with one or more control units, then we have an exact balance with little inefficiency. This means that a (weighted) difference between the average outcomes of matched treated and control units is sufficient to obtain an unbiased estimate of the ATT. We added "weighted" in parentheses because, since each treated unit can be matched with more than one control unit, a weighted difference in means across exactly matched subclasses is suggested to account for the difference in the number of treated and control units. Beware that if some treated units cannot be matched because there is not at least one control unit with exactly the same covariates values, the exact matching procedure drops these treated units. By dropping some treated units, we alter the *estimand*: it is no longer the ATT, but a more local version of it (Crump et al., 2009; Rubin, 2010). As discussed in Chap. 3, Sect. 3.3.3, this may weaken the external validity of the estimates. This choice is reasonable as long as the researcher is transparent about it and its consequences in terms of the new set of treated units over which the causal effect is defined (Iacus et al., 2012: 5).

If an insufficient number of exact matches are found, and thus, many treated units have to be discarded, the researcher has to switch to other matching procedures that preprocess the data so that each treated unit is matched with all the available control units that have approximately the same covariates values.

4.3.3 Propensity Score Tautology

The best practice for approximate matching procedures involves two steps. The first step drops treated and control units outside the so-called "common support" of both groups. Informally, the common support assumption requires that for any treated unit with given covariate values, it is also possible to observe a control unit with the same (or approximately the same) covariate values. Thus, ensuring common support requires the researcher to drop observations where the empirical density of treated and control units does not overlap since including these observations would require extrapolation from the data, which can generate considerable model dependence.

To accomplish this first step, King and Zeng (2007) suggest pruning observations from the control group that are outside of the "convex hull" of the treatment group. Informally, with one pretreatment covariate X, the convex hull of the treatment group is the range of the subset of observations of X that are in the treatment group so that control units with values of X greater than $max(X|T = 1)$ or lower than $min(X|T = 1)$ are discarded. Similarly, if any treated units fall outside the convex hull of the control units, these are also discarded (see also Iacus & Porro, 2009 for another conservative way of identifying common support). Remember once more

that dropping treated units changes the *estimand*: it is no longer the ATT, but a more local version of it.

The second step matches treated units with control units so that they are as close as possible according to some metric. However, as anticipated, establishing on which dimensions the degree of closeness between treated and control units has to be evaluated (i.e. selecting the pretreatment covariates to be included into X_i) is not easy: the researcher might be willing to include a large set of covariates, many of them multivalued or continuous. This problem is known as "the curse of dimensionality."

Rosembaum and Rubin (1983) addressed this problem by developing a matching procedure based on the propensity score, defined as the conditional probability of receiving the treatment given the pretreatment covariates selected by the researcher. They start from the usual selection on observables assumption: once we condition on X_i, the average potential outcome under control for those units in the treatment condition should be equal to the average potential outcome under control for those units in the control condition. Thus, once we condition on X_i, the average potential outcome under control should be the same irrespective of the treatment condition:

$$E\left(Y_i\left(0\right) \mid D_i = 1, X_i\right) = E\left(Y_i\left(0\right) \mid D_i = 0, X_i\right) = E\left(Y_i\left(0\right) \mid X_i\right)$$

They move on by demonstrating that if potential outcomes are independent of treatment status conditional on the set of covariates X_i, then potential outcomes are also independent of treatment status conditional on a scalar function of the same covariates X_i, labelled "propensity score." They collapsed the set of covariates X_i into a monodimensional variable that measures, for each unit i, the probability of receiving treatment given the values of its set of covariates X_i, $P(D_i = 1 \mid X_i)$. Usually, it is estimated through a logit or a probit function, which regresses D_i on a constant term and the set of covariates X_i chosen by the researcher, without looking at Y_i:

$$E\left(Y_i\left(0\right) \mid D_i = 1, P\left(X_i\right)\right) = E\left(Y_i\left(0\right) \mid D_i = 0, P\left(X_i\right)\right) = E\left(Y_i\left(0\right) \mid P\left(X_i\right)\right)$$

Approximate matching methods based on the propensity score tend to skip the first step and to check for common support only after having estimated the propensity score for each observation i. Indeed, they drop control units that have a propensity score lower than the minimum or higher than the maximum of the propensity score of the treated units (Khandker et al., 2010).

However, the reader may have already realized that the propensity score solution by Rosembaum and Rubin (1983) is a tautology. The propensity score has been developed to solve the course of dimensionality problem (i.e. too many dimensions to be controlled for to match treated and control units). However, since we do not know the "true" propensity score, it has to be estimated through a probability model that adds the same dimensions as independent variables. Moreover, the only way to check the validity of the specification of the estimated propensity score (i.e. to check whether the estimated propensity score is a consistent estimate of the "true"

propensity score) is to stratify the sample over small propensity score intervals and then, for each covariate in each interval, test whether the means of the treated and control units are not statistically different. If this is not the case, the researcher has to improve the specification of the *probit* or *logit* function he/she used to estimate the propensity score and start again (Dehejia & Wahba, 1999; Becker & Ichino, 2002). Unfortunately, there is no way out from the propensity score tautology: "[I]t works when it works [when matching on the propensity score balances the raw covariates], and when it does not work, it does not work (and when it does not work, keep working at it)" (Ho et al., 2007: 219).

4.3.4 How to Choose Among Matching Procedures?

Once the researcher has estimated the propensity score for each unit *i*, they have to choose a metric to match treated and control units. Several metrics are available: they vary in the strategy they follow to select the matches and in the weight they associate with each match. Table 4.1 lists the most widely used approximate matching procedures based on the propensity score and provides references for further readings (see also Caliendo & Kopeinig, 2008).

Given this long and non-exhaustive list of approximate matching procedures, how can we choose among them? The methodological literature does not provide a clear-cut answer. Since the main diagnostics of success in matching are balance (i.e. the degree to which the treatment and the control group covariate distributions resemble each other) and the number of observations remaining after preprocessing

Table 4.1 Commonest approximate matching techniques based on the propensity score

Technique	Description	Further readings
Nearest neighbor matching	For each treated unit, the algorithm finds the control unit with the nearest propensity score. This can be done with or without replacement. In the former case, an untreated unit can be used more than once as a match. In the latter case, if the nearest control unit has already been matched to another treated unit, the algorithm does not consider it and searches for a new one.	Smith (1997), Smith and Todd (2005)
Caliper and radius matching	For each treated unit, the caliper matching algorithm finds the closest control unit whose propensity score falls within a radius *r* chosen by the researcher. The radius version matches each treated unit with all the control units within the radius *r*.	Smith and Todd (2005), Dehejia and Wahba (2002)
Stratification matching	The algorithm partitions the sample into a set of intervals (strata) so that in each stratum, the propensity score of treated and control units have the same mean value.	Imbens (2004)
Kernel matching	The algorithm matches every treated unit with a weighted average of (nearly) all control units with weights that are inversely proportional to the distance between the propensity scores.	Heckman et al. (1997, 1998)

the raw data, a rule of thumb is to preprocess raw data by running as many approximate matching procedures as possible. To avoid any confirmation bias, it is crucial that the researcher performs this comparison without consulting Y. Then, they have to choose the procedure that maximizes balance while keeping n as large as possible (Ho et al., 2007). As the reader may have foreseen, this search for the matching procedure that maximizes balance and the number of observations may be tedious as the researcher has to manually iterate between the available algorithms (Ho et al., 2007; Iacus et al., 2009; Heinmueller, 2012; King & Nielsen, 2019). Section 4.4 describes two techniques that address this problem.

To assess balance, Ho et al. (2007: 221) suggest the following options: first, comparing the mean of each variable X_i in the treatment group with the mean of each variable in the control group (if one or more of these differences differ by more than a quarter of a standard deviation of the respective X_i variable, a better balance is needed) (Cochran, 1968); second, comparing treatment and control histograms one variable at a time; third, using a quantile–quantile plot (QQ plot) for each variable to compare the full empirical distributions of each variable for the treatment and control groups; and lastly, the same QQ plot can be used for the propensity scores of the treatment and control groups. Even if tautological (it relies on the propensity score as a summary of the data to check whether the chosen propensity score matching is adequate), it may be a good low-dimensional summary (Ho et al., 2007: 221–223; see also Rubin, 2001; Austin & Mamdani, 2006; Imai et al., 2008).

One might object that increasing balance by throwing away unmatched observations will reduce statistical efficiency (i.e. the mean squared error of the estimated effect might increase). However, "efficiency should be a secondary concern for observational students" (Keele, 2015: 325). In a randomized experiment, where selection bias is known to be zero, adding observations simply increases power. On the other hand, in an observational study, increasing the sample size may shrink the confidence intervals to a point that excludes the "true" treatment effect point estimate (Cochran & Chambers, 1965). Moreover, Rosenbaum (2004, 2005) demonstrated that in observational studies, reducing unit heterogeneity reduces both sampling variability and sensitivity to bias from unobserved covariates. Thus, as a rule of thumb, there are reasons for preprocessing raw data through matching procedures in order to reduce heterogeneity between the treatment and control groups according to a set of observable covariates (for theoretical and simulation results, see also Rubin & Thomas, 1992, 1996; Imai & Van Dyk, 2004; Imbens, 2004; Morgan & Winship, 2014; Stuart, 2010).

4.3.5 The End: The Parametric Outcome Analysis

Having selected the matching algorithm that maximizes balance while keeping n as large as possible, the researcher has to move to the usual parametric analysis to obtain a causal effect estimate. Indeed, matching is just a non-parametric statistic tool for reweighting or simply discarding units in the raw data so that the treatment

and control groups become as similar as possible on a set of observable covariates or, to put it differently, so that the treatment variable becomes as close as possible to being independent of the background characteristics.

The causal effect can be estimated through a simple (weighted) difference in means between the observed outcomes of the treatment and control groups only if they are exactly balanced. Indeed, the difference in means is equivalent to regressing Y_i on D_i without any control variables, thus assuming that D_i and X_i are unrelated. This assumption is plausible only if exact matching has been achieved for the treated units, which is very unlikely. By computing a simple difference in means on a preprocessed sample where there is some remaining imbalance between the treatment and the control groups, we would certainly incur in an omitted variable bias.

Thus, whenever the treatment and control groups are not exactly balanced, the researcher is better off using the same parametric model he/she would have also used on the raw data without preprocessing. Preprocessing data with matching makes causal effect estimates based on the subsequent parametric analyses far less dependent on modeling choices and specifications (Ho et al., 2007; (Iacus et al., 2019).

4.4 Empirical Illustration

LaLonde (1986) was the first to assess the performance of several non-experimental estimators by using experimental data as a benchmark. His experimental data came from the National Supported Work Demonstration (NSWD), a subsidized work experience program that took place in 1975–1976 in the United States. The program consisted into providing trainees with work in a sheltered training environment and then assisting them in finding regular jobs. To take part in the NSWD, potential participants had to satisfy a set of eligibility criteria intended to identify individuals with significant barriers to employment. Then, actual treatment (i.e. the subsidized work experience) was randomized among applicants meeting the eligibility criteria.

Using a simple difference in means between the observed post-intervention earnings of the treatment and control groups, LaLonde (1986) obtained an unbiased estimate of the effect of the subsidized work experience: the program was estimated to increase post-intervention earnings by $1,794 with a 95% confidence interval of [551; 3,038]. Thus, according to this experimental result, the program was successful. Then, he compared this experimental result to those obtained from several non-experimental estimators applied to the NSWD observations that received training (treated units only) and a set of control observations constructed ex post from two standard population survey data sets (i.e. CPS and PSID). His findings show that alternative non-experimental estimators produce very different estimates, most of which deviate substantially from the experimental benchmark.

Several subsequent studies have reanalyzed LaLonde's results, using more recent statistical procedures (e.g., Dehejia & Wahba, 1999; Becker & Ichino, 2002; Smith & Todd, 2005; Iacus et al., 2009, 2012, 2019). Notably, Dehejia and Wahba (1999)

restricted LaLonde's data set to individuals from whom data on previous earnings were available in 1974 and compared several matching estimations to a fully saturated in *X* OLS regression (original samples and replication materials are available on Dehejia's page: https://users.nber.org/~rdehejia/nswdata2.html). They concluded that matching procedures dominated fully saturated in *X* regression. However, Smith and Todd (2005) showed that Dehejia and Wahba's findings came from the specific sample chosen by the authors, but they did not hold on other samples. Thus, they argued that estimating the causal effect by simply preprocessing data with matching and then computing a (weighted) difference in mean between the treatment and control groups seems not to perform better than a fully saturated in *X* OLS regression. Thus, as explained in the Sect. 4.3.5, after having preprocessed data with the matching procedure that maximizes balance while saving enough of *n*, a method of estimation should be applied. Smith and Todd (2005), for example, found that a combination of matching and difference-in-differences performs the best.

This section summarizes and simplifies for the reader the very latest contribution in this long *querelle* about LaLonde results and matching procedures. Indeed, we focus on the theoretical refinements by Heinmueller (2012) and Iacus et al. (2019) and on the algorithms they, respectively, developed: entropy balancing (EB; Heinmueller & Xu, 2013) and coarsened exact matching (CEM; Blackwell et al., 2009).

EB and CEM are similar from several points of view. Both of these techniques are used in observational studies to preprocess the raw data prior to the estimation of a binary treatment effect under the assumption of selection on observables, and both of them are aimed at improving the covariate balance between the treatment and control groups. Moreover, both techniques overcome the propensity score tautology by requiring the researcher to establish the desired degree of covariate balance before the preprocessing adjustment. Lastly, both of them are computationally efficient and have been proved to reduce model dependence for the subsequent estimation of the treatment effect via parametric outcome analysis.

However, they also differ in important ways. As explained below, CEM coarsens each covariate into substantively meaningful categories identified ex ante by the researcher and then matches units exactly on this coarsened scale. Treated and control units that cannot be exactly matched are discarded. As the reader already knows, by discarding treated units, CEM changes the *estimand* from the ATT to a more local treatment effect for the remaining treated units (see Iacus et al., 2009 for reasons for why this can be beneficial). On the other hand, EB leaves the *estimand* unchanged because it does not discard treated units. Sections 4.4.1 and 4.4.2. assist readers in getting familiar with these two algorithms.

4.4.1 Entropy Balancing

EB is a data preprocessing method proposed by Heinmueller (2012). Crudely put, the algorithm works as follows. As usual, the researcher has to identify a set of pretreatment covariates according to his/her substantive knowledge, previous studies,

and data availability. Then, for each covariate, the researcher has to pre-specify a potential large set of balance constraints to equate the moments of the covariate distribution between the treatment and the control groups. The moments refer to the mean (first moment), the variance (second moment), and the skewness (third moment). For example, the researcher can request that the mean values (first moments) of a set of covariates in the control group exactly equate to the mean values of the same set of covariates in the treatment group. Moreover, they can also include interaction terms such that, for example, the mean of one covariate is balanced across subgroups of another covariate. Lastly, the algorithm searches for a set of entropy weights to satisfy the balance constraints imposed by the researcher, while remaining as close as possible to the uniformly distributed base weights to prevent loss of information.

EB has several attractive features. Its reweighting scheme directly incorporates the researcher's knowledge about the moments in the treatment group and adjusts the weights to balance the covariate distribution exactly in finite samples, without discarding any treated unit. These are key improvements as they overcome the time-consuming search over propensity score models without changing the *estimand*. Moreover, the weights that result from EB can be easily incorporated into any standard statistical model the researcher would have used even without the preprocessing step.

To illustrate the functioning of EB, Heinmueller and Xu (2013) rely on the subset of the original LaLonde data set (1986) already used by Dehejia and Wahba (1999). The data set provides information on 185 treated units from the NSWD that were involved in the subsidized work experience and 15,992 non-participants from the Current Population Survey Social Security Administration File (CPS-1). The former constitutes the treatment group, and the latter the control group. Remember that this control group is not the one identified through randomization during the NSWD. Instead, this control group is built ex post by using the CPS.

The treatment variable, *treat*, is 1 for participants and 0 for nonparticipants. The outcome variable is real earnings in 1978 US dollars (*re78*). The available pretreatment covariates include age (*age*), years of education (*educ*), marital status (*married*), lack of a high school diploma (*nodegree*), race (*black*, *hispanic*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), and real earnings in 1974 (*re74*) and 1975 (*re75*). The *estimand* is the increase in earnings in 1978 due to the subsidized work experience.

By simply regressing *re78* on the treatment variable and all the controls, it seems that being exposed to the subsidized work experience increased earnings in 1978 by $1,068 (Fig. 4.1). However, the 95% confidence interval is large enough that the relative estimate is not statistically different from 0. Remember that in this lucky case, we know from the NSWD experimental result that being exposed to the treatment increased earnings in 1978 by $1,794 with a 95% confidence interval of [551; 3,038]. Thus, the OLS estimate on the raw data is substantially lower than the benchmark effect established on the experimental data.

Thus, the authors preprocess the raw data using EB. The basic syntax of the command *ebalance* requires the researcher to list the treatment variable (*treat*) and the

```
. reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75
```

Source	SS	df	MS	Number of obs	=	16,177
				F(11, 16165)	=	1343.88
Model	7.2418e+11	11	6.5835e+10	Prob > F	=	0.0000
Residual	7.9190e+11	16,165	48988567.3	R-squared	=	0.4777
				Adj R-squared	=	0.4773
Total	1.5161e+12	16,176	93724175.2	Root MSE	=	6999.2

re78	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	1067.546	554.0595	1.93	0.054	-18.47193	2153.564
age	-94.54102	6.000283	-15.76	0.000	-106.3022	-82.7798
educ	175.2255	28.69658	6.11	0.000	118.977	231.474
black	-811.0888	212.8488	-3.81	0.000	-1228.296	-393.8815
hispan	-230.5349	218.6098	-1.05	0.292	-659.0344	197.9646
married	153.2284	142.7748	1.07	0.283	-126.626	433.0828
nodegree	342.9265	177.8778	1.93	0.054	-5.733561	691.5866
re74	.2914332	.0127311	22.89	0.000	.2664789	.3163875
re75	.4426945	.0128868	34.35	0.000	.417435	.467954
u74	355.5564	231.6004	1.54	0.125	-98.40599	809.5189
u75	-1612.758	239.803	-6.73	0.000	-2082.798	-1142.717
_cons	5762.18	445.6145	12.93	0.000	4888.726	6635.634

Fig. 4.1 OLS regression on the raw data

pretreatment covariates he/she will focus on (e.g., *age*, *educ*, *black*, and *hispan*). The most important option in *ebalance* is *targets(numlist)* as it allows the researcher to impose the balance constraints for the included covariates. In detail, the researcher has to specify a number (1, 2, or 3) that corresponds to the highest covariate moment that should be adjusted for each covariate.

For example, this code requests that the mean, variance, and skewness of the variables *age*, *educ*, *black*, and *hispan* are adjusted: ebalance treat age educ black hispan, targets (3).

As shown in Fig. 4.2, the command returns the number of treated and control units. Note that EB does not discard treated units (185), thus keeping the original *estimand*. Then, it reports descriptive statistics on the mean, variance, and skewness of the selected covariates in the treatment and in the control groups, before and after the reweighting procedure. As requested, the algorithm perfectly balances the two groups on first-, second-, and third-order moments by fitting the EB weights. By default, the EB weights are stored in a variable named *_webal* and can be readily used for subsequent analysis.

By writing 2 instead of 3 in parentheses, the algorithm would have balanced only the mean and variance of the same variables; by writing 1, it would have balanced only the mean of the same variables. The command also allows to specify specific constraints to each variable (see Fig. 4.3). For example, according to the command:

ebalance will adjust the first moment for *age* and *educ*, the first and the second moments for *black* and the first, second, and third moments for *hispan*.

To reweight the original LaLonde (1986) data set, Heinmueller and Xu (2013) adjust the sample by including the means, variances, and skewness of all of the 10

```
Treated units: 185      total of weights: 185
Control units: 15992    total of weights: 185

Before: without weighting
```

	Treat			Control		
	mean	variance	skewness	mean	variance	skewness
age	25.82	51.19	1.115	33.23	122	.3478
educ	10.35	4.043	-.7212	12.03	8.242	-.4233
black	.8432	.1329	-1.888	.07354	.06813	3.268
hispan	.05946	.05623	3.726	.07204	.06685	3.311

```
After:  _webal as the weighting variable
```

	Treat			Control		
	mean	variance	skewness	mean	variance	skewness
age	25.82	51.19	1.115	25.8	51.16	1.122
educ	10.35	4.043	-.7212	10.34	4.04	-.7119
black	.8432	.1329	-1.888	.8421	.1329	-1.877
hispan	.05946	.05623	3.726	.05966	.05611	3.718

Fig. 4.2 The output of the *ebalance* command

```
. ebalance treat age educ black hispan, targets(1 1 2 3)

Data Setup
Treatment variable:   treat
Covariate adjustment: age educ black hispan (1st order). black hispan (2nd order). hispan (3rd order).
```

Fig. 4.3 Options of the *ebalance* command

pretreatment covariates plus squared terms and first-order interactions of the same 10 covariates and cubed terms for *age*, *educ*, *re74*, and *re75*.

By running the initial OLS regression on the reweighted data, the treatment effect estimate suggests that being exposed to the subsidized work experience increased earnings in 1978 by $1,761 with a 95% confidence interval of [333; 3,190]. Thus, the simple OLS estimate on the reweighted data is very close to the experimental target answer ($1,794 with a 95% confidence interval of [551; 3,038]). A similar conclusion may be achieved by regressing *re78* on *treat* only (Fig. 4.4).

4.4.2 Coarsened Exact Matching

All the matching procedures based on the propensity score (see Table 4.1) assume that the data generation process is based on simple random sampling, which means that drawing repeated hypothetical samples of fixed size $n < \infty$ at random from a population of θ units with covariates X, each sample of n observations has an equal probability of selection.

```
. svy: reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         1                  Number of obs      =    16,177
Number of PSUs     =    16,177                  Population size    =       370
                                                Design df          =    16,176
                                                F(  11,  16166)    =      3.08
                                                Prob > F           =    0.0004
                                                R-squared          =    0.0942
```

re78	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	1761.951	729.079	2.42	0.016	332.8754	3191.027
age	-18.3346	49.97697	-0.37	0.714	-116.295	79.62578
educ	446.504	214.4348	2.08	0.037	26.188	866.82
black	-629.7442	848.1102	-0.74	0.458	-2292.134	1032.646
hispan	1263.509	1486.528	0.85	0.395	-1650.25	4177.269
married	-288.1295	908.5693	-0.32	0.751	-2069.026	1492.767
nodegree	-76.56057	1127.997	-0.07	0.946	-2287.559	2134.438
re74	.1277705	.1639249	0.78	0.436	-.1935404	.4490815
re75	.394469	.169975	2.32	0.020	.0612992	.7276387
u74	2585.847	1635.288	1.58	0.114	-619.4993	5791.193
u75	-1570.203	1585.939	-0.99	0.322	-4678.818	1538.413
_cons	-755.398	3829.273	-0.20	0.844	-8261.196	6750.4

Fig. 4.4 OLS regression on the reweighted data

CEM modifies this assumption by theorizing that the data generation process guarantees stratified random sampling. Informally, the adjective "stratified" means that random sampling does not apply directly to the population of θ units, but to strata or partitions, within this population, that are identified by the researcher according to his/her knowledge of the set of covariates X. For example, if the set of covariates X includes age, gender, and earnings, a stratum may refer to young males making more than $25,000. Inside this stratum, sample selection should be random (Iacus et al., 2019: 48–49). Then, as with all the other matching procedures, CEM is grounded on the selection on observables and on the common support assumptions (even if inside each stratum; see Iacus et al., 2019: 50–51).

As the reader may have already realized, the emphasis is on the definition of strata by the researcher. The authors underline that this step is case specific and critically reflects "the knowledge the investigator must have" (Iacus et al., 2019: 54). Indeed, the CEM algorithm helps the researcher in coarsening each variable among the set of pretreatment covariates judged as relevant into substantively meaningful categories that reduce variability while at the same time preserving information. The easiest example is the variable reporting the years of education that can be easily coarsened into categories such as high school, some college, college graduates, etc.

Starting from the LaLonde's data set (1986), Iacus et al. (2009, 2011, 2012, 2019) show that CEM, on average, dominates commonly used matching procedures in a large variety of real and simulated data sets because it reduces imbalance, model

dependence, estimation error, bias, variance, and mean square error. Moreover, it usually produces more matched units. Furthermore, while to improve propensity score matching, the researcher has to marginally change and rerun the model, recheck imbalance, and rerun the model again several times (King & Nielsen, 2019), and CEM makes it easier to find a specification that improves balance. Indeed, strata are explicitly defined ex ante by the researcher according to his/her substantive knowledge on the covariates: reducing maximum imbalance on one variable never has any effect on the maximum imbalance specified for any of the other variables (Iacus et al., 2012: 21). Let us apply this algorithm to the subset of the original LaLonde data set (1986) already used by Dehejia and Wahba (1999). For an application on the original experimental LaLonde's data set, see Blackwell et al. (2009).

First, we have to assess the imbalance in the original unmatched data through the λ^1 statistic (Iacus et al., 2008). This statistic ranges from 0, meaning perfect global balance between the treatment and the control groups, to 1, meaning complete separation between the two (Fig. 4.5).

The *imb* (meaning "imbalance") command works as follows. The researcher has to list the pretreatment covariates they want to focus on (in the example, *age*, *educ*, *black*, and *hispan*), followed by the indication of the treatment variable (*treat*). First, the Stata output shows the λ^1 statistic. In our example, $\lambda^1 = 0.893$, thus signaling that the original unmatched data are highly unbalanced. Note that the λ^1 value is not valuable on its own: it is as a point of comparison between matching solutions. The value 0.893 is a baseline reference for the unmatched data. The researcher has to compare the λ^1 value obtained on the matched data to the value 0.893 obtained on the unmatched data and verify whether there has been an increase in balance due to the matching solution (Blackwell et al., 2009: 531).

Then, the output shows additional unidimensional measures of imbalance. The first column, labelled *L1*, reports the statistics λ^1 computed for each variable separately. The second column, *mean*, reports the difference in means between the treatment and control groups. The remaining columns report the difference in the empirical quantiles of the distributions of the two groups for the 0th, 25th, 50th, 75th, and 100th percentiles for each variable (Fig. 4.6).

```
. imb age educ black hispan, treatment(treat)
(using the scott break method for L1 distance)

Multivariate L1 distance: .89338487

Univariate imbalance:

             L1      mean       min       25%       50%       75%       max
   age   .34379   -7.409         1        -4        -6       -13        -7
  educ   .43776   -1.6816        4        -2        -1        -1        -2
 black   .76971    .76971        0         1         1         1         0
hispan   .01258   -.01258        0         0         0         0         0
```

Fig. 4.5 The output of the *imb* command

```
. cem age educ black hispan, treatment(treat)

Matching Summary:
-----------------
Number of strata: 495
Number of matched strata: 73

                0      1
      All   15992    185
  Matched    4942    183
Unmatched   11050      2


Multivariate L1 distance: .34363655

Univariate imbalance:

               L1      mean       min       25%       50%       75%       max
   age     .14045    .06542         1         0         0         1        -1
  educ     .03644   -.03644         0         0         0         0         0
 black   5.4e-15   6.3e-15         0         0         0         0         0
hispan   3.2e-15   4.6e-16         0         0         0         0         0
```

Fig. 4.6 The output of the *cem* command

Having obtained our baseline reference λ^1 value for the unmatched data, we apply the CEM algorithm by calling the *cem* command. Crudely put, CEM (1) begins with the covariates X and makes a copy X^*, (2) coarsens X^* according to user-defined cut-points (or CEM's automatic binning algorithm), (3) creates one stratum per unique observation of X^* and places each observation in a stratum, and (4) assigns these strata to the original data, X, and drops any observation whose stratum does not contain at least one treated and one control unit. Note that (4) may drop both treated and control units, thus changing the *estimand*. However, it does it transparently. Obviously, fewer strata will result in more heterogeneous observations within the same stratum and thus higher imbalance and vice versa (Blackwell et al., 2009: 527).

According to this basic coding, *cem* performs an automated coarsening. The output provides a small table reporting the number of observations in total (*All*), matched and unmatched by treatment group. Notably, two treated observations have been discarded because there were no good matches (thus, the *estimand* is changed).

Then, the output provides information about the imbalance in the matched data. The imbalance in the preprocessed data set is equal to 0.343, which means that the common ground between treated and control units is equal to 66%. Since our baseline reference λ^1 value for the unmatched data is 0.893, this matching solution increases the balance between the two groups. Note that *cem* also generates weights (stored in *cem weights*) for use in the subsequent analysis (Fig. 4.7).

As anticipated, the added value of *cem* is that it allows the researcher to set the coarsening for each variable such that substantively indistinguishable values are grouped together. For example, the code below asks *cem* to match all binary

```
. cem age (19.5 24.5 34.5 44.5) educ black hispan, treatment(treat)
(using the scott break method for imbalance)

Matching Summary:
-----------------
Number of strata: 188
Number of matched strata: 47

                0      1
      All   15992    185
  Matched    7781    185
Unmatched    8211      0


Multivariate L1 distance: .43109143

Univariate imbalance:

             L1       mean        min        25%        50%        75%        max
   age     .22288   -.53236          1          0          0         -2         -7
  educ      .0274    -.0274          0          0          0          0          0
 black   4.0e-15  -5.7e-15          0          0          0          0          0
hispan   1.1e-15  -3.3e-16          0          0          0          0          0
```

Fig. 4.7 The output of the *cem* command with specific coarsening

```
. reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75 [iweight=cem_weights]

      Source |       SS           df       MS      Number of obs   =     7,965
-------------+----------------------------------   F(11, 7953)     =    707.33
       Model |  2.9823e+11         11  2.7112e+10   Prob > F        =    0.0000
    Residual |  3.0488e+11      7,953  38334972.2   R-squared       =    0.4945
-------------+----------------------------------   Adj R-squared   =    0.4939
       Total |  6.0311e+11      7,964  75729411.4   Root MSE        =    6191.1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |    1499.672   473.9449     3.16   0.002     570.6154    2428.728
         age |   -12.28058   11.1687    -1.10   0.272    -34.17417    9.613014
        educ |    214.2673   48.6097     4.41   0.000     118.9796    309.5551
       black |   -1110.799   238.654    -4.65   0.000    -1578.624   -642.9746
      hispan |    375.2776   366.6572     1.02   0.306    -343.4666    1094.022
     married |   -1135.783   166.2893    -6.83   0.000    -1461.753   -809.8118
   nodegree |   -41.36208   215.1226    -0.19   0.848    -463.0588    380.3346
        re74 |    .2799715   .0180831    15.48   0.000     .2445239    .3154191
        re75 |    .5133666   .0183447    27.98   0.000     .4774062     .549327
         u74 |    15.95361   239.9555     0.07   0.947    -454.422     486.3293
         u75 |   -379.1638   243.8983    -1.55   0.120    -857.2685    98.94082
       _cons |    2951.233   734.1814     4.02   0.000     1512.044    4390.421
------------------------------------------------------------------------------
```

Fig. 4.8 OLS regression with *cem* weights

variables and education exactly and *age* according to standard labor force classes (i.e. 15–19, 20–24, 25–34, 35 and over).

This matching solution differs from that resulting from the automated approach: the balance is worse (from 0.343 in the automated preprocessed data set to 0.431 in the data set preprocessed according to user choices), but all the treated units have been matched. Since we have not achieved a perfect balance between treatment and control groups, it a good idea to adjust for the remaining imbalance via a statistical model. This can be done by taking advantage of the *cem weights* (Fig. 4.8).

By running the initial OLS regression on the reweighted data, the treatment effect estimate suggests that being exposed to the subsidized work experience increased earnings in 1978 by $1,499 with a 95% confidence interval of [571; 2,428]. Thus, the OLS estimate on the *cem* reweighted data is quite close to the experimental target answer ($1,794 with a 95% confidence interval of [551; 3,038]).

4.5 Conclusion

This chapter discussed the necessary assumptions for statistical correlation to justify a causal interpretation when, as is usually the case in practice, controlled randomization is unfeasible or politically sensitive and there are no convincing natural experiments providing a substitute for randomization.

First, the chapter recognized that in observational studies, causal inference is always hazardous due to the strong assumption of selection on observables, which is not easily testable by looking at the raw data (see Oster, 2019 on evaluating OLS robustness to the omitted variable bias). The chapter clarified that, ultimately, the reliability of the estimates obtained by preprocessing the raw data depends on the validity of the selection on observables assumption, which should be discussed on a case-by-case basis by the researcher. Simply put, once you have identified a set of covariates X_i, you should ask yourself whether there are additional unobservable variables capable of pushing units into treatment. If the answer is "No," then the assumption of selection on observables is theoretically met and matching and weighting procedures may credibly help you in finding out causal relationships.

Second, the chapter endorsed the practice of preprocessing the raw data through weighting and matching techniques in order to generate well-balanced samples and then applying the same familiar methods of estimation the researcher would have used anyway on the original data set, without preprocessing. In fact, even if these implementation steps do not overcome the selection on observables assumption (i.e. even if your answer to the previous question is "Yes"), weighting and matching techniques will reduce model dependence for the subsequent estimation of the treatment effect via parametric analysis. This means that effect estimates become far less sensitive to seemingly arbitrary choices in model specification: if the treatment and control groups are well balanced, slightly different model specifications are less likely to alter the substantial empirical conclusion of the analysis. Thus, preprocessing the raw data through weighting and matching techniques to generate well-balanced samples is strongly suggested. In this regard, remember that CEM may discard treated units, while EB leaves the *estimand* unchanged. Even if dropping unmatched treated units can be beneficial (Iacus et al., 2009), also this choice should be openly discussed on a case-by-case basis by the researcher: for example, dropping a treated respondent in a survey may be easier to justify than dropping an entire geographical region.

The hands-on section provided practical guidance for the implementation of the EB and CEM algorithms, respectively. This exercise was performed on the well-known LaLonde (1986) data set, a lucky case in which we know the "true" average treatment effect from an RCT and we have to match or weight the observations and to adjust the model specification so that the estimation becomes as close as possible to the experimental result (see also Costalli & Negri, 2021 for the application of CEM to the evaluation of the effectiveness of peacekeeping missions in the Bosnian civil war).

This is not what usually happens in practice. Since researchers do not know the "true" average treatment effect, they face several decisions during the implementation of the statistical analysis, and there are not always rules of thumb to be applied. The most desirable feature of the implementation steps suggested here is that they force researchers to take the assumptions that have to be met out of the shadows and make them explicit before looking at the outcomes.

Several things may go wrong. For example, researchers may miss a higher dimensional aspect of imbalance when checking lower dimensional summaries. This may affect the estimates. However, since this may also happen without preprocessing, following the steps suggested here should at least not make things worse. Moreover, when the preprocessing implies the loss of some treated unit, researchers should openly discuss the consequences in terms of external validity.

Lastly, as with the techniques covered in Chaps. 3 and 5, the research design discussed here are suitable for establishing a causal relationship between a given variable of interest, the treatment, and an outcome variable, while controlling for confounders. The implementation steps described here are not designed to investigate the paths linking a factor of interest to the outcome (see Chap. 6), to identify the full set of conditions under which the positive outcome is observed (see Chap. 7) or the mechanisms (see Chap. 8) behind the uncovered effects. While recognizing these limitations, these implementation steps help researchers in evaluating whether they are meeting the necessary conditions for generating valid inferences in their applications or how far they go. Good luck with your applied research.

Review Questions

1. Discuss the reasons why statistical association is not a sufficient, but still a necessary, condition to make a causal claim.
2. Formalize the causal inference identification problem through the lens of the potential outcomes framework and discuss it.
3. Do matching procedures overcome the inferential problems related to the selection on observables assumption?
4. What are the differences between exact and approximate matching procedures? List the aforementioned four approximate matching procedures based on the propensity score and describe two of them.
5. Why can the propensity score solution to the curse of dimensionality be seen as a tautology?
6. Once treated units have been matched to control units according to one among the available matching algorithms, is it correct to estimate the causal effect

through a simple difference in means between the observed outcomes of the treatment and control groups?
7. Compare EB and CEM preprocessing techniques by highlighting how they, respectively, address the propensity score tautology.
8. Define the following keywords:

- Confirmation bias
- Selection on observables
- Model dependence
- Common support
- Propensity score
- Balance

Replication Material

- Data and replication materials for Section 4.4 are available at https://github.com/FedraNegri/CorrelationIsNotCausationYet-.git

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press.

Arceneaux, K., Gerber, A. S., & Green, D. P. (2006). Comparing experimental and matching methods using a large-scale voter mobilization study. *Political Analysis, 14*(1), 37–62.

Atkinson, R. L., Atkinson, R. C., Smith, E. E., Bem, D. J., & Nolan-Hoeksema, S. (1996). *Hilgard's introduction to psychology* (12th ed.). Harcourt Brace Jovanovich.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*, 2084–2106.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies* (Vol. 5, pp. 43–59). Sage Publications.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*, 358–377.

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *The Stata Journal, 9*(4), 524–546.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295–313.

Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of Royal Statistical Society, Series A, 128*(2), 234–265.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Houghton Mifflin.

Costalli, S., & Negri, F. (2021). Looking for twins: How to build better counterfactuals with matching. *Italian Political Science Review/Rivista Italiana Di Scienza Politica, 51*(2), 215–230.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*, 187.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151–161.

Goldberger, A. (1991). *A course in econometrics*. Harvard University Press.
Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impacts of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data*. Cambridge University Press.
Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.
Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–294.
Heinmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46.
Heinmueller, J., & Xu, Y. (2013). ebalance: A Stata package for entropy balancing. *Journal of Statistical Software, 54*(7), 1–18.
Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health, 108*, 616–619. https://doi.org/10.2105/AJPH.2018.304337
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199–236.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.
Iacus, S. M., & Porro, G. (2009). Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics, 24*, 163–185.
Iacus, S. M., King, G., & Porro, G. (2008). *Matching for causal inference without balance checking*. http://gking.harvard.edu/files/cem.pdf
Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened exact matching software. *Journal of Statistical Software, 30*(9) http://gking.harvard.edu/cem
Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106*(493), 345–361.
Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*, 1–24.
Iacus, S. M., King, G., & Porro, G. (2019). A theory of statistical inference for matching methods in causal research. *Political Analysis, 27*(1), 46–68.
Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(September), 854–866.
Imai, K., King, G., & Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A, 171*, 481–502.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*, 4–29.
Imbens, G. M., & Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. *NBER Working Paper No. 14251*. http://www.nber.org/papers/w14251
Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis, 23*, 313–335.
Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: Quantitative methods and practices*. World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/2693 License: CC BY 3.0 IGO.
King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 435–454.
King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis, 14*, 131–159.
King, G., & Zeng, L. (2007). Detecting model dependence in statistical inference: A response. *International Studies Quarterly, 51*, 231–241.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.
LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review, 76*, 604–620.
Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.
Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.
Martini, A., & Sisti, M. (2009). *Valutare il successo delle politiche pubbliche*. Il Mulino.
Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics, 6*, 5307–5368.
Miles, J., & Shevlin, M. (2001). *Applying regression & correlation. A guide for students and researchers* (pp. 113–135). Sage Publications.
Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.
Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics,* 37(2), 187–204.
Pearl, J. (2009a). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Pearl, J. (2009b). Letter to the editor. *Statistics in Medicine, 28*, 1415–1416.
Robins, J. M., & Rotnitzky, A. (2001). Comment on the Peter J. Bickel and Jaimyoung Kwon, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica, 11*, 920–936.
Rosenbaum, P. R. (1984). The consequences of adjusting for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, 147*, 656–666.
Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.
Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika, 91*(1), 153–164.
Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician, 59*(2), 147–152.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 6*, 688–701.
Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34–58.
Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*(December), 169–188.
Rubin, D. B. (2010). On the limitations of comparative effectiveness research. *Statistics in Medicine, 29*(19), 1991–1995.
Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika, 79*, 797–809.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics, 52*, 249–264.
Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12*, 487–508.
Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27*, 325–353.
Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Suggested Readings

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Heinmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46.

Iacus, S. M., King, G., & Porro, G. (2019). A theory of statistical inference for matching methods in causal research. *Political Analysis, 27*(1), 46–68.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis, 23*, 313–335.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.

Chapter 5
Getting the Most Out of Surveys: Multilevel Regression and Poststratification

Joseph T. Ornstein

Abstract Good causal inference requires good measurement; even the most thoughtfully designed research can be derailed by noisy data. Because policy scholars are often interested in public opinion as a key dependent or independent variable, paying careful attention to the sources of measurement error from surveys is an essential step toward detecting causation. This chapter introduces multilevel regression and poststratification (MRP), a method for adjusting public opinion estimates to account for observed imbalances between the survey sample and population of interest. It covers the history of MRP, recent advances, an example analysis with code, and concludes with a discussion of best practices and limitations of the approach.

Learning Objectives
By the end of this chapter, you will be able to:

- Explain the motivation for MRP and the circumstances under which it is appropriate to implement.
- Describe the two steps in producing MRP estimates: model fitting and postsratification.
- Generate MRP estimates by adapting the provided sample code.
- Implement more sophisticated variants of MRP, including stacked regression and postratification (SRP) or multilevel regression and synthetic poststratification (MrsP) where appropriate.

J. T. Ornstein (✉)
Department of Political Science, University of Georgia, Athens, GA, USA
e-mail: jornstein@uga.edu

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_5

5.1 Introduction

The book you are reading is a testament to the "credibility revolution" in the social sciences (Angrist & Pischke, 2010), a wide-ranging effort spanning multiple disciplines to develop credible, design-based approaches to causal inference. It is difficult to overstate the influence this revolution has had on empirical social science, and the increasing emphasis that policymakers place on informing policy with good research design is a welcome trend.

But as the ongoing replication crisis in experimental psychology (Button et al., 2013) has made clear, good research design alone is insufficient to yield good science. After all, double-blind randomized control trials are the "gold standard" of credible causal inference, but small sample sizes and noisy measurement have created a situation where many published effect estimates fail to replicate upon further scrutiny (Loken & Gelman, 2017). To confidently detect causation, one needs both good research design *and* good measurement.

Often policy researchers are interested in public opinion on some issue, either as an independent or dependent variable. But the surveys we use to measure public opinion are frequently unrepresentative in some important way. Perhaps their respondents come from a convenience sample (Wang et al., 2015), or non-response bias skews an otherwise random sample. Or perhaps the data is representative of some larger population (i.e., a country-level random sample) but contains too few observations to make inferences about a subgroup of interest. Even the largest US public opinion surveys do not have enough respondents to make reliable inferences about lower-level political entities like states or municipalities. Conclusions drawn from low frequency observations – even in a large sample survey – can be wildly misleading (Ansolabehere et al., 2015).

This presents a challenge for researchers: how to take unrepresentative survey data and adjust it so that it is useful for our particular research question. In this chapter, I will demonstrate a method called *Multilevel Regression and Poststratification* (MRP). Using this approach, the researcher first constructs a model of public opinion (multilevel regression) and then reweights the model's predictions based on the observed characteristics of the population of interest (poststratification). In the sections that follow, I will describe this approach in detail, accompanied by replication code in the R statistical language.

As we will see, the accuracy of our MRP estimates depends critically on whether the first-stage model makes good out-of-sample predictions. The best first-stage models are *regularized* (Gelman, 2018) to avoid both over- and underfitting to the survey data. Regularized ensemble models (Ornstein, 2020) with group-level predictors tend to produce the best estimates, especially when trained on large survey datasets.

5.2 How It Works

MRP was first introduced by Gelman and Little (1997), and in the subsequent decades, it has helped address a diverse set of research questions in political science. These range from generating election forecasts using unrepresentative survey data (Wang et al., 2015) to assessing the responsiveness of state (Lax & Phillips, 2012) and local policymakers (Tausanovitch & Warshaw, 2014) to their constituents' policy preferences.

To demonstrate how the method works, the next section will introduce a running example drawn from the Cooperative Election Study (Schaffner et al., 2021), a 50,000+ respondent study of voters in the United States. The 2020 wave of the study includes a question asking respondents whether they support a policy that would "decrease the number of police on the street by 10 percent, and increase funding for other public services." Since police reform is a policy issue on which US local governments have a significant amount of autonomy, it would be useful to know how opinions on this issue vary from place to place without having to conduct separate, costly surveys in each area.

The problem is that even a survey as large as CES has relatively few respondents in some small areas of interest. If we wanted to know, for example, what voters in Detroit thought about police reform, a survey of 50,000 people randomly sampled from across the United States will have, on average, only 100 people from Detroit. Estimates from such a small sample will not be very precise. And more importantly, those 100 people are unlikely to be representative of the population of Detroit, since the survey was designed to be representative of the country at large.

The core insight of the MRP approach is that we can use similar respondents from similar areas – e.g., Cleveland or Chicago or Pittsburgh – to improve our inferences about public opinion in Detroit. The way we do so is to first fit a statistical model of public opinion, using both individual-level predictors (e.g., race, age, gender, education) and group-level predictors (e.g., median income, population density) from our survey dataset. Then, we reweight the predictions of the model to match the observed demographics and characteristics of Detroit. In this way, we get the most out of the information contained in our survey and produce a better estimate of what Detroit residents think than our small sample from Detroit alone could produce.

5.3 Running Example

To help demonstrate this process, we will draw a small random sample from the CES survey, and, using that sample alone, attempt to estimate state-level public opinion on police reform in each US state. In this way, we can evaluate the accuracy

of our MRP estimates and explore how various refinements to the method improve predictive accuracy. This approach mirrors Buttice and Highton (2013), who use disaggregated responses from large-scale US survey of voters as their target estimand to evaluate MRP's performance. The Cooperative Election Study data is available here, and we'll be using a tidied version of the dataset created by the R/cleanup-ces-2020.R script.[1]

```
library(tidyverse)
library(ggrepel)

load('data/CES-2020.RData')
```

This tidied version of the data only includes the 33 states with at least 500 respondents. First, let's plot the percent of CES respondents who supported "defunding" the police[2] by state.

```
truth <- ces %>%
  group_by(abb) %>%
  summarize(truth = mean(defund_police))

truth %>%
  mutate(abb = fct_reorder(abb, truth)) %>%
  ggplot(mapping = aes(x=truth, y=abb)) +
  geom_point(alpha = 0.7) +
  labs(x = 'Percent Who Support Police Reform Policy',
       y = 'State') +
  theme_minimal()
```

Oregon is the only state where a majority of respondents supported this policy proposal. And note that Fig. 5.1 likely *overstates* the percent of the total population that support such a policy, since self-identified Democrats are overrepresented in the CES sample. But nevertheless, these population-level parameters will be a useful target to evaluate the performance of our MRP estimates.

[1] All replication code and data is available on a public repository (https://github.com/joeornstein/mrp-chapter). Throughout, I will use R functions from the "tidyverse" (Wickham et al., 2019) to make the code more human readable.

[2] Obviously that phrase means different things to different people. In this case, we'll stick with the CES proposed policy of reducing police staffing by 10% and diverting those expenditures to other priorities.

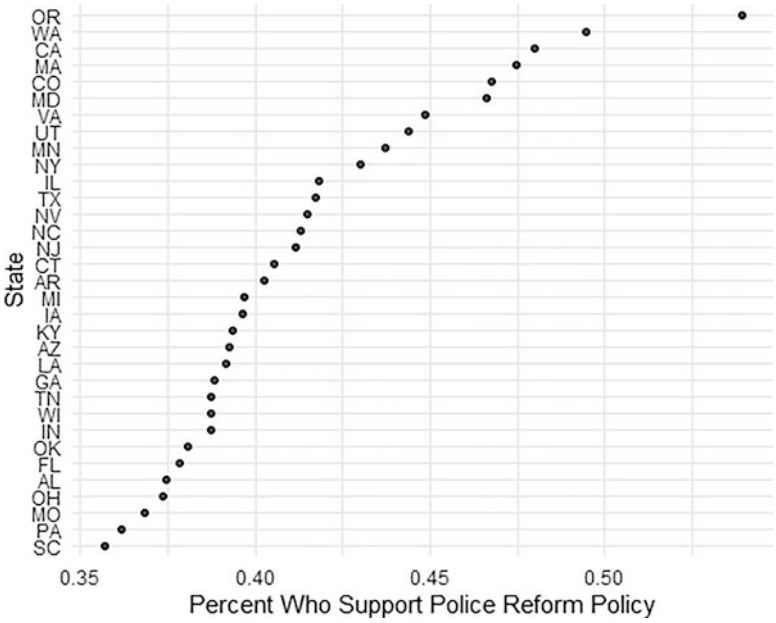

Fig. 5.1 The percent of CES respondents in each state who support reducing police budgets. These are our target estimands

5.3.1 Draw a Sample

Suppose that we did not have access to the entire CES dataset, but only to a random sample of 1,000 respondents. How good of a job can we do at estimating those state-level means?

5.3.1. Draw a Sample

```
sample_data <- ces %>%
  slice_sample(n = 1000)

sample_summary <- sample_data %>%
  group_by(abb) %>%
  summarize(estimate = mean(defund_police),
            num = n())


sample_summary
```

```
## # A tibble: 33 x 3
##    abb   estimate   num
##    <chr>    <dbl> <int>
##  1 AL       0.55     20
##  2 AR       0         4
##  3 AZ       0.438    16
##  4 CA       0.435    85
##  5 CO       0.478    23
##  6 CT       0.375     8
##  7 FL       0.402    87
##  8 GA       0.346    26
##  9 IA       0.308    13
## 10 IL       0.28     50
## # ... with 23 more rows
```

In a sample with only 1,000 respondents, there are several states with very few (or no) respondents. Notice, for example, that this sample includes only four respondents from Arkansas, of whom zero support reducing police budgets. Simply disaggregating and taking sample means is unlikely to yield good estimates, as you can see by comparing those sample means against the truth (Fig. 5.2).

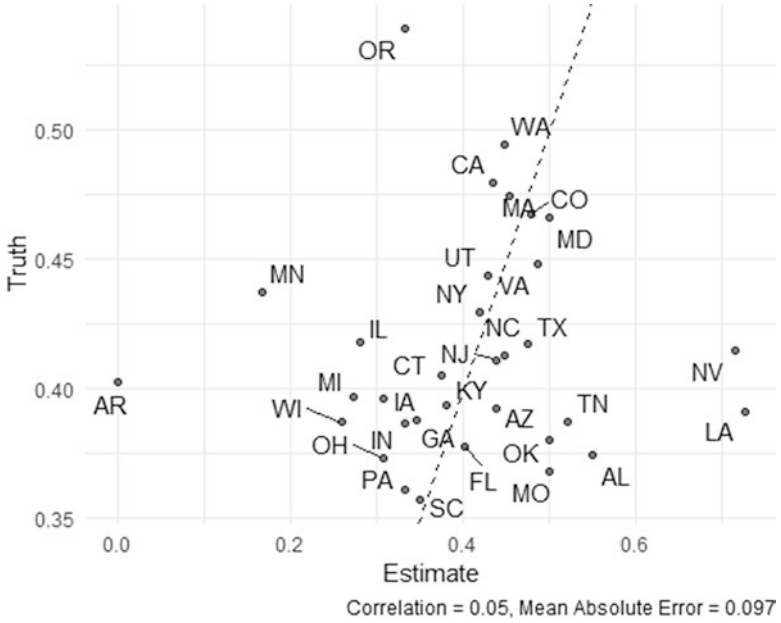

Fig. 5.2 Estimates from disaggregated sample data

```
# a function to plot the state-level estimates against the truth
compare_to_truth <- function(estimates, truth){

d <- left_join(estimates, truth, by = 'abb')

ggplot(data = d,
       mapping = aes(x=estimate,
                     y=truth,
                     label=abb)) +
  geom_point(alpha = 0.5) +
  geom_text_repel() +
  theme_minimal() +
  geom_abline(intercept = 0, slope = 1, linetype = 'dashed') +
  labs(x = 'Estimate',
       y = 'Truth',
       caption = paste0('Correlation = ', round(cor(d$estimate, d$truth), 2),
                        ', Mean Absolute Error = ', round(mean(abs(d$estimate - d$
truth)), 3)))
}

compare_to_truth(sample_summary, truth)
```

These are clearly poor estimates of state-level public opinion. The four respondents from Arksansas simply do not give us enough information to adequately measure public opinion in that state. But one of the key insights behind MRP is that the respondents from Arkansas are not the only respondents who can give us information about Arkansas! There are other respondents in, for example, Missouri, that are similar to Arkansas residents on their observed characteristics. If we can determine the characteristics that predict support for police reform using the entire survey sample, then we can use those predictions – combined with demographic information about Arkansans – to generate better estimates. The trick, in essence, is that our estimate for Arkansas will be borrowing information from similar respondents in other states.

The method proceeds in three steps.

5.3.1.1 Step 1: Fit a Model

First, we fit a model of our outcome, using observed characteristics of the survey respondents as predictors. To demonstrate, let's fit a simple logistic regression model including only four demographic predictors: gender, education, race, and age.

```
model <- glm(defund_police ~
               gender + educ + race + age,
             data = sample_data,
             family = 'binomial')
```

5.3.1.2 Step 2: Construct the Poststratification Frame

The poststratification stage requires the researcher to know (or estimate) the joint frequency distribution of predictor variables in each state. This information is stored in a "poststratification frame," a matrix where each row is a unique combination of characteristics, along with the observed frequency of that combination. Often, one constructs this frequency distribution from Census micro-data (Lax & Phillips, 2009). For our demonstration, I will compute it directly from the CES.

```
psframe <- ces %>%
  count(abb, gender, educ, race, age)

head(psframe)
```

```
## # A tibble: 6 x 6
##   abb   gender educ   race    age     n
##   <chr> <chr>  <chr>  <chr> <dbl> <int>
## 1 AL    Female 2_year Black    26     1
## 2 AL    Female 2_year Black    27     2
## 3 AL    Female 2_year Black    29     1
## 4 AL    Female 2_year Black    31     1
## 5 AL    Female 2_year Black    34     2
## 6 AL    Female 2_year Black    35     2
```

5.3.1.3 Step 3: Predict and Poststratify

With the model and poststratification frame in hand, the final step is to generate frequency-weighted predictions of public opinion. For each cell in the poststratification frame, append the model's predicted probability of supporting police defunding.

```
psframe$predicted_probability <- predict(model, psframe, type = 'response')
```

Then, the poststratified estimates are the frequency-weighted means of those predictions.

```
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))
```

Let's see how these estimates compare with the known values (Fig. 5.3).

```
compare_to_truth(poststratified_estimates, truth)
```

These estimates, though still imperfectly correlated with the truth, are much better than the previous estimates from disaggregation. Notice, in particular, that the estimate for Arkansas went from 0% to roughly 39%, reflecting the significant improvement that comes from using more information than the four Arkansans in our sample can provide.

Fig. 5.3 Underfit MRP estimates from complete pooling model

But we can still do better. In the following sections, I will show how successive improvements to the first-stage model can yield more reliable poststratified estimates.

5.3.2 Beware Overfitting

A common instinct among social scientists building models is to take a "kitchen sink" approach, including as many explanatory variables as possible (Achen, 2005). This is counterproductive when the objective is out-of-sample predictive accuracy. To illustrate, let's estimate a model with a separate intercept term for each state – a "fixed effects" model. Because our sample contains several states with very few observations, these state-specific intercepts will be overfit to sampling variability (Fig. 5.4).

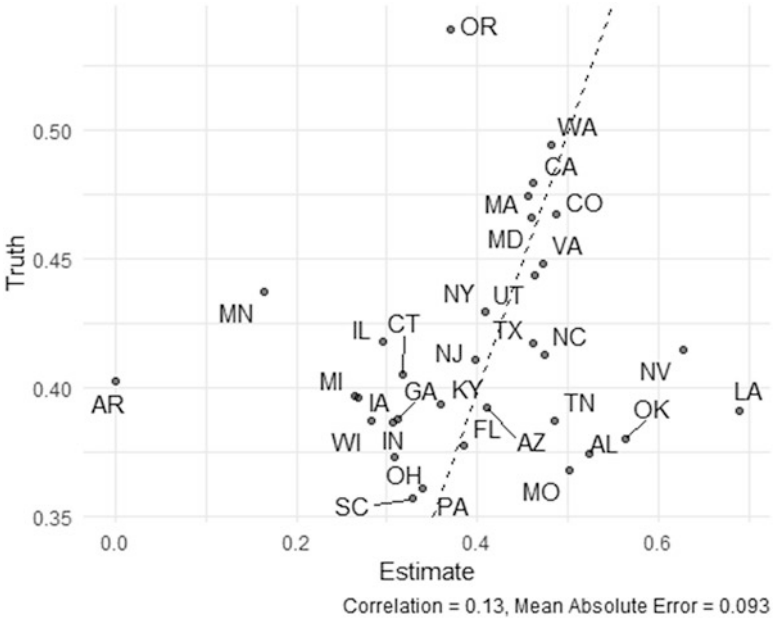

Fig. 5.4 Overfit MRP estimates from fixed effects model

```
# fit the model
model2 <- glm(defund_police ~
              gender + educ + race + age +
                abb,
            data = sample_data,
            family = 'binomial')

# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, educ, race, age)

# make predictions
psframe$predicted_probability <- predict(model2, psframe, type = 'response')

# poststratify
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))


compare_to_truth(poststratified_estimates, truth)
```

These poststratified estimates perform about as well as the disaggregated estimates from Fig. 5.2. Because each state's intercept is estimated separately, the overfit model foregoes the advantages of "partial pooling" (Park et al., 2004), borrowing information from respondents in other states. Note that the estimate for Arkansas is once again 0%.

5.3.3 Partial Pooling

A better approach is to estimate a multilevel model (alternatively known as "varying intercepts" or "random effects" model), including group-level covariates. In the model below, I estimate varying intercepts by US Census division, including the state's 2020 Democratic vote share as a covariate. The result is a marked improvement over Fig. 5.3 (particularly for West Coast states like Oregon, Washington, and California) (Fig. 5.5).

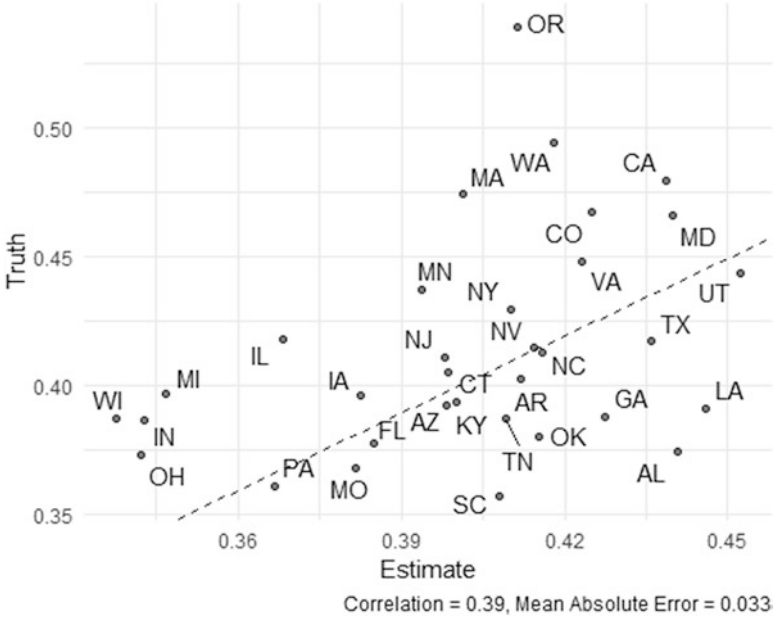

Fig. 5.5 MRP estimates from model with partial pooling

```
library(lme4)

# fit the model
model3 <- glmer(defund_police ~  gender + educ + race + age +
                (1 + biden_vote_share | division),
                 data = sample_data,
                 family = 'binomial')

# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, educ, race, age, division, biden_vote_share)

# make predictions
psframe$predicted_probability <- predict(model3, psframe, type = 'response')

# poststratify
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))

compare_to_truth(poststratified_estimates, truth)
```

5.3.4 *Sample Size Is Critical*

MRP's performance depends heavily on the quality and size of the researcher's survey sample. Up to now, we've been working with a random sample of 1,000 respondents, and though the resulting estimates are better than the raw sample means, their performance has been somewhat underwhelming. Suppose instead we had a sample of 5,000 respondents (Fig. 5.6).

```
sample_data <- ces %>%
  slice_sample(n = 5000)

# fit the model
model3 <- glmer(defund_police ~  gender + educ + race + age +
                (1 + biden_vote_share | division),
                 data = sample_data,
                 family = 'binomial')

# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, educ, race, age, division, biden_vote_share)

# make predictions
psframe$predicted_probability <- predict(model3, psframe, type = 'response')

# poststratify
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))

compare_to_truth(poststratified_estimates, truth)
```

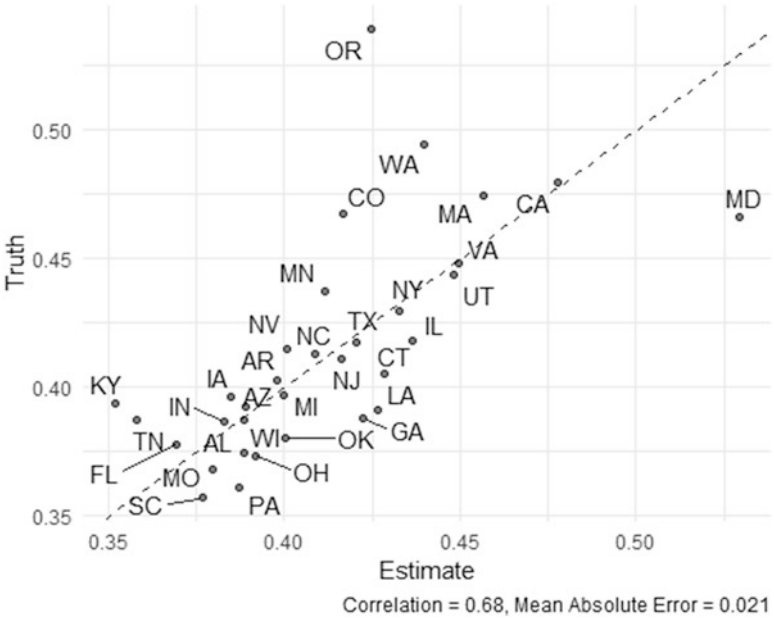

Fig. 5.6 Poststratified estimates with a survey sample of 5,000

Now MRP really shines. With more observations, the first-stage model can better predict opinions of out-of-sample respondents, which dramatically improves the poststratified estimates.

5.3.5 *Stacked Regression and Poststratification (SRP)*

Ultimately, the accuracy of one's poststratified estimates depends on the out-of-sample predictive performance of the first-stage model. As we've seen above, the challenge is to thread the needle between overfitting and underfitting. Several recent papers (Bisbee, 2019; Broniecki et al., 2022; Ornstein, 2020) have shown that approaches from machine learning can help to automate this process, particularly with large survey samples.

In the code below, I'll demonstrate how an *ensemble* of models – using the same set of predictors but different methods for combining them into predictions – can yield superior performance to a single multilevel regression model. In particular, I will fit a "stacked regression" (Breiman, 1996), which makes predictions based on a weighted average of multiple models, where the weights are assigned by cross-validated prediction performance (van der Laan et al., 2007). The literature on ensemble models is extensive, but for good entry points, I recommend Breiman (1996), Breiman (2001), and Montgomery et al. (2012) (Fig. 5.7).

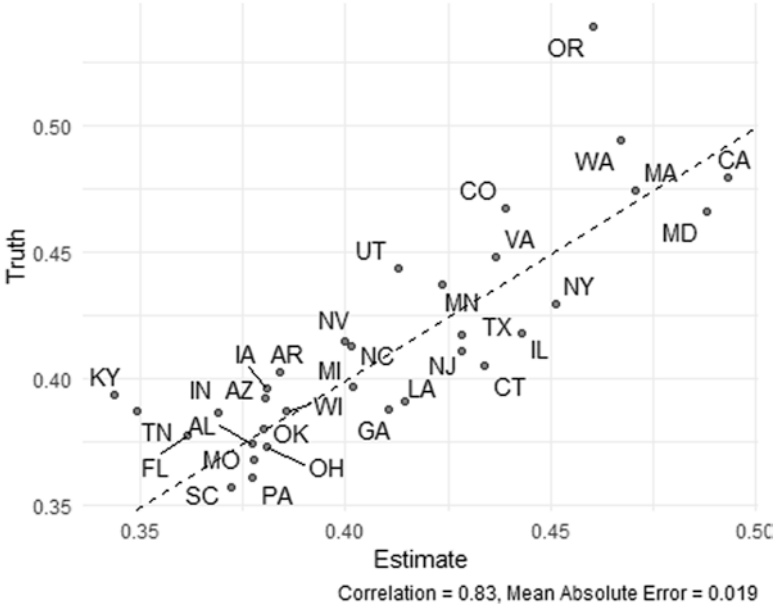

Fig. 5.7 Estimates from an ensemble first-stage model

```
# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, educ, race, age, division, biden_vote_share)

# fit the model (an ensemble of random forest and logistic regression)
library(SuperLearner)

SL.library <- c("SL.ranger", "SL.glm")

X <- sample_data %>%
  select(gender, educ, race, age, division, biden_vote_share)

newX <- psframe %>%
  select(gender, educ, race, age, division, biden_vote_share)

sl <- SuperLearner(Y = sample_data$defund_police,
                   X = X,
                   newX = newX,
                   family = binomial(),
                   SL.library = SL.library, verbose = FALSE)
# make predictions
psframe$predicted_probability <- sl$SL.predict

# poststratify
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))

compare_to_truth(poststratified_estimates, truth)
```

The performance gains in Fig. 5.7 reflect the improvement that comes from modeling "deep interactions" in the predictors of public opinion (Ghitza & Gelman, 2013). If, for example, income better predicts partisanship in some states but not in others (Gelman et al., 2007), then a model that captures that moderating effect will produce better poststratified estimates than one that does not. Machine learning techniques like random forest (Breiman, 2001) are especially useful for automatically detecting and representing such deep interactions, and stacked regression and poststratification (SRP) tends to outperform MRP in simulations, particularly for training data with large sample size (Ornstein, 2020).

5.3.6 Synthetic Poststratification

Researchers rarely have access to the entire joint distribution of individual-level covariates. This can be limiting, since there may be a variable that one would like to include in the first-stage model but cannot because it is not in the poststratification frame. Leemann and Wasserfallen (2017) suggest an extension of MRP, which they (delightfully) dub Multilevel regression and synthetic Poststratification' (MrsP). Lacking the full joint distribution of covariates for poststratification, one can instead create a *synthetic* poststratification frame by assuming that additional covariates are statistically independent of one another. So long as the first-stage model is linear additive, this approach yields the same predictions as if you knew the true joint distribution![3] And even if the first-stage model is not linear additive, simulations suggest that the improved performance from additional predictors tends to overcome the error introduced in the poststratification stage.

Here are some CES covariates that we might want to include in our model of police reform:

- How important is religion to the respondent?
- Whether the respondent lives in an urban, rural, or suburban area.
- Whether the respondent or a member of the respondent's family is a military veteran.
- Whether the respondent owns or rents their home.
- Is the respondent the parent or guardian of a child under the age of 18?

These variables are likely to be useful predictors of opinion about police reform, and the first-stage model could be improved by including them. But there is no dataset (that I know of) that would allow us to compute a state-level joint probability distribution over every one of them. Instead, we would typically only know the marginal distributions of each covariate (e.g., the percent of a state's residents that are military households or the percent that live in urban areas). So a synthetic poststratification approach may prove helpful.

To create a synthetic poststratification frame, we create a set of marginal probability distributions and multiply them together.[4]

[3] See Ornstein (2020) Appendix A for mathematical proof.

[4] The SRP package contains a convenience function for this operation (see the vignette for more information).

```
# fit the model
model4 <- glmer(defund_police ~  gender + educ + race + age +
                  pew_religimp + homeowner + urban +
                  parent + military_household +
                  (1 + biden_vote_share | division),
                data = sample_data,
                family = 'binomial')

# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, educ, race, age,
        division, biden_vote_share) %>%
  # convert frequencies to probabilities
  group_by(abb) %>%
  mutate(prob = n/sum(n))

# find the marginal distribution for each new variable
marginal_pew_religimp <- ces %>%
  count(abb, pew_religimp) %>%
  group_by(abb) %>%
  mutate(marginal_pew_religimp = n/sum(n))

marginal_homeowner <- ces %>%
  count(abb, homeowner) %>%
  group_by(abb) %>%
  mutate(marginal_homeowner = n/sum(n))

marginal_urban <- ces %>%
  count(abb, urban) %>%
  group_by(abb) %>%
  mutate(marginal_urban = n/sum(n))

marginal_parent <- ces %>%
  count(abb, parent) %>%
  group_by(abb) %>%
  mutate(marginal_parent = n/sum(n))

marginal_military_household <- ces %>%
  count(abb, military_household) %>%
  group_by(abb) %>%
  mutate(marginal_military_household = n/sum(n))
```

```
# merge the marginal distributions together
synthetic_psframe <- psframe %>%
  left_join(marginal_pew_religimp, by = 'abb') %>%
  left_join(marginal_homeowner, by = 'abb') %>%
  left_join(marginal_urban, by = 'abb') %>%
  left_join(marginal_parent, by = 'abb') %>%
  left_join(marginal_military_household, by = 'abb') %>%
  # and multiply
  mutate(prob = prob * marginal_pew_religimp *
           marginal_homeowner * marginal_urban *
           marginal_parent * marginal_military_household)
```

Then, poststratify as normal using the synthetic poststratification frame (Fig. 5.8).

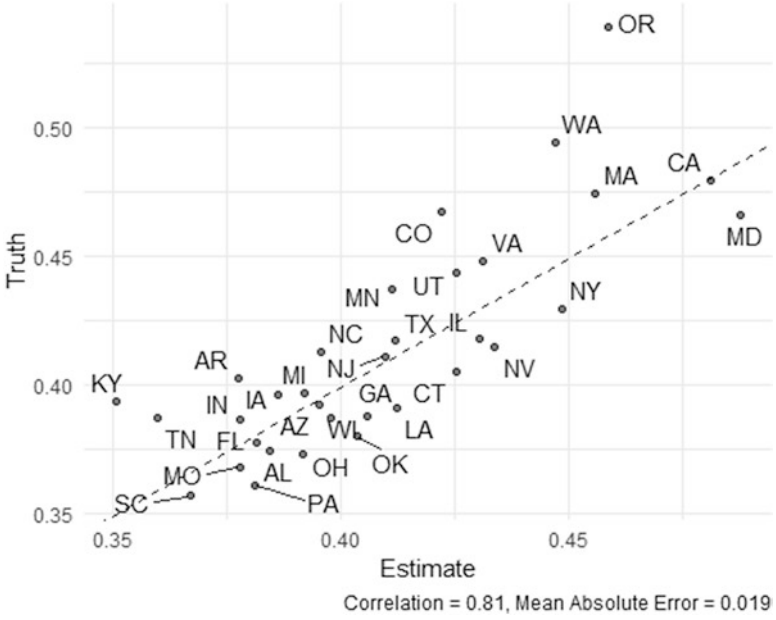

Fig. 5.8 Estimates from synthetic poststratification, including additional covariates

```
# make predictions
synthetic_psframe$predicted_probability <- predict(model4, synthetic_psframe,
                                                   type = 'response')

# poststratify
poststratified_estimates <- synthetic_psframe %>%
  group_by(abb) %>%
  # (note that we're weighting by prob instead of n here)
  summarize(estimate = weighted.mean(predicted_probability, prob))

compare_to_truth(poststratified_estimates, truth)
```

5.3.7 *Best Performing*

As a final demonstration, suppose we had access to the entire joint distribution over those covariates, *and* our first-stage model was a Super Learner ensemble. This combination yields the best-performing estimates yet (Fig. 5.9).

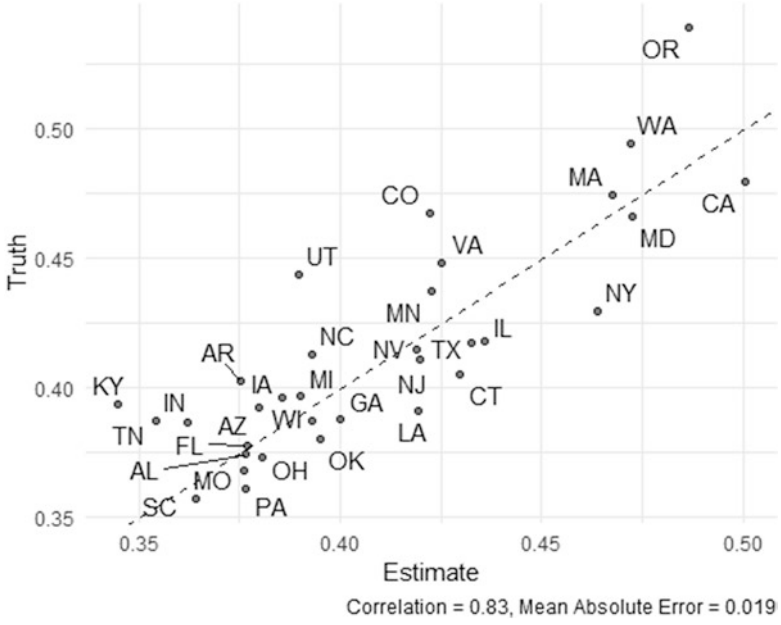

Fig. 5.9 The best performing estimates, using a large survey sample, ensemble first-stage model, and full set of predictors

```
# construct the poststratification frame
psframe <- ces %>%
  count(abb, gender, race, age, educ,
        division, biden_vote_share,
        pew_religimp, homeowner, urban,
        parent, military_household)

# fit Super Learner
SL.library <- c("SL.ranger", "SL.glm")

X <- sample_data %>%
  select(gender, race, age, educ,
        division, biden_vote_share,
        pew_religimp, homeowner, urban,
        parent, military_household)

newX <- psframe %>%
  select(gender, race, age, educ,
        division, biden_vote_share,
        pew_religimp, homeowner, urban,
        parent, military_household)

sl <- SuperLearner(Y = sample_data$defund_police,
                   X = X,
                   newX = newX,
                   family = binomial(),
                   SL.library = SL.library,
                   verbose = FALSE)

# make predictions
psframe$predicted_probability <- sl$SL.predict

# poststratify
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))

compare_to_truth(poststratified_estimates, truth)
```

The results shown in Fig. 5.9 reflect all the gains from a larger sample size, ensemble modeling, and a full set of individual-level and group-level predictors.

5.4 Conclusion

For policy researchers interested in public opinion, MRP and its various refinements offer a useful approach to get the most out of survey data. The results I've presented in this chapter suggest a few lessons to keep in mind when applying MRP to one's own research.

First, be wary of first-stage models that are underfit or overfit to the survey data. As we saw in Fig. 5.3, MRP estimates with too few predictors tend to over-shrink toward the grand mean.[5] Using such estimates to inform subsequent causal inference would understate the differences between regions. Conversely, models that are overfit to survey data (e.g., Fig. 5.4) will tend to exaggerate regional differences.

Second, new techniques like synthetic poststratification and stacked regression can help researchers manage the trade-off between underfitting and overfitting. Synthetic poststratification allows for the inclusion of more relevant predictors, and regularized ensemble models help ensure that the predictions are not overfit to noisy survey samples. The best estimates often come from combining these two approaches.

Finally, recall that the most significant performance gains in our demonstration came not from more sophisticated modeling techniques, but from more data. As we saw in Fig. 5.6, working with a larger survey yielded greater improvements than any tinkering around with the first-stage modeling choices. MRP is not a panacea, and one should be skeptical of estimates produced from small-sample surveys, regardless of how they are operationalized.

In the code above, I emphasize "do-it-yourself" approaches to MRP – fitting a model, building a poststratification frame, and producing estimates separately. But there are a now number of R packages available with useful functions to help ease the process. In particular, I would encourage curious readers to explore the *autoMrP* package (Broniecki et al., 2022), which implements the ensemble modeling approach described above and performs quite well in simulations when compared to existing packages.

Further Suggested Readings

- McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. Boca Raton: Taylor and Francis, CRC Press. (particularly chapter 13).
- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2021. *Regression and Other Stories*. Cambridge, United Kingdom: Cambridge University Press. (particularly chapter 17).

[5] In the limit, a first-stage model with zero predictors would yield identical poststratified estimates for each state, equal to the survey sample mean.

Review Questions

1. What other individual-level or group-level variables might be useful to include in the first-stage model of opinion on police reform, if they were available?
2. Why is regularization crucial for constructing good first-stage MRP models?
3. What are the benefits and potential downsides of using a synthetic poststratification frame?

References

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science, 22*(4), 327–339. https://doi.org/10.1080/07388940500339167

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Ansolabehere, S., Luks, S., & Schaffner, B. F. (2015). The perils of cherry picking low frequency events in large sample surveys. *Electoral Studies, 40*(December), 409–410. https://doi.org/10.1016/j.electstud.2015.07.002

Bisbee, J. (2019). BARP: Improving mister P using Bayesian additive regression trees. *American Political Science Review, 113*(4), 1060–1065. https://doi.org/10.1017/S0003055419000480

Breiman, L. (1996). Stacked regressions. *Machine Learning, 24*, 49–64. https://doi.org/10.17485/ijst/2016/v9i28/98380

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broniecki, P., Leemann, L., & Wüest, R. (2022). Improved multilevel regression with poststratification through machine learning (autoMrP). *The Journal of Politics, 84*(1). https://doi.org/10.1086/714777

Buttice, M. K., & Highton, B. (2013). How does multilevel regression and poststratification perform with conventional National Surveys? *Political Analysis, 21*(4), 449–467. https://doi.org/10.1093/pan/mpt017

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Gelman, A. (2018, May 19). Regularized prediction and poststratification (The Generalization of Mister p). *Statistical Modeling, Causal Inference, and Social Science (Blog)*. https://statmodeling.stat.columbia.edu/2018/05/19/

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierachical logistic regression. *Survey Methodology, 23*(2), 127–135.

Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science, 2*(June 2006), 345–367. https://doi.org/10.1561/100.00006026

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science, 57*(3), 762–776. https://doi.org/10.1111/ajps.12004

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? *American Journal of Political Science, 53*(1), 107–121. https://doi.org/10.1111/j.1540-5907.2008.00360.x

Lax, J. R., & Phillips, J. H. (2012). The democratic deficit in the states. *American Journal of Political Science, 56*(1), 148–166. https://doi.org/10.1111/j.1540-5907.2011

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science, 61*(4), 1003–1022.
Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585. https://doi.org/10.1126/science.aal3618
Montgomery, J. M., Hollenbach, F., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. *Political Analysis, 20*(3), 271–291.
Ornstein, J. T. (2020). Stacked regression and Poststratification. *Political Analysis, 28*(2), 293–301. https://doi.org/10.1017/pan.2019.43
Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis, 12*(4), 375–385. https://doi.org/10.1093/pan/mph024
Schaffner, B., Ansolabehere, S., & Luks, S. (2021). *Cooperative election study common content, 2020*. Edited by YouGov and Add your team name(s) here. https://doi.org/10.7910/DVN/E9N6PH.
Tausanovitch, C., & Warshaw, C. (2014). Representation in municipal government. *The American Political Science Review, 108*(03), 605–641. https://doi.org/10.1017/S0003055414000318
van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1).
Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Chapter 6
Pathway Analysis, Causal Mediation, and the Identification of Causal Mechanisms

Leonce Röth

Abstract This chapter presents the systematic analysis of causal mechanisms from the perspective of pathway analysis as an essential complement to conventional approaches to causation. It builds on the evidence that credible causal identification defies design-based strategies such as randomization or linear mediation analysis unless their research designs are supported by reliable mechanistic knowledge. The chapter reasons that the reliable causal identification of a mechanism requires the concept of 'natural indirect effect' and a double-nested counterfactual strategy. It discusses the empirical quantification of causal mechanisms and its underlying assumptions, offers empirical examples that clarify them, and reviews the conditions and limits of the strategy.

Learning Objectives
After studying this chapter, you will be able to:

- Understand the meaning of a mechanism from the pathway perspective.
- Learn how a counterfactual perspective on causality relates to mechanistic thinking.
- Learn how to identify and quantify causal mechanisms using non-parametric procedures.
- Understand why randomization alone does not suffice to identify causal mechanisms.
- Learn how to identify mechanisms when treatment and mediator interact.
- Understand the crucial assumptions under which indirect natural effect estimates equal identified causal mechanisms.

L. Röth (✉)
University of Cologne, Cologne, Germany
e-mail: Leonce.Roeth@uni-koeln.de

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_6

6.1 Introduction

An increasingly popular postulate of causal analysis maintains that good research includes some account of *how* one variable generates another to underpin a causal claim. Causal mechanisms are at the center of research in small-n analyses, often are a crucial part of the theoretical argument in large-n studies, and prove indispensable for scholars of systematic pathway analysis. In some accounts, a credible causal mechanism makes the difference between explanatory and non-explanatory propositions (Waldner, 2007, 146; Kiser & Hechter, 1991, 5; Mayntz, 2004, 14; Hedström, 2008).

Asking not just for a cause of an effect but also for the intermediate process in between is a deeper or second form of asking *why* (Pearl & Mackenzie, 2018, 299–300). The response to this deeper *why* always complements other types of evidence but remains crucial for qualifying the external and internal validity of causal relations. Indeed, mechanisms can raise our confidence in the established validity of a causal association – or undermine it (internal validity). Moreover, their knowledge can change the inference on evidence even from well-executed trials and improve the next experimental setup. This is because mechanisms convey information on the scope conditions of a causal association, which expose the limits of causal effects and their underlying processes (external validity). Besides, knowledge of mechanisms can reveal multiple pathways between cause and outcome, thus guiding us to more effective interventions.

A textbook illustration of these points comes from one of the earliest documented controlled experiments. In 1747, James Lind observed that eating citrus fruits prevents scurvy; understanding and validating the mechanism between citrus intake and scurvy prevention took another 183 years. In the meantime, the link from citrus to scurvy was discredited because the mechanism and its scope conditions remained unknown.[1]

The central intuition about the citrus treatment was that it involved vitamin C – a particular type of acid, later called 'ascorbic' in recognition of its scurvy preventive properties. We now know that vitamin C oxidizes when exposed to heat and light or put in contact with copper. In other words, the citrus treatment only works under specific scope conditions. Back then, however, the juice was heated for conservation, copper pipes were in widespread use, and exposure to light was regular. Thus, many attempts to produce lime juice for sea travels proved ineffective against scurvy.

Furthermore, mechanisms take time to unfold. Today we know that the intake of ascorbic acid activates the synthesis of the enzyme collagen IV. Collagen is a structural protein necessary for healthy blood vessels, muscle, skin, bone, cartilage, and other connective tissues. Ascorbic acid is required for various biosynthetic pathways; when these pathways decay, humans develop a series of symptoms

[1] The startling history of the cure for scurvy is well told in Lewis (1972). Pearl and Mackenzie (2018) recall it to illustrate mediation. This chapter's version enriches the history with some recent knowledge about the causal mechanism, and gives center stage to its scope conditions.

collectively assembled in the diagnosis of scurvy. Moreover, humans cannot synthesize collagen without ascorbic acid and have a low capacity to store it. As collagen IV synthesis stops 4–12 weeks after the last intake of ascorbic acid, symptoms of scurvy start to be visible after 4 weeks. The citrus intake also appeared ineffective for sea travels as the diffusion of steam navigation made many sea trips too short for the symptoms to show. However, Arctic expeditions remained long enough, and many seafarers suffered from scurvy in expeditions until the early twentieth century.[2]

For long, the wrong inference that citrus intake is ineffective for scurvy prevention survived due to the lack of knowledge of the mechanism of activation of collagen IV synthesis. Filling this gap proved crucial for restoring the causal association, as the mechanism disclosed many necessary scope conditions required for it to hold – namely, time, temperature, and exposure to light or copper. These conditions imply that the link between the effect of the treatment and the outcome can only be established in a study period of at least 4 weeks and if the ascorbic acid is kept intact. Moreover, they suggest that the link blurs whenever equivalent pathways are activated – for instance, if seafarers can eat raw meat or any fresh food containing sufficient ascorbic acid. Thus, perfect randomization of citrus intake may not reveal its preventive effect when its design does not take the relevant scope conditions of the mechanism into account.

In short, the knowledge of mechanisms improves three vital criteria of scientific inference – reliability and internal and external validity. But how to study mechanisms systematically?

In the following, I present the answer provided by the particular version of pathway analysis that merges graph theory with a counterfactual model of causality into a powerful framework for identifying mechanisms. This development is roughly 15 years old and still in full swing. It has taken computer science and biology by storm: biostatisticians now usually run millions of pathway models a minute to analyze gene expressions and understand the mechanisms linking a drug treatment and its effect. In comparison, social scientists still seem hesitant to embrace the many benefits that such a pathway perspective can bring. This chapter's first and foremost intention is to reduce hesitation.[3]

To this end, Sect. 6.2 locates the mechanistic why-question in the philosophy of science and discusses the assumptions under which a generic definition of a pathway or mediator[4] can be called 'a mechanism'. Then, Sect. 6.3 discusses how to distinguish between mechanistic associations and causal mechanisms. To this end, it dwells upon a remarkable strength of this method for pathway analysis – a

[2] Notably, the two expeditions of Robert Falcon Scott to Antarctica in 1903 and 1911 suffered greatly from scurvy.

[3] Excellent discussions of causal identification of mechanisms using graph theory are in Morgan and Winship (2015, Chap. 10); Pearl and Mackenzie (2018, Chap. 9); VanderWeele (2015, Part One). This chapter owes almost everything to these contributions. However, it takes a more specific angle on the causal identification of mechanisms in the social sciences.

[4] Note that, in some disciplines, the identification of mechanism is synonymous with causal mediation analysis. Here, instead, mediation is considered a special instance of pathway analysis.

graphical rendering of causal assumptions that helps to lay out the structural conditions under which pathways are causally identified or mistaken. Thus, it clarifies how the graph perspective improves on one of the most applied and cited methods in the history of the social sciences – the so-called Baron-Kenny approach to mediation analysis – and, in so doing, enhances our conditioning strategies.

Section 6.4 discusses the innovative core of pathways analysis – namely, the 'decomposition' and the quantification of the total, direct, and indirect effects on observational data. Indeed, Judea Pearl and others spearheaded a causal revolution when they defined the conditions of causally identified pathways and developed non-parametric formulae to decompose total effects into direct and indirect ones (Pearl, 2022). This quantification strategy of pathway effects took time to be accepted and faced some deep-rooted skepticism from the more conventional quarters of causal analysis (e.g., Rubin, 2004; Rubin, 2005). Nevertheless, social science scholars are slowly getting familiar with indirect effects and their underlying counterfactual theory of causation (see Imbens, 2020).

Section 6.5 replicates one influential model from development economics and sketches another from educational research. The first example demonstrates how strong supposedly mechanistic inference based on innovative cluster randomization in Kenya can be misleading. The second example shows how pathways analysis can draw important mechanistic lessons from a randomized controlled trial run in the United States to seemingly no effect. These examples prove mechanistic knowledge essential to validate and refine even causal evidence from compelling research designs.

The last section of this chapter intends to keep the promises of the pathway approach in check and dispel the illusion that causal identification is a simple technical exercise. As randomized controlled trials or instrumental variable applications show, the devil lies in the detail of the exclusion restrictions; in this respect, pathway causal identification is even more demanding than total effects via randomization or quasi-randomization. Pathway analysis reminds us that our models seldom ensure the perfect causal identification of a mechanism. Indeed, the complexity of the real world typically defies our attempts to draw exhaustive causal maps with analytic tools that require exclusion restrictions. Nonetheless, these restrictions ensure the transparent rigor that qualifies evidence as causal and distinct from mere association.

6.2 Can Pathways Be Mechanisms?

Sometimes, the concepts of mechanism, pathway, and mediation can be confusing. All three terms adhere to the general idea of increasing causal depth by diminishing the contiguity of time and space between cause and outcome. However, what exactly is considered a cause–effect framework and a mechanistic framework is subject to the relative status of a research field and is constantly in flux (see also Chap. 2, Sect. 2.3.1).

What appears to be a sufficiently deep causal mechanism in one particular research tradition and time can be perceived as a superficial association in another. Ideally, research fields increase causal depth over time and remain cautious about the trade-off between desirable specificity and useful parsimony (Craver & Kaplan, 2020). The balance of specificity and parsimony changes while research progresses, and what was considered a mechanism once might be addressed as separate cause–effect relations. Recall from the introduction that it took 183 years to detect the crucial acid for the mechanism between citrus intake and scurvy prevention. During the attempts to isolate ascorbic acid, the intake of vitamin C could have been appropriately described as the causal mechanism. In light of new knowledge, researchers today focus on way more specific biosynthesis pathways as distinct causal relationships. In short, researchers have approached the old mechanism to more causal depth. Philosophers of science call this kind of deepening process "bottoming-out" (see Fig. 6.1) or, in simpler terms, delivering on the demand for the explanation that can stop the infinite regress in causal analysis.

Aiming at fundamental explanations has had a strong appeal for a long time now in the social sciences (see Elster, 1989; Goldthorpe, 2001; Hedström et al., 1998; Hedström & Ylikoski, 2010; Knight & Winship, 2013). Nonetheless, causal mechanisms are also seen as the least understood kind of causal claim (Gerring, 2010; Hedström & Ylikoski, 2010; Waldner, 2012).

Some scholars use the term "mechanism" to refer to a series of events between the original cause and the outcome (Abell, 2004; Mahoney, 2012; Morgan & Winship, 2015; Pearl, 2009, Pearl & Mackenzie, 2018). The concept of "pathway", too, indicates a chain of mediators connecting a cause to an outcome. Thus, some have embraced the term "mechanism" for the analysis of pathways across cases (see Gerring, 2010; Imai et al. 2011; Weller & Barnes, 2014; Woodward, 2003, 350–58; Runhardt, 2015; Morgan & Winship, 2015, 325–352). Other scholars, however, try to exclusively use the term "causal mechanism" for process tracing within single cases (for example, Beach, 2017). These scholars adhere to the "process" or "physical" theories of causation that provide a substantive account of what causal processes are in light of what science tells us about the world (Dowe, 2000, 1–11 and Chap. 10).

Far from a terminological subtlety, these usages point to a fundamental divide over the concept of mechanism. The first group considers causality a matter of epistemology that can be addressed with probabilistic or counterfactual models. From this standpoint, establishing causation is an exercise in logic that many techniques

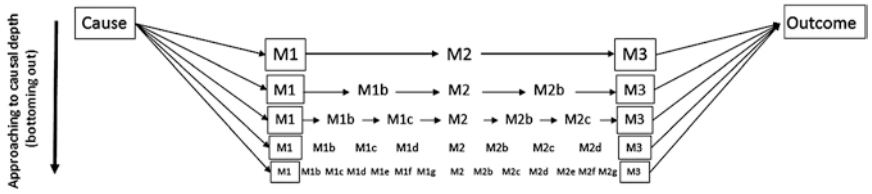

Fig. 6.1 Approaching to causal depth

can perform – provided that they afford comparisons ("type" causality; see Rohlfing & Zuber, 2021, 1634–35). In contrast, the holders of the process theory of causation maintain that causality is necessarily local – which means that it is manifest only in individual cases ("token" causality). Following the process view, within every unique case, causality exists in fine-grained sequences of entities' activities that have to satisfy the criterion of seamless productive continuity (Dowe, 2000). From the perspective of bottoming-out, the process viewpoint on mechanistic causation raises the highest possible demand on causal depth.

A pathway as a sequence of mediators (or interactions) cannot satisfy the ontological criteria established by the process view of mechanistic causation. First, seamless productive continuity can hardly be demonstrated by pathway analysis. Second, the very strength of pathway analysis lies in inferences from comparisons across cases or samples. In short, from the process view on causation, pathways do not deserve the term "mechanism". However, this reservation is a relative rarity in the social sciences. Most scholars are satisfied with an evidential view on mechanisms as a cause-to-effect pathway that at least includes one mediator. Even without satisfying the high demands from the process view, pathway analysts also approach causal depth as they want to know what connects a supposed cause and its outcome at the fundamental level, hence in a general form. As we will see in the next part, the biggest strength of pathway analysis in that ambition for deeper explanations is epistemological. Pathway analysis has developed clear and transparent criteria to distinguish causal mechanisms from mechanistic associations.

6.3 Identifying Causal Mechanisms with Graphs

Causal identification is a general problem independent of the commitment to a mechanistic theory (Pearl, 2009). Pearl's metaphor of a "ladder of causation" renders the solutions to the identification problem as a historical endeavor to more reliable causal knowledge (Pearl & Mackenzie, 2018, 23–52). In this line of thought, scientists moved from the regularity theory over probabilistic theory to the interventionist theory before reaching the top level of the counterfactual theory. As Pearl's argument goes, counterfactuals win the highest pitch as they synthesize and improve on previous solutions to causal identification problems.

From a regularity viewpoint, only the perfect sequence of the candidate cause and outcome constitutes evidence for causation. In our scurvy example, the regularity criterion requires that every citrus intake prevents scurvy without exceptions. The scope conditions of the mechanism demonstrated this bare inference mostly wrong. Under some circumstances, citrus can fail, or the causal effect might be observed without citrus. In Pearl's account, the limits of perfect regularity motivate the shift toward the probabilistic account of causality.

The probabilistic account admits that a causal relation unfolds or fails due to scope conditions and alternative mechanisms but maintains that many of them remain unknown. Hence, our best knowledge about citrus intake can focus on

whether it affects the probability of getting scurvy net of contextual vagaries – that is, on average. However, evidence that a factor affects the probability of an outcome does not constitute evidence for causation either. A limit of the probabilistic approach is that it cannot establish the direction of causation – a problem known as "asymmetry" or "endogeneity". In light of observed probability, for instance, it might also be that scurvy causes lemon intake.

The problem of asymmetry is solved when the candidate cause precedes the outcome. The best way of ensuring this order is to get some control over the candidate causal factor. So, if we prescribe citrus intake to healthy and compliant seafarers once on board, we can gather more convincing evidence of its contribution to the probability of getting scurvy. This approach is at the heart of the 'interventionist' school of causality.

With the asymmetry problem being solved, the thorniest issue of causal identification takes center stage. Even in an interventionist framework, confounders can bias the identification. Thus, we might mistake the sequence of two events as causal despite it being due to a third unobserved factor instead. Logically, the counterfactual theory of causation can discriminate between a confounded relationship and a causal one. The observed event is the real cause when it precedes the outcome, *and* its manipulation resonates with a change in the outcome that would not have occurred without the intervention. Thus, the counterfactual subsumes all preceding approaches to causal identification. Moreover, it embraces the 'would haves' and, on this basis, can offer a single theoretical solution to both asymmetry and confounding problems.

The counterfactual approach is deeply embedded in pathway analysis with graphs. Its notation responds to the problem of asymmetry by using directed arrows to clarify the direction of causality in contrast to the equal sign typical of the regression framework. Directed arrows connect "nodes" or variables in structures of dependency that recall family trees. Thus, the nodes in a path of directed arrows can be indicated as "grand-parent", "parent", "child", and "grand-child." These structures embody strong and weak causal assumptions. An arrow between two nodes indicates a weak causal assumption. It renders the direction of dependency – the fact that values of the child variable change in response to the values taken by the parent variable – but neither its sign[5] nor the size of the causal effect. The strongest causal assumption is the absence of an arrow between two nodes, as it signals that the corresponding variables take their values independently of one another. Furthermore, pathway analysts have introduced the so-called "*do*-operator" to mimic an intervention on an arrow and model the effect of its removal on observational data. This operator marks a relevant difference from conventional counterfactual studies based on non-intervention.

[5] However, some biologists introduced a distinction in the notation of the positive and the negative effects.

6.3.1 Closing the Backdoor

Graph theory offers a transparent strategy to tackle the two crucial problems of causal identification, namely, asymmetry and confounding. Figure 6.2 illustrates the task in its simplest form.

On the left-hand side of Fig. 6.2, we see the identification for the total effect framework, as in a typical correlation or regression analysis. To declare the association between X and Y causal, we first need to demonstrate that X precedes Y and not the other way around. This assumption is embodied in the direction of the arrows. The second task is to check that the association between X and Y is not confounded by third factors such as C. Path X ← C → Y is a so-called "open back-door path" and can be seen as a pipe where non-causal variance is flowing that confounds the true relationship between X and Y. Back-door paths can be closed in two ways. First, by conditioning on C. If we can hold C constant, the back-door paths between X and Y are closed, and the association between X and Y is not confounded anymore. To hold confounders constant is a common identification strategy – for example, in multivariate regressions where we regress Y on X and condition on C (Pearl & Mackenzie, 2018, 157). A second widespread approach is the randomization of X. If we assign the treatment condition of X randomly, all associations running into X are broken, and, therefore, all back-door paths are closed (compare middle part of Fig. 6.2). Experimental designs build on the randomization of the treatment. In quasi-experimental designs – such as regression discontinuity or instrumental variables – randomness in the assignment to treatment arises indirectly from natural factors or events independently of the causal channel of interest (see Chap. 3). If we can rule out both reversed causality and confounding, the associations between X and Y imply causation by necessity. The power of the back-door criterion is that it reveals under which conditions associations are causal even based on observational data.

In a mechanistic framework, the two conditions for a causal interpretation of associations are the same: X needs to precede Y, and all back-door paths between X and Y need to be closed, as on the right-hand side of Fig. 6.2. However, these conditions allow the causal interpretation of the total effect between X and Y, not the causal interpretation of the other quantities of interest to a mechanistic framework – namely, the effect of X on M (X → M, M being the mediator), and the effect of M on Y (M → Y; Y being the outcome). More conditions must be fulfilled to allow for a causal interpretation of the associations b and c on the right-hand side of Fig. 6.2.

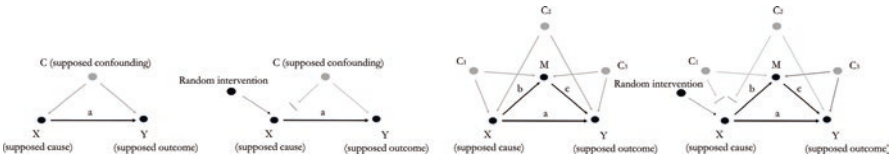

Fig. 6.2 Causal identification with and without a mechanism

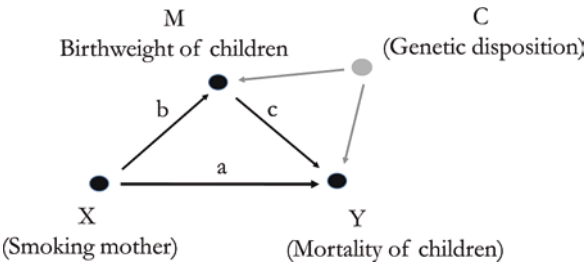

Fig. 6.3 Collider bias in mediation analysis

X has to precede M, and M has to precede Y. Furthermore, all three associations (a, b, and c) have to be un-confounded to reveal the 'true' causal effect from X → M, from M → Y, and the remaining effect of X → Y. In that framework, the total effect equals the sum of the effect from X over M to Y (the *indirect* effect) and the remaining effect of X on Y (the *direct* effect).

If we randomize the treatment X of a mediation model, the randomized treatment blocks all arrows running into X. In the example on the right-hand side of Fig. 6.2, the randomization means ruling out the confounding of C1 and C2 so that the total effect of X on Y still is the true causal effect. However, even with a randomized treatment, we are still unable to quantify the indirect effect. The reason is that C3 is left unconditioned and confounds the relationship between M and Y (path c). Randomization of the treatment does close all back-door paths running into X but does not suffice to identify mechanisms. Unfortunately, the problem of potential confounding between M and Y runs even deeper.

Figure 6.3 represents a famous causal model of the effect of smoking on child mortality. It represents precisely the constellation described on the right-hand side of Fig. 6.2 and represents a fundamental problem of mechanistic identification, the collider bias. The collider bias has troubled statisticians for centuries and led to uncountable false inferences, the birth-weight paradox just being a prominent example.[6]

Let us consider the example in Fig. 6.3. In the mid-1960s, Jacob Yerushalmy pointed out that smoking during pregnancy seemed to benefit the health of children if the baby happened to be born underweight – the so-called "birth-weight paradox" (see Yerushalmy, 1971).[7] Until 2006, this paradox remained unexplained.

In an extensive data set, Yerushalmy found unexpected relationships. Babies of smokers were lighter than babies of non-smokers. However, within the group of low-birth-weight babies, the babies of smoking mothers had a better survival rate than those of non-smokers. It was as if the mother's smoking had a protective effect within the group of babies being born underweight. The inference was that "there is no causal path from smoking to mortality" (Yerushalmy, 1971). How come?

Yerushalmy's findings are the consequence of a problematic conditioning strategy. He was unaware of the importance of genetic disposition and operated under

[6] It likely was Barbara Burks who first modeled the problem using causal graphs in 1926.

[7] An excellent discussion of the birthweight paradox can be found in Wilcox (2006).

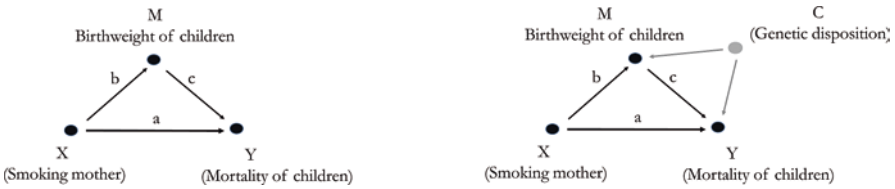

Fig. 6.4 Collider bias in mediation analysis

the assumption of the left model in Fig. 6.4. However, even within that model, it does not make sense to condition on birthweight. Birthweight is not a confounder, but a mediator. Conditioning on the mediator means correcting for the variance that runs through it. In the example, it means controlling for the *indirect* effect of birthweight. The remaining effect of X on Y is typically seen as the *direct* effect.

Conditioning on a mediator is justified to separate the indirect effect (X → M → Y) from the direct one (X → Y). As such, it lies at the heart of the conventional mediation analysis. Indeed, conventional mediation analysis compares effect estimates of the cause based on two separate regressions. The crucial difference runs between the estimate of the coefficient of X on Y in a model without a mediator and in one conditioned on the mediator. As an illustration, if 100% of the variance of the effect from cause X runs through mediator M, conditioning on M leads to a null coefficient of the cause. Baron and Kenny (1986) define three necessary, but not sufficient, conditions for detecting mediation along these lines[8]:

- X has to be significantly related to M.
- M has to be significantly related to Y.
- The total association between X and Y has to decrease when M is kept in the model.

This reasoning allows inferring four types of mediations based on how the effect between X on Y changes when we condition on M (see Fig. 6.5).

Conventional mediation analysis speaks of 'full mediation' when the total variance is associated with the path from X via M to Y (indirect effect), and the direct effect of X on Y leaves nothing unexplained. "Partial mediation" is inferred from a reduced direct effect of X on Y after conditioning on the mediator. "No evidence for mediation" is inferred when the conditioning on the mediator does not affect the direct effect from X on Y. Finally, "inconsistent mediation" is inferred when the adjustment on the mediator reverses the direction of the effect of X on Y.

The birth weight paradox is an instructive example of inconsistent mediation. The reason is that the most prominent factor for low birth weight is a specific genetic disposition that sorts an even higher impact on mortality than smoking. Genetic dispositions confound the path M → Y, as illustrated on the right-hand side of

[8] Note that this paper is one of the most cited papers in scientific history.

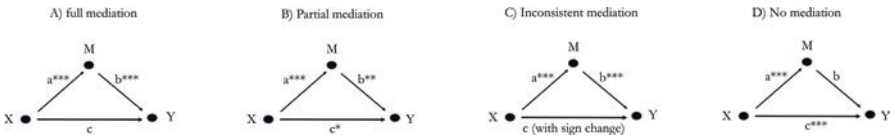

Fig. 6.5 Types of mediation. (**Note:** *** refers to the level of significance)

Fig. 6.4. It is easy to see that Yerushalmy overlooked an important confounder; what is not so easy to see is that Yerushalmy conditioned on a *collider*.

A collider is given when the same outcome depends on two different causes or, in graphical terms, when at least two arrows point to the same node. In Fig. 6.4, birthweight is a mediator (X → M → Y) and a collider (X → M ← C). Adjusting for the collider means opening a closed back-door path from X over C to Y. In other words, conditioning on birthweight creates a spurious positive association between the smoking of mothers and children's survival because genetic dispositions confound the relationship between birth weight and child mortality.

In short, Yerushalmy's surprising findings follow from this troublesome conditioning strategy. Conditioning on birth weight leads to an entirely new comparison within the stratum of children with low weight at birth. Within this new stratum, smoking mothers seem to affect babies' survival positively. However, this association is spurious. Genetic disposition has an even stronger effect on birth weight than smoking, and unless controlled for, it biases the association between birth weight and child mortality.

The graph-theoretical solution of the birth weight paradox offers at least two important lessons. First, while conditioning on confounders closes back-door paths and yields unbiased associations, conditioning on mediators and/or collider variables leads to biased associations. Second, and more important for the causal identification of mechanisms, standard mediation analysis proves unreliable. Conditioning on a collider has caused uncountable "mediation fallacies" (Pearl & Mackenzie, 2018, 315). Despite the increased awareness, the pervasiveness of the problem can still be underestimated. Indeed, mediation fallacies are not limited to the cases of inconsistent mediation. Instead, they may affect all types of conventional mediation with significant consequences. If a collider cannot be ruled out, regression-based mediation analysis cannot be trusted to produce reliable effect estimates as we cannot quantify the bias introduced by conditioning on the mediator.

Figure 6.6 illustrates a more complex causal system where we might be interested in the relative importance of pathway X → M1 → M2 → Y versus pathway X → M3 → Y. This identification task clearly falls beyond the possibilities of the regression framework and demands the more powerful approach to pathway analysis that graphs afford instead.

The overall model entails 11 variables and consists of 16 paths. The back-door criteria guide us to an effective conditioning strategy. There is no confounding between X and Y and the total effect represents the true causal effect, as we declare the causal system exhaustive. However, estimating the indirect effect of the two

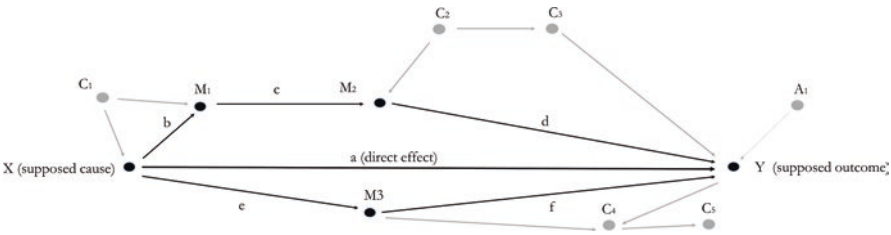

Fig. 6.6 More complex pathways

pathways of interest requires conditioning. The effect of path b is biased unless we condition on C1. The effect of path d is biased unless we condition on C2, C3, or C2 and C3 – conditioning on any of these confounders blocks the back-door path M2 ← C2 → C3 → Y effectively. A1 could be considered an alternative explanation for Y on which it is unnecessary to condition because it does not affect the quantities of interest. C4 and C5 should not be conditioned on: C4 is a collider and would open the non-active backdoor path M3 → C4 → C5 → Y; similarly, C5 should not be conditioned because of the extended collider rule that even 'descendants' of colliders, too, activate back-door paths.

The overall goal of the conditioning strategy guided by the back-door criterion is to block all the paths that generate non-causal associations between the cause and the outcome without inadvertently blocking any of the paths that generate the causal effect itself (Morgan & Winship, 2015, 109). Conditioning on C in Fig. 6.2 is a viable option whereas conditioning on M in Fig. 6.3 opens an otherwise closed back-door path. Eventually, with Morgan and Winship (2015, 109), the back-door criterion can be defined as follows:

> If one or more back-door paths connect the causal variable to the outcome variable, the causal effect is identified by conditioning on a set of variables Z if
>
> ***Condition 1*:** All back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z, which will always be the case if each back-door path
>
> (a) Contains a chain of mediation, where the middle variable is in Z or
> (b) Contains a fork of mutual dependence, where the middle variable is in Z or
> (c) Contains an inverted fork of mutual causation, where the middle variable and all of its descendants are not in Z
>
> and
>
> ***Condition 2*:** No variables in Z are descendants of the causal variable that lie on any of the directed paths that begin at the causal variable and reach the outcome variable.

However, closing the back-doors is only one of two possible identification strategies.

6.3.2 Closing the Front Door

The front-door criterion provides another interesting identification strategy derived from causal graph theory in cases where essential confounders remain unobserved. For example, let us turn to the prize-winning paper on skills and the labor market by

Glynn and Kashin (2018). Glynn and Kashin applied the front-door criterion to a well-known dataset on the effect of the Job Training Partnership Act (JTPA). The Act institutes a job training program to equip participants with different skills. The dataset contains data on the people who applied for the program, whether they showed up, and their earnings over 18 months. The study includes a randomized control trial (RCT) and an observational component. Figure 6.7 provides the causal graphs of the general problem (left), the example (middle), and the front-door approach (right).

The variable *signed up* records whether a person did enroll to the job training, the variable *showed up* whether the enrollee did use the services. The program can only affect the earnings if users showed up, so the absence of a direct arrow between *signed up* to *earnings* can be easily justified. In other words, the entire effect is mediated. Let us say cause, outcome, and mediator are all affected by the general motivation of an applicant, but unfortunately, we have not measured motivation. In a causal graph, an unmeasured variable is typically depicted by a hollow node.

The logic of the front door is to block all paths running into M – in other words, to shield the mediator. In the example of Fig. 6.7, we might randomly call applicants off and compare the randomly canceled applicants with those given real training. With all front-door paths being closed, the estimates of paths b and c can be calculated and are unbiased by definition. In that example, absent a direct effect, the indirect effect equals the total effect, and the estimate using the front-door equals the estimate based on the randomization of X. Glynn and Kashin compared the front-door predictions with those from a randomized controlled experiment, and found the results very similar (Glynn & Kashin, 2018).

The front-door approach could remove almost all of the bias introduced by the omission of the confounder of motivation. In contrast, a simultaneous estimation using the back-door without the possibility of conditioning on motivation showed substantial differences to both the experimental results and the front-door approach (Glynn & Kashin, 2017, 2018).

With Morgan and Winship (2015, 333–334), the front-door criterion can be defined as follows:

> If one or more unblocked back-door paths connect a causal variable to an outcome variable, the causal effect is identified by conditioning on a set of observed variables, M, that make up an identifying mechanism if

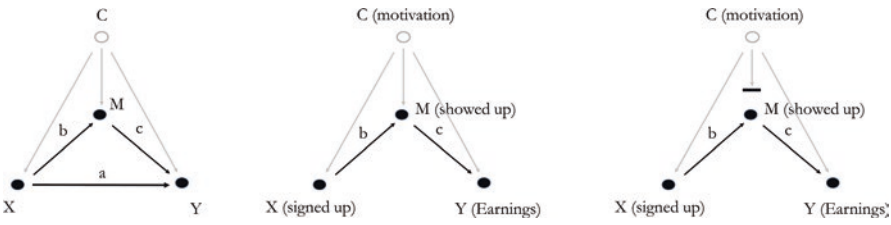

Fig. 6.7 How to shield a mediator

Condition 1 (*exhaustiveness*): The variable in the set M intercepts all directed paths from the causal variable to the outcome variable.

and

Condition 2 (*isolation*): No unblocked back-door paths connect the causal variable to the variables in the set M, and all back-door paths from the variables in the set M to the outcome variable can be blocked by conditioning on the causal variable.

At this point, we have learned two different ways to identify causal mechanisms. By definition, closing all back-door paths or closing all front-door paths leads to causal estimates even with observational data. The logic of back-door paths explains why the identification of indirect effect is neither ensured by the randomization of the cause nor by conditioning on the mediator as applied by conventional regression-based mediation analysis. The next section discusses how indirect and direct effects can nonetheless be identified.

6.4 Identifying Indirect Effects

For a long time, mediation analysts defined:

$$\text{Total Effect} = \text{Direct Effect} + \text{Indirect Effect}$$

This formula understands the indirect effect as a residual category. The Baron-Kenny approach (1986) is entirely built upon this logical pillar. As a straightforward consequence, the conventional approach advised conditioning on the mediator to arrive at the direct effect and, in force of the composition assumption, calculating the indirect effect of mediation as the total minus the direct effect.

The first problem, as already seen, is that the composition stands if M and Y are not confounded or, in other words, if a collider bias can be ruled out. The second problem is that the estimate of the residual is only credible in strictly linear systems. Once we relax the linearity assumption, the composition rule fails (Pearl & Mackenzie, 2018, 322–336).[9]

6.4.1 Indirect Effect in Non-linear Systems

The language of indirect, direct, and total effects evolved in the 1970s, but only recently was the indirect effect defined in causal terms. This shift entailed embracing counterfactual thinking.

[9] The problem of conventional mediation analysis is very fundamental. Mediation analysis based on the difference methods (Baron & Kenny, 1986; Judd and Kenny, 1981) and linear regression models suffer from problems in the presence of interactions, non-linearities, binary outcomes, unobserved confounders, and other modeling complications (see Shpitser, 2013).

Let us start with the direct effect using the *do*-calculus. In the simple graph of treatment (X), mediator (M), and outcome (Y), we get the direct effect of X on Y when we intervene on X without allowing M to change. We *do*(M = 0) and randomly assign units to *do*(X = 1) or *do*(X = 0). We call this the 'controlled direct effect' or CDE.

CDE(0) raises when we force the mediator to take on the value of zero and can be computed as

$$CDE(0) = Pr\left(\mathrm{Y} = 1 \mid do\left(\mathrm{X} = 1\right), do\left(\mathrm{M} = 0\right)\right) - Pr\left(\mathrm{Y} = 1 \mid, do\left(\mathrm{X} = 0\right) \mid, do\left(\mathrm{M} = 0\right)\right)$$

Had we forced the mediator to be 1, we would have denoted the resulting controlled direct effect as CDE(1). In practice, however, this alternative strategy could prove unwise as it forces M on instances of X that are potentially implausible to observe. Moreover, inferring the direct effect from the difference between CDE(1) and CDE(0) is to infer from an over-controlled experiment.

The so-called 'natural direct effect' or NDE offers an alternative perspective. We randomize X, but let M take the value it would naturally do. The 'would' indicates that a counterfactual is required and can be calculated as follows:

$$NDE = Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 1\right)\right) - Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 0\right)\right).$$

The NDE subtracts the probability of having a positive outcome without the treatment (X = 0) under M equal to zero from the probability of having a positive outcome with the treatment (X = 1) again under null M. In short, the NDE holds the mediator constant while the treatment is forced toward specific values. Indirect effects, unlike direct effects, have no controlled version because there is no way to disable the direct path by holding some variable constant.

Indirect effects have a natural version, too, which again requires thinking in counterfactual terms. The natural indirect effect (NIE) is when we would abstain from the treatment, but allow the mediator to be present. Understanding the causal properties of the indirect effect requires a double-nested counterfactual. In formal terms, we can define the natural indirect effect as follows:

$$NIE = Pr\left(\mathrm{Y}_{\mathrm{M=M1}} = 1 \mid do\left(\mathrm{X} = 0\right)\right) - Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 0\right)\right)$$

The first term indicates the probability of a positive outcome under absent treatment and present mediator. From this quantity, we subtract the probability of the positive outcome under the 'natural' situation where both the treatment and mediator are given.

The counterfactual M1 must be computed for each observation on a case-by-case basis. This requirement places the natural indirect effect out of the experimenters' reach as they may not know the value of the mediator M1 for any particular

treatment X at the level of the individual unit. However, assuming there is no confounding between X and M as well as M and Y (i.e., ruling out the confounding and the collider bias), the NIE can still be computed on observational data. The natural indirect effect entails denying the treatment to anyone, and letting the mediator take the value it would have in the presence of the counterfactual treatment for each individual. The difference yields Pearl and Mackenzie (2018, 333) mediation formula as follows:

$$NIE = \sum_{m} \left[Pr\left(\mathrm{X} = 1\right) - Pr\left(\mathrm{X} = 0\right) \right] \cdot Pr\left(\mathrm{Y} = 1 \mid, \mathrm{X} = 0 \mid, \mathrm{M} = m\right)$$

The expression stands for the effect of X on M in the subset of the units where the mediator takes the value *m* (in square brackets) times the probability that Y = 1 when X = 0 and the mediator takes the value *m*. So formulated, the NIE exposes the source of the product-of-coefficients idea and casts the product of two non-linear effects. Moreover, this formula allows calculating what is *explained by* mediation and the percentage *owed to mediation*.

6.4.2 Indirect Effect When the Cause and the Mediator Interact

The identification of indirect effects becomes more complex when the mediator and the supposed cause (or "exposure") interact. A unified perspective on the decomposition of the total effect in a case where the independent variable of interest interacts with the mediator has been provided by VanderWeele (2014).

So far, effect decomposition has meant to split a total effect into an indirect and direct one. In the presence of exposure-mediator interaction, two components need to be added: the one due to interaction only; the other due to mediation and interaction (see VanderWeele, 2014, 751). The counterfactual assumptions to identify the effect quantities are similar to those required to analyze causal mediation without interaction. As in the case of causal mediation, indirect effects including interactions require double-nested counterfactuals, whereas the direct effect requires weaker assumptions. The attribution of the interaction quantities to either the indirect or direct effect, instead, remains an empirical question. Figure 6.8 illustrates two possible response strategies based on VanderWeele (2014, 757).

The fourfold decomposition depicted in Fig. 6.8 encompasses both decompositions for mediation and interaction.

For interaction, the reference interaction (INT*ref*) and the mediated interaction (INT*med*) combine to the portion attributable to interaction (PAI). The portion attributable to interaction (PAI) combines with the controlled direct effect (CDE) and the pure indirect effect (PIE) to give the total effect (TE).

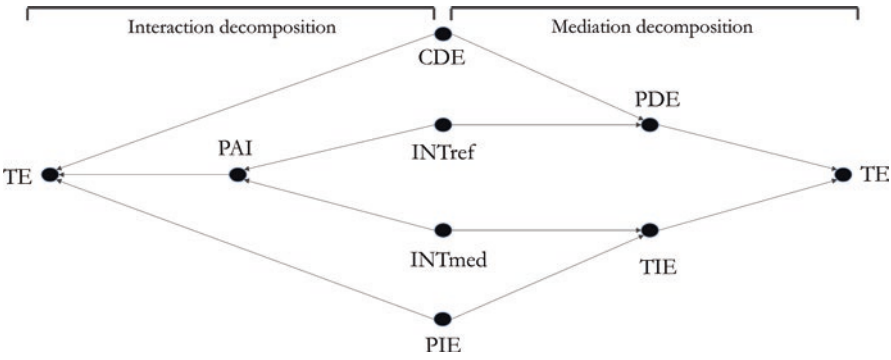

Fig. 6.8 Fourfold decomposition

For mediation, the controlled direct effect and the reference interaction (INT*ref*) combine to give the pure direct effect (PDE); the pure indirect effect (PIE) combines with the mediated interaction (INT*med*) to give the total indirect effect (TIE), and the pure direct effect (PDE) combines with total indirect effect (TIE) to give the total effect (TE).

6.4.3 Wrapping Up

The graph theory reveals that the identification of causal mechanisms requires counterfactuals. The natural indirect effect is when we abstain from the treatment, but the mediator is present. Contrasted with the state where both the treatment and the mediator are present, we can quantify how much of the effect of X on Y is captured by the mediator M, and how much of Y is owed to the mediator M alone. Such a natural indirect effect gauges a causal mechanism once the back-door criterion is satisfied, e.g., all back-door paths are closed.

The consequences of this definition are far-reaching. The identification of causal mechanisms appears as out of reach to the conventional mediation analysis than to randomization. What appears as bad news can also be a good insight, as the natural indirect effect yields a mediation formula stripped of any parametric assumptions. Under some assumptions, this formula allows quantifying the causal mechanism based on observational data. Section 6.5 demonstrates this claim with the example of a renowned identification debate.

6.5 Applications

6.5.1 A Mechanistic View on the Worm Wars

In this application case, I add a causal mediation view to the "worm wars" – a famous debate over the interpretation of influential cluster randomization in Kenya that, besides other studies, brought one of its authors, Michael Kremer, the Nobel Memorial Prize in Economic Sciences in 2019.

The study originates from the evidence that nearly two billion people worldwide – mostly children – are infected by intestinal worms. These species inhabit the human digestive tract; they spread by expelling their eggs via the body waste of infected people. Without good sanitation, these microscopic eggs can find their way, unnoticed, onto the skin or food of another person. Once someone ingests an egg, the reinfection cycle continues. Poor sanitation facilities and hygiene practices allow infections to spread locally.

In 2004, a landmark study showed that an inexpensive medication to treat parasitic worms could improve health and school attendance for millions of children in many developing countries (Miguel & Kremer, 2004). Eleven years later, a headline in *The Guardian* reported that the deworming treatment had been debunked. In 2021, a carefully exercised replication study restated the original findings (see Ozier, 2021). Why so?

Miguel and Kremer convincingly argued that, due to the infectiousness of the worms, individual treatments are unlikely to be effective because children will quickly re-infect themselves with other children. Consequently, they run an encompassing field experiment in Kenya using cluster randomization at the school level. The experiment compared more than 25,000 treated children across three waves to a control group for each wave with similar attributes except for the suppressed treatment. They found a remarkable effect of the treatment on school attendance not only in the treatment area (up to 3 km) but also in the surrounding areas (3–6 km from the treatment).

Replication analyses have mainly confirmed the direct effect in the treatment areas. However, the spillover effects became subject to debate and turned insignificant in some specifications (for example, Aiken et al., 2014). The debate about the replication involved many influential scholars, was covered by several blogs, and eventually came to be known as the "worm wars". A systematic review of the debate seemed to restore the trust in the key findings of the original study. Ozier (2021) concluded that, if anything, years of debates and replication have reinforced his belief in the main effect. In short, it appeared as if the treatment of Miguel and Kremer had indeed sorted a substantial positive impact on children's school attendance.

However, there is a second line of skepticism, less concerned with the significance levels of the total effects but with the plausibility of the indirect effect. The indirect effect, as we have learned, considers the probability of a positive outcome (school attendance) given that we do not have a treatment (no de-worming drug

intake), but we set the mediator (being, in fact, de-wormed) to the values as if we would have had treatment (de-worming drug intake). We contrast this with the probability of a positive outcome (school attendance) under natural conditions where the treatment is given (de-worming drug intake) and the mediator too (being de-wormed). Based on Pearl's mediation formulae, we can compute the natural indirect effect using observational data. The results can be given a causal interpretation if we can exclude confounding between the mediator (being de-wormed) and the outcome (school attendance).

This mechanistic perspective on the study is of great interest for at least two reasons. First, experts in deworming cast considerable doubt on the findings. Epidemiologists refused to include the paper in a meta-study for methodological reasons (no blinded treatment was performed) and referred instead to existing epidemiological studies that, if at all, showed very modest effects of deworming on school attendance. In other words, the authors of a Cochrane review were unconvinced that de-worming could have had such a substantial effect as reported in Miguel and Kremer (Taylor-Robinson et al., 2015). Second, the authors of the original experiment framed their study and their results as if they had strong evidence for the entire mechanism. In the words of the authors' abstract, "[*d*]*eworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment*." (Miguel & Kremer, 2004, 159). In short, the authors' inference is that their evidence point to a clear recommendation for subsidizing de-worming treatments because de-wormed students have a higher likelihood of attending school. Is it the de-worming via the drug intake that causes students to attend school more often?

Based on the original data, the mediation formulae can be used to put the mechanistic claim under scrutiny. Table 6.1 includes all probabilities required to compute the natural indirect, natural direct, and the total effect based on the replication data of Miguel and Kremer (2014), Miguel et al. (2014).[10] By relating indirect and direct effect quantities to the total effect, we can draw valuable conclusions. The natural indirect effect supports the suspicion of the epidemiologists. Only 1.8% of the total effect would be achieved by worm-free students alone. In contrast, 94.2% of the total effect is related to the natural direct effect of the treatment other than

[10] For the replication, I use a very simple model based on the drug treatment in the first period of the field experiment. The experiment had three waves, but the comparison groups changed during the waves and because the effect on school attendance is predominantly a result of the first wave, I focus on the first wave only. For the mediator, I use the reversed indicator of any moderate or heavy worm infection based on the WHO standard in 1999. I see the mechanism present when a treated student is indeed free of worms. For the outcome, I use a dummy of students being present in school at times of the surprise visit. The current documentation of the data is exemplary (see Miguel and Kremer, 2014; Miguel et al. 2014; Hicks and Nekesa, 2014).

Table 6.1 Probabilities of the treatment, the mechanism, the outcome and the natural direct (NDI), indirect (NIE), and total effect (NTE)

Treatment condition, mediator condition, and outcome probabilities						
Treatment	Dewormed	Present in school (in %)			Treatment	Dewormed (in %)
Yes	Yes	0.90			No	0.55
Yes	No	0.86			Yes	0.59
No	Yes	0.86				
No	No	0.85				
				Inference		
NIE	0.05	NIE/TE	1.8	1.8% of the school attendance effect would be achieved by worm-free students alone		
NDE	2.7	NDE/TE	94.2	94.2% of the attendance effect is related to the treatment other than deworming students		
TE	2.9	1-NDE/TE	5.8	5.8% of attendance effect is owed to the capacity of the treatment to deworm students		

Note: Compare equations for NIE, NDE, and TE above.

deworming students. Finally, 5.8% of the effect on attendance is owed to the capacity of the treatment to deworm students.[11]

How do we make sense of these numbers?

Humphreys (2015) documented and commented on the worm wars in close detail, driven by concerns for the mechanistic element of the study. He points to several important aspects that can be learned from the documentation of the experiment. Based on background information and the skeptical comments of epidemiologists, we might add several pathways between treatment and outcome (see Fig. 6.9). The causal graph reveals that the estimate above of the natural indirect effect is not identified. There is nothing identified in this system of pathways because too many nodes are unobserved. Let us briefly describe the pathways in Fig. 6.9.

One element of the treatment is the drug intake that seems to effectively de-worm students. The effect of de-worming alone is relatively weak, as the path analysis in Table 6.1 confirms. The drug intake has as least two more effects on attendance that cannot be isolated given the existing data. De-wormed students create spillovers, and spillovers might feedback to the treated. This feedback is problematic because it undermines the assumption of the independence of the treatment group and the control group – the problem that compelled resorting to cluster randomization in the first place.

Beyond spillovers, the drug intake can create placebo effects. Students feel better because of the drug, irrespective of being de-wormed, which might increase school

[11] An alternative way of modeling these numbers would be to use readymade packages in software such as R or Stata. In Stata, you would use the model builder and simple graph the mediation model. After the estimation of all path-coefficients, the effects can be decomposed into total, direct, and indirect effects using the teffects command (see Bollen, 1989; Sobel, 1987). Note that this command still assumes linearity and leads to biased estimates in this case.

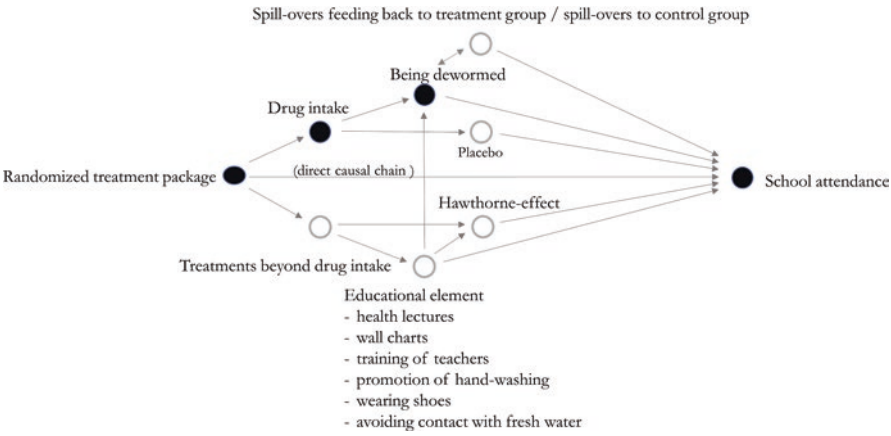

Fig. 6.9 Mechanisms in the worm wars

attendance. Since the control group was not treated with a placebo, we cannot estimate the placebo effect. More worrisome is how the research group treated the treatment group beyond the drug intake. The documentation files list health lectures, wall charts in the schools, training of teachers in the treatment schools, encouragements of the treated students for handwashing, wearing shoes, and avoiding freshwater (see Hicks & Nekesa, 2014, 7).[12] This extensive treatment had obvious health effects – including a contribution to de-worming – which suggests that the treated students likely became well aware of being subject to an encompassing treatment package. Thus, at least three more paths follow from that treatment beyond drug intake.

First, the educational elements on health issues might have affected the well-being of students besides de-worming, which raises their probability to be present in school. Second, being so obviously treated might activate the Hawthorne effect, the rising willingness of participants to make the experiment a success in light of the efforts experimenters provided for the treated. For example, teachers might just encourage students in the treatment group to show up because they know that school attendance is an important measure (although it has to be noted that the measurement of school attendance was achieved by surprise visits). Third, health education

[12] The educational treatments at the school level were part of a separate intervention of the same NGO and could in principle be controlled based on the data (see Hicks & Nekesa, 2014, 5). In fact Miguel and Kremer condition on those interventions. They write "None of these programs involved health treatments for pupils, and given the cross-cutting design, are unlikely to complicate the identification of average treatment effects across PSDP program and comparison schools." Nonetheless, in many specifications Miguel and Kremer (2004) control for assignment to assistance through these other programs'. Only a page later, they write without considering any potential bias "[t]he educational component of the intervention focused on teaching children about avoiding the disease. Health educators explained the transmission vectors for different types of helminths [one of the relevant worm types] and also promoted hand-washing, wearing shoes, and avoiding contact with fresh water" (2014, 7).

affects the likelihood of being de-wormed besides de-worming drug intake and school attendance. Accordingly, the effect of being de-wormed on school attendance, including the spillover effects, is confounded. Knowing about the direction of the influence of health education (increasing de-worming and school attendance), the already weak indirect effect of de-worming via drug-intake on school attendance is most likely biased upwards. This perspective reveals that the authors make strong mechanistic inference without ever quantifying the importance of their hypothesized mechanism and without noticing that the indirect effect cannot be precisely identified, given the observable data at hand.

Such a mechanistic perspective also reveals the standing of the main criticism of the epidemiologists. The Cochrane reviewers classified the study as very weak in terms of evidence, predominantly because of the lack of placebo treatment of the control group. Indeed, except for the spillover path, all alternative paths between treatment and outcome could have been closed by placebo treatment. The consideration also applies to the educational health elements.

Thus, the mechanistic view qualifies the inference of this landmark study substantially. First, there is a confirmation of a significant indirect effect running from the treatment over being de-wormed to higher school attendance. However, this indirect effect explains a very marginal part of the increased school attendance. Way more important are the indirect effects triggered by the entire treatment package beyond the ability to de-worm students. The rise in school attendance is predominantly a composite of different pathways from the Hawthorne pathway over the health education pathway to a potential placebo pathway, combined around 54 times more powerful for school attendance than the de-worming effect. The overall inference to recommend the distribution of cheap drugs might be replaced by the recommendation to offer supposedly more expensive health education.

To be very clear about it, the study of Miguel and Kremer is comparatively well-executed and deserves to be praised for the logic of cluster randomization alone. Nonetheless, the mechanistic view on this experiment demonstrates that randomization does not allow for mechanistic inference. While the total effect of the treatment package might still be perfectly identified, the mechanistic view helps identify which elements of the treatment have created more or less powerful pathways to the outcome. It is extremely interesting to know how much Hawthorne, placebo, or health education contributed to the substantial rise in school attendance, as such effect decomposition can help to improve similar experiments in the future. Like in the lemon-scurvy example, experimenters need to disable these alternative pathways (exclusion restriction) for getting to the correct inference.

A mechanistic view may help to understand supposedly strong effects in well-executed experiments. Moreover, it can reveal causal mechanisms where experiments seem to yield nothing.

6.5.2 A Mechanistic View on a Chicago School Reform

In 1998, US secretary of education, William Bennet, called Chicago's public school the worst of the nation. However, several reforms in the late 1990s moved them from the worst to 'innovators of the nation'.[13] One of the core reforms involved a program called 'Algebra for All', compulsory prep courses for ninth graders in high school. At first sight, the program seemed a success as math scores rose significantly. However, the qualification of incoming ninth-grade students was already improving due to changes in the K-8 curriculums (an important confounder). Once controlled for this confounder, the reform turned out to be insignificantly related to the math performance of ninth graders. Here, the story would have typically found its end.

Luckily, Professor Guanghei Hong remained curious because she knew that when Algebra for All was introduced, more than the curriculum changed. The lower-achieving students found themselves in classrooms with higher-achieving students and could not keep up. Detrimental effects for students and teachers caused by mixed classes compared to remedial classes are well-known. In short, Mrs. Hong was suspicious of the unanticipated side effects of the treatment package. Testing the classroom environment as a mediator between reform and outcome clearly showed that this pathway had negative consequences. Once taken into consideration, the direct effect turned positive. The lesson seemed clear: removing the mixed classes and keeping the prep courses was the logical consequence and created a success story of the modified Algebra for All program.

Students in Chicago significantly benefited from a mechanistic view on an education program that has, at first sight, falsely been considered a failure. We learn from this example that different mechanisms can cancel each other out ("opposing mediation" as in Kenny [1998]), which demonstrates that even a null finding based on a randomized treatment can be worth considering with closer scrutiny on the level of mechanisms. The Algebra for All example is similar to the discredited causal link between lemons and scurvy prevention, although its revitalization took place in a substantially shorter period.

6.6 Thou Shall Not Raise Causal Illusions

Scholars of pathways have revolutionized our view on causal identification. The counterfactual perspective on pathways reveals that fundamental problems of causality – asymmetry and confounding – can logically be solved by closing either the back- or the front-door. This perspective embraces conventional counterfactual causal inference such as randomization or quasi-experiments. Causal graphs help to make its logic and assumptions very transparent. Applying the logic of the

[13] One of its inventors, Arne Duncan, became secretary of education under Barack Obama.

back-door to generally defined causal mechanisms reveals two things. First, conventional approaches are ill-suited for identifying causal mechanisms as they can mistake their structure. Pathway analysis solved that issue by focussing on indirect effects. This perspective reveals that causal mechanisms can be quantified by non-parametric comparisons of observable with counterfactual probabilities. To lend these numbers a causal meaning depends on a simple assumption: path estimates in a system of pathways must be unconfounded.

This unconfoundedness can unfortunately not be fully ensured by randomization – although the randomization of the treatment helps a lot to block all paths running into the candidate cause. Moreover, causal mechanisms can only be identified if a theoretically exhaustive causal system is given and all confounders are observed and conditioned on. Based on a theoretically defined causal system, effective strategies of de-confounding can be determined. The complexity of the task becomes apparent when we remind ourselves of the problem of the collider bias. The collider bias is an instance of a single confounded path in a system of pathways, leading in the worst of events to completely misleading estimates of the indirect and direct effects – such as when smoking mothers are understood to increase the survival rate of their children. Besides, complex pathways with sequences of many mediators can complicate the identification task and the chances for false inference multiply.

The pathways perspective on the identification of causal mechanisms is logically simple. However, mechanisms can only be identified given a theoretically exhaustive causal system where all the variables required to close the back-doors are measured, free of error, and conditioned. Empirically, these assumptions are hard to meet. Thus, research relying on pathways or causal mechanisms should avoid creating the causal illusion that the back-door criterion will easily tackle identification tasks.

The greater strength of the pathway approach is not to deliver a readymade tool for causal inference but a perspective that can boost the transparency over what is needed to identify a mechanism causally. It complements standard approaches of causal inference that typically seek to identify total effects. Analyses of mechanisms searching for indirect effects ask a deeper form of why. Preliminary answers to these deeper questions can at times be very generic, such as a single mediator connecting cause and outcome, and at times can also span to very complex systems of pathways. However, even the most generic mechanism can reveal a great deal. Thinking of lemons' ability to prevent scurvy, smoking mothers to decrease the survival rate of their children, the capacity of de-worming to increase school attendance or preparation courses to improve school performance. In all examples of this chapter, evidence on a single mediator considerably qualified the inference of a cause–effect relationship.

Despite the capacity of a mechanistic view to qualify the inference of even well-executed experiments, the added values are complementary. Randomized treatments facilitate the identification of causal mechanisms because important sources of confounding are erased by design. Mechanisms, in turn, improve the exercise and

inference on well-executed experiments too. The more we know about the mechanisms, the better we can identify total effects.

Suggested Readings

There are three books of great help to understand causal mediation. The most encompassing work on causal mediation analysis, including moderated mediation, is most likely VanderWeeles' book *Explanation in causal inference: methods for mediation and interaction*, published in 2015 by Oxford University Press. Although probably the most encompassing, it addresses the issue from the perspective of biostatistics. Easier access to causal mediation can prove Chapter 9 on *Mediation: The search for a mechanism* in Pearl and Mackenzie (2018), published by Basic Books. The entire textbook can be highly recommended to cast light on recent developments in causal identification against the background of the history of statistics. Finally, Chapter 10 on *Mechanisms and causal explanation* in Morgan and Winship (2015) lies somehow in between VanderWeeles' equation-based insights and Pearl and Mackenzie's captivating narrative. Their entire book on *Counterfactuals and causal inference* can be recommended, as it covers virtually all causal identification tasks from the perspective of the social sciences while preserving a deep commitment to graph theory and counterfactual thinking.

Helpful Websites

Beyond books, there are two highly informative websites on causal mediation. The one by David Kenny provides regular updates on mediation analysis and also covered issues in causal mediation (http://davidakenny.net/cm/mediate.htm). Alternatively, Columbia University provides information on causal mediation, including a recorded lecture of VanderWeele based on the Harvard Seminar Series in Biostatistics (https://www.publichealth.columbia.edu/research/population-health-methods/causal-mediation#websites).

Software Recommendations

Causal mediation, the identification of mechanisms, or causal pathway analysis are relatively new and characterized by rapid development. Formulas, methods, and software applications change accordingly. Nonetheless, several software packages have proven extremely useful.

1. R *mediation* package (Tingley et al. 2014):
   - the *mediate()* function estimates the natural direct and indirect effects based on Pearl's mediation formula,
   - X-M interaction may be conducted by the function test *TMint()* (significant finding implies that the no X-M interaction assumption does not hold).
   - the sensitivity analysis function *medsens()* allows for investigators to examine, through simulations, the robustness of their findings to potential unmeasured M-Y confounders.

   Results for all analyses are displayed using the *summary()* and *plot()* functions

2. SAS macro:
   - The SAS macro is a regression-based approach to estimating controlled direct and natural direct and indirect effects.
   - The macro can handle virtually every distributional and link assumption (compare Valeri et al., 2013).
3. Stata:
   - *paramed* package (no sensitivity analysis) (Emsley et al., 2013).
   - *ldecomp* (no sensitivity analysis) (Buis, 2010).
   - *medeff* (sensitivity analysis) (Hicks and Tingley, 2011).
   - *gformula* (helpful in case of post-treatment and time-varying confounding) (Daniel et al., 2011).

Review Questions

1. Under which conditions can mechanisms be causally identified?
2. What is a natural indirect effect in comparison to a controlled indirect effect?
3. Why randomization might identify cause-effect relationships but not neccessarily indirect effects?
4. Why might conventional mediation analysis be misleading for the causal identification of the mechanism?
5. How does mechanistic evidence help to improve the implementation of experiments?
6. What are the consequences of treatment-mediator interactions for the identification of mechanisms?
7. What are the limits of mechanistic causal identification?

References

Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology, 30*, 287–310. https://doi.org/10.1146/annurev.soc.29.010202.100113

Aiken, A, Davey, C, Hayes, R, & Hargreaves, J. (2014). Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: A pure replication. *3ie replication paper* 3, part 1. Washington, DC: International initiative for impact evaluation (3ie). https://doi.org/10.1093/ije/dyv127.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Beach, D. (2017). What are we actually tracing? Process tracing and the benefits of conceptualizing causal mechanisms as systems. *Qualitative & Multi-Method Research, 14*(1/2), 15–22. https://doi.org/10.5281/zenodo.823306

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Buis, M. (2010). Direct and indirect effects in a logit model. *The Stata Journal, 10*(1):11–29.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Dowe, P. (2000). *Physical causation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511570650

Daniel, R. M., De Stavola, B. L., & Cousens, S. N. (2011). gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal, 11*(4), 479–517.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812255

Emsley, R. & Liu, H. (2013). "PARAMED: Stata module to perform causal mediation analysis using parametric regression models," Statistical Software Components S457581, Boston College Department of Economics.

Gerring, J. (2010). Causal mechanisms: Yes, but…. *Comparative Political Studies, 43*(11), 1499–1526. https://doi.org/10.1177/0010414010376911

Glynn, A. N., & Kashin, K. (2017). Front-door difference-in-differences estimators. *American Journal of Political Science, 61*(4), 989–1002. https://doi.org/10.1111/ajps.12311

Glynn, A. N., & Kashin, K. (2018). Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association, 113*(523), 1040–1049. https://doi.org/10.1080/01621459.2017.1398657

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review, 17*(1), 1–20. https://www.jstor.org/stable/522622

Hedström, P. (2008). Studying mechanisms to strengthen causal inferences in quantitative research. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 319–335). https://doi.org/10.1093/oxfordhb/9780199286546.003.0013

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67. https://doi.org/10.1146/annurev.soc.012809.102632

Hedström, P., Swedberg, R., Hernes, G., & (Eds.). (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511663901

Hicks, J., & Nekesa, C. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Codebooks. Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Hicks, R., & Tingley, D. (2011). mediation: STATA package for causal mediation analysis.

Humphreys, M. (2015). *What has been learned from the deworming replications: A nonpartisan view*. http://www.columbia.edu/~mh2245/w/worms.html [retrieved 01.11.2021].

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature, 58*(4), 1129–1179. https://doi.org/10.1257/jel.20191597

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789.

Judd, C. & Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*(5), 602–619.

Kiser, E., & Hechter, M. (1991). The role of general theory in comparative-historical sociology. *American Journal of Sociology, 97*(1), 1–30. https://doi.org/10.1086/229738

Knight, C., & Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 275–299). Springer. https://doi.org/10.1007/978-94-007-6094-3_14

Lewis, H. E. (1972). Medical aspects of polar exploration: Sixtieth anniversary of Scotts last expedition: State of knowledge about scurvy in 1911. *Proceedings of the Royal Society of Medicine, 65*(1), 39–42. https://doi.org/10.1177/003591577206500116

Mahoney, J. (2012). The logic of process tracing tests in the social sciences. *Sociological Methods & Research, 41*(4), 570–597. https://doi.org/10.1177/0049124112437709

Mayntz, R. (2004). Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences, 34*(2), 237–259. https://doi.org/10.1177/0048393103262552

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72*(1), 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x

Miguel, E. & Kremer, M. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Guide to Replication of Miguel and Kremer (2004). Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Miguel, E., Kremer, M., Hicks, J. & Nekesa, C. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Data User's Guide. Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.. https://doi.org/10.1017/CBO9781107587991

Ozier, O. (2021). Replication Redux: The reproducibility crisis and the case of deworming. *The World Bank Research Observer, 36*(1), 101–130. https://doi.org/10.1093/wbro/lkaa005

Pearl, J. (2009). *Causality*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2022). Direct and indirect effects. In Geffner, H., Dechter, R., & Halpern, J. Y. (Eds.). Probabilistic and Causal Inference: The Works of Judea Pearl (pp. 373–392).

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Rohlfing, I., & Zuber, C. I. (2021). Check your truth conditions! Clarifying the relationship between theories of causation and social science methods for causal inference. *Sociological Methods & Research, 50*(4), 1623–1659. https://doi.org/10.1177/0049124119826156

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31*(2), 161–170. https://doi.org/10.1111/j.1467-9469.2004.02-123.x

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331. https://doi.org/10.1111/j.1467-9469.2004.02-123.x

Runhardt, R. W. (2015). Evidence for causal mechanisms in social science: Recommendations from Woodward's manipulability theory of causation. *Philosophy of Science, 82*(5), 1296–1307. https://doi.org/10.1086/683679

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science, 37*(6), 1011–1035. https://doi.org/10.1111/cogs.12058

Sobel, M. E. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods and Research, 16*, 155–176. https://doi.org/10.1177/0049124187016001006

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, D., & Garner, P. (2015). Deworming drugs for soil-transmitted intestinal Worms in children: Effects on nutritional indicators, haemoglobin, and school performance (review). *Cochrane Database of Systematic Reviews, 7*. https://doi.org/10.1002/14651858.CD000371.pub6

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis

VanderWeele, T. J. (2014). A unification of mediation and interaction: A four-way decomposition. *Epidemiology, 25*(5), 749–761. https://doi.org/10.1097/EDE.0000000000000121

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18*(2), 137.

Waldner, D. (2007). Transforming inferences into explanations: Lessons from the study of mass extinctions. In R. N. Lebow & M. I. Lichbach (Eds.), *Theory and evidence in com-*

parative politics and international relations (pp. 145–175). Palgrave Macmillan. https://doi.org/10.1057/9780230607507_6

Waldner, D. (2012). Process tracing and causal mechanisms. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science* (pp. 65–84). Oxford University Press.

Weller, N., & Barnes, J. (2014). *Finding pathways: Mixed-method research for studying causal mechanisms*. Cambridge University Press. https://doi.org/10.1017/CBO9781139644501

Wilcox, A. J. (2006). Invited commentary. The perils of birth weight—A lesson from directed acyclic graphs. *American Journal of Epidemiology, 164*(11), 1121–1123. https://doi.org/10.1093/aje/kwj276

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yerushalmy, J. (1971). The relationship of parents' cigarette smoking to outcome of pregnancy – Implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology, 93*(6), 443–456. https://doi.org/10.1093/oxfordjournals.aje.a121278

Chapter 7
Testing Joint Sufficiency Twice: Explanatory Qualitative Comparative Analysis

Alessia Damonte

Abstract Standard Qualitative Comparative Analysis (QCA) applies an eliminative cross-case algorithm to identify which combinations of factors are logically associated with an outcome in a population. As such, it suits the purpose of pinpointing the conditions under which an outcome occurs or fails. However, the explanatory import of its findings only follows if the algorithm identifies theoretically *interpretable*, logically *valid*, and empirically *plausible* causal compounds.

The chapter provides an essential guide to designing an explanatory QCA that meets the three credibility requirements at once. Section 7.2 addresses how to develop starting hypotheses consistent with the assumptions of complex causation to preserve theoretical interpretability. Section 7.3 introduces the Boolean algebra required to model a hypothesis and find which part supports the explanatory claim in the cases at hand. Section 7.4 addresses the issue of gauging conditions to ensure the empirical plausibility of the analysis. Last, Sect. 7.5 summarizes the protocol, illustrated by the replicable example in the online R file.

Learning Objectives
After studying this chapter, you will be able to:

- Understand causation in terms of individual necessity and joint sufficiency of many factors.
- Develop a configurational hypothesis.
- Apply Boolean algebra to formalize configurational hypotheses and establish criteria of fit.
- Gauge factors as sets that are suitable to logical formalizations.
- Identify and discuss credible configurational solutions.

A. Damonte (✉)
University of Milan, Milan, Italy
e-mail: alessia.damonte@unimi.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_7

7.1 Introduction

Qualitative Comparative Analysis (QCA: Ragin, 1987/2014, 2000, 2008; Duşa, 2019; Oana et al., 2021; Mello, 2021) stands amid the suite of causal techniques for three main reasons that drive as many questions.

First, QCA moves from the default assumption that causation lies in compounds or teams of conditions. Its solutions entail that things happen when all the "right" conditions are given together, like in a chemical reaction (Mackie, 1965, 1974; Cartwright & Hardie, 2012). The first question of explanatory QCA asks how to ensure that results are interpretable "recipes" for the outcome.

Second, QCA originally revolves around a pruning algorithm. It compares configurations that meet regularity requirements of association with an outcome to drop irrelevant conditions, along the lines of a most-dissimilar case design (e.g., De Meur & Berg-Schlosser, 1994), albeit run twice. The second question asks how the technique can be geared toward pinpointing valid causal compounds despite the shortcomings of such a design (e.g., Geddes, 1990; Most & Starr, 2015; Krogslund et al., 2015).

Third, QCA's solutions hold at the levels of both the population and individual cases. Such a peculiarity is based on gauging operations that preserve quantitative and qualitative information. These operations are an integral part of the analysis and bind findings to analytic units. The third question asks how these operations affect the tenability of solutions.

These three questions are addressed in Sects. 7.2, 7.3, and 7.4, respectively. Section 7.5 summarizes the protocol illustrated by the online R file.

7.2 Interpretability

The recognized hallmark of QCA lies in its assumptions that causation is an asymmetric, conjunctural, and equifinal phenomenon (Ragin, 2008; see also Rosenberg et al., 2017). *Asymmetric* means that causation has a direction and proceeds from "causes" to "effects" as a relationship of dependence or conditionality ahead of temporal considerations. *Conjunctural* refers to the first reason for asymmetry: the actual cause is a compound and consists of a team, bundle, or package of contributing factors. *Equifinal* recalls the second reason for asymmetry: different compounds can yield the same outcome. These assumptions chime with mechanistic considerations on the ultimate shape of causation (e.g., Befani, 2013; Mahoney, 2021; Chap. 2).

7.2.1 Mechanisms and Machines

QCA assumes that the factors responsible for an outcome are many and related to each other as the constituting parts are to their whole. Moreover, it allows factors have substitutes without loss of effectiveness for the causal compound (Mackie, 1966; Cheng, 1997; Cartwright & Hardie, 2012).

The textbook illustration of such a parts-to-whole relationship offers heat, oxygen, fuel, and defective or no sprinklers as the compound accounting for fire. These circumstances provide the complete set of relevant conditions under which the process of combustion must initiate (Salmon, 2020). Thus, they form a causal team based on the process that they explain.

The process also clarifies the general relationship between components, teams, and outcomes. In the textbook example, combustion results in a fire when the whole team of circumstances is given in the same place and the right state—present heat, fuel, oxygen; absent or defective sprinklers. The surefire or *sufficient cause* of the outcome is the right bundle. However, the right circumstances can take many actual shapes. For instance, a lightning bolt, a short circuit, or a lit match can all be equivalent sources of heat. Any actual bundle, then, is *unnecessary* as such. Besides, the process fails when any circumstance is given in the wrong state—poor oxygen, no fuel, or no heat all prevent combustion, while a working fire system suffocates it. Any element of the compounds, then, is a counterfactually vital—and hence, *necessary*—component of the team, despite it alone being insufficient to yield the outcome. The elements of the compound are "partial causes" or "*inus* conditions"—*inus* being the acronym of the *Insufficient* but *Necessary* part of an *Unnecessary* but *Sufficient* team.

Bundles of *inus* conditions seldom capture a generative process directly (see Chaps. 8, 9, and 10). Instead, they can capture the set of right circumstances as "nomological machines"—that is, as "sufficiently stable" arrangements of triggering, enabling, sustaining, and shielding conditions underlying the generative process (Cartwright, 1999: 49, 2017). A nomological machine is such that its components together make other factors irrelevant before the same type of outcome across time and space. Therefore, a nomological machine is the specified explanation of a regular behavior independent of the remaining context (Craver & Kaplan, 2020). Moreover, it provides the theoretical construct that affords counterfactual evidence about the contribution of single components across cases.

7.2.2 Operationalizing Typological Theories

Typological theories provide a renowned starting point for developing configurational explanations (e.g., Elman, 2005). Such theories prove especially fruitful as they enable modeling of the alternative causal bundles as different settings of the same factors.

Some theories are consistent "explications" of a driving concept. For instance, Pahl-Wostl (2008) takes "regimes" as the driving concept. She defines water management regimes as the alignment of governance style, type of sectoral integration, scale of analysis and operation, information management, plus finance and risk management. Huntjens et al. (2011) operationalize the setting of these structural dimensions for two polar types of regimes—the "market-based" and the "integrated adaptive"—then run a QCA to establish the features that account for the diversity in the policy-learning capacity of water management systems when faced with climate

change challenges. In a similar vein, Colby (1991) builds on the concept of "policy paradigms." He stipulates that the compatibility of environmental and economic policy goals depends on the alignment of policy ideas and policy tools. Thus, "frontier economics" and "deep ecology" establish the trade-off between economic growth and environmental preservation, while "environmental protection," "resource management," and "eco-development" make room for their coexistence and integration. Damonte (2013) operationalizes these alternative paradigms as different settings of the same bundle of policy tools and identifies the configurations that account for the green decoupling of economic growth from pollution.

Other configurational hypotheses integrate heterogeneous streams of literature into a consistent explanatory whole. For instance, Sabatier and Mazmanian (1980) reason that the many accounts of the success and failure of policy implementation can be reduced to the consistent interplay of three dimensions: problem tractability, administrative effectiveness, and political support. Hinterleintner et al. (2016) operationalize the components of each dimension and run a QCA that explains the differences in the IMF's evaluation of austerity programs as differences in the credibility of national implementations. Theoretical integration can also be purposefully operated within the study. As an example, Lauri et al. (2020) integrate theories linking the defamiliarization of care work and gender equality with theories on the gender division of labor as embedded in different types of welfare systems. On this basis, they provide a thorough operationalization of childcare policies as bundles of tools that enforce different gender norms. QCA is applied to identify which tools, linked to the norms of which type of welfare system, yield high gender equality and which endanger the goal instead.

7.2.3 Assembling Configurational Hypotheses

A configurational hypothesis can also be crafted after a reasoned selection and integration of statistical "determinants." Surveys of scholars' practices (Amenta & Poulsen, 1994; Berg Schlosser & De Meur, 2009) pinpointed four selection strategies. The "comprehensive approach" includes all the factors from all the relevant theories; the "perspective approach" selects single variables that represent major theories; the "significance approach" only focuses on statistically significant variables; the "second look" approach mixes statistically significant variables with theoretically meaningful factors that did not survive those same tests.

However, none of these strategies is proven to yield proper configurational hypotheses unless the selected factors can be related to the unfolding of a generative process as actors' constraints and opportunities. To witness, Stiller (2017) explains governments' success in adopting major welfare reforms as the interplay of policymakers' strategies—identified in ideational leadership, concession making, and blame avoidance—with key background features that make these strategies adequate—namely, the stage of the election cycle and the government's position toward the national welfare system. Similarly, Ansell et al. (2020) account for stakeholders' participation in collaborative governance as the result of motivations—that is,

perceived incentives, interdependence, trust, and purpose—and governance's support of motivations—through leadership services, opportunities to build relationships, and structures for pooling information.

A configurational hypothesis may also follow from problematizing correlational theories. Kogut and Ragin (2006) focus on the theory linking high economic development, thriving financial markets, and common law institutions. The configurational hypothesis develops from the consideration that the causal chain is underspecified. National economies, they reason, may still thrive despite poor financial markets if legality is ensured. Moreover, the effectiveness of common law institutions beyond their original contexts depends on their interplay with existing legal traditions. Thus, they run two QCAs that employ common law, features of the institutional "transplant," and commitment to the rule of law to account for differences in GDP per capita and, separately, in the dimension of the domestic financial markets, to check whether the two explanations overlap.

In short, the fundamental criterion for selecting an interpretable candidate *inus* factor is functional. It consists of whether one can develop *directional expectations* about the factor's contribution to the setting that compels and protects some causal process of interest. The expectation should support the claim that, were the factor given in the right state and in the right team, the process to the outcome would certainly follow. As we will see in Sect. 7.3.2, these directional expectations play a crucial role in the analysis as they establish the plausibility of counterfactual assumptions.

7.3 Validity

The validity of inferences about *inus* hypotheses depends on the algebra deployed to make them testable. Such a suitable algebra should allow factors to

- Have observable states, such as presence and absence;
- Form compounds as configurations of states;
- Have equifinal alternatives;
- Establish relationships of dependence.

Boolean algebras can easily render these states and relationships. Introduced as primary devices to analyze human reasoning about the world (De Morgan, 1847; Boole, 1853), their structures support a twofold reading (Stone, 1936)—logical, and set-theoretical.

7.3.1 QCA's Algebra

Like any other, QCA's algebra is a language of literals and operators suitable to render complex relationships according to fundamental rules.

7.3.1.1 Literals

Boolean algebras use "literal symbols" to indicate factors as attributes or states of a unit of observation. A literal stands for a name or an adjective denoting "either a thing or some quality or circumstance belonging to it" (Boole, 1853:27). QCA borrows the convention and indicates a state with an uppercase letter. Thus, A reads ⌜A present⌝ or ⌜A positive⌝ or the predicate ⌜is A⌝. The literal provides an empty placeholder for whatever attribute we consider as the candidate *inus* condition—such as "inflammable" referred to a material; "hierarchical" to a governance structure; "affluent" to a society; "independent" to a voter.

Once defined, a literal establishes the similarity of any units of observation u_i to which it applies. In Boole's original proposal, and all the basic operations of QCA, such a recognition raises a class, that is, an *idempotent* collection of units. Idempotency means that, in contrast to probabilistic samples, classes satisfy the logical rule dubbed *dictum de omni*: that which can be said of the whole, it also holds for each of its parts. Boole renders idempotency as in Eq. (7.1):

$$A^2 := A \tag{7.1}$$

where := indicates a stipulation and reads ⌜is by definition equal to⌝. As the only two numerical values that satisfy the stipulation are 1 and 0, Boole's literals can only take these two values—and the basic operations in QCA share this bivalent assumption, too.

These values convey two separate readings of the relationship between a unit and a literal:

- When the literal is understood as a *predicate*, 1 and 0 are the *truth values* that a literal can take in the actual unit u_i from the *universe of discourse* $\mathbb{U} = \{u_1, \ldots, u_N\}$. 1 reads ⌜true⌝ for ⌜it is the case that⌝, while 0 reads ⌜false⌝ for ⌜it is not the case that⌝.
- When the literal is understood as a *class*, 1 and 0 are read as *membership values*. Thus, $A_i = 1$ means that the i-th unit belongs to class A, while $A_i = 0$ indicates that the same unit does not belong to it.

The logical understanding captures the literal as the *intension* or quality of a unit. In contrast, the set-theoretical understanding captures the literal as the *extension* of the quality across the units in a universe. Operationally, the intension is decided by gauging rules—for instance, on defining which manifestations and intensity make it true that a unit ⌜is A⌝. Extension, on the other hand, is decided by counting—for instance, the number of units in the universe that ⌜are A⌝, which corresponds to the *cardinality* of class A. In bivalent Boolean algebra, the two readings overlap, making logical inferences especially straightforward.

7.3.1.2 Operators

The Boolean operators relevant to *inus* hypotheses correspond to the logical connectives ⌜*not*⌝, ⌜*and*⌝, ⌜*or*⌝, ⌜*only if*⌝, ⌜*if*⌝ and the set-theoretical relationships of *difference*, *intersection*, *union*, and *superset/subset*.

Negation

The connective ⌜*not*⌝ denies the literal. The Boolean notation renders it with a bar above the uppercase literal to which it applies; in QCA, also common is the tilde before the uppercase literal, or the use of the lowercase literal. Thus, $\overline{A}, \sim A, a$ all read ⌜is not-A⌝.

The logical negation transforms a unit's truth value into its opposite, calculated as in Eq. (7.2). The set-theoretical reading establishes the negation of a set is the collection of units that are excluded from that set. Therefore, the negated set $\overline{A}$ corresponds to the difference (indicated by the backslash \) between the universe U and set A, as in Eq. (7.3):

$$\overline{A}_i := 1 - A_i \tag{7.2}$$

$$\overline{A} := \mathbb{U} \setminus A \tag{7.3}$$

Equations (7.2) and (7.3) indicate that, by definition, a literal and its negation are mutual *complements*. The enforcement of this definition depends on gauging operations—an issue addressed in Sect. 7.4.

Joint Occurrence

These correspond to bundles of literals connected by the ⌜*and*⌝ operator. In logic, the operator is a wedge (∧); in set theory, it is a cap (∩). In QCA, the operator is a dot (●) or a star (∗) although the connecting symbol may be omitted.

Two implications are worth noting. Permutation and grouping are irrelevant to ⌜*and*⌝ bundles: $AB\overline{C}$ means the same as $A\overline{C}B$ and $A\left(B\overline{C}\right)$ as the resulting class clusters the same units. In short, the Boolean ⌜*and*⌝ supports the commutative and the associative rule. Therefore, bundles are blind to the time dimension of sequences; instead, they emphasize the joint occurrence or interaction of attributes in a unit.

Logically, the ⌜*and*⌝ operator raises a *conjunction*. The underlying rule establishes a conjunction as true when each of its conjuncts is true. The rule is also known as "*the weakest link*": the conjunct with the lowest truth value defines the truth value of the compound.

Applied to a single predicate and its negation, the rule renders the logical *principle of non-contradiction*. As summarized by Eq. (7.4), the principle states that a predicate and its negation cannot be true of the same unit at the same time in the same sense. Set-theoretically, the principle is met when the intersection of a set and its negation is empty (∅), as in Eq. (7.5). The principle offers the first criterion of validity: it commits to rejecting inferences that build on, or lead to, *contradictions*.

$$A \wedge \overline{A}_i := 0 \tag{7.4}$$

$$A \cap \overline{A} := \varnothing \tag{7.5}$$

More generally, the weakest link of the i-th unit can be calculated as the minimum of its truth values in any of the $1 \leq j \leq K$ conjuncts, as in Eq. (7.6):

$$\bigwedge A_{ij} = \min\left(A_{i1}, \ldots, A_{iK}\right) \tag{7.6}$$

Therefore, in a universe of N units, the cardinality of the intersection of the k literals of interest corresponds to the sum of the $1 \leq i \leq N$ units' weakest links as in (7.7):

$$\bigcap A_j = \sum_{i=1}^{N} \min\left(A_{i1}, \ldots, A_{iK}\right) \tag{7.7}$$

Alternatives

These arise when literals are connected by the operator ⌜*or*⌝. In QCA, the operator is a plus symbol (+) and never omitted. Logic indicates it with a vee (∨); set theory with a cup (∪). Class idempotency makes permutation and grouping irrelevant to alternatives, too.

Logically, the ⌜*or*⌝ operator raises a *disjunction*. The underlying rule establishes the disjunction as true when at least one of its disjuncts is true. The rule can be dubbed "*the strongest link*": the disjunct with the highest truth value defines the truth value of the whole compound.

Applied to a single predicate and its negation, the rule renders the logical *principle of the excluded middle*. As summarized by Eq. (7.8), the principle states that, necessarily, either a predicate or its negation is true in a unit, so that the disjunction of the two raises a non-informative tautology. Set-theoretically, the principle is met when the union of the set and its negation returns the universe, as in Eq. (7.9).

$$A_i \vee \overline{A}_i := 1 \tag{7.8}$$

$$A \cup \overline{A} := \mathbb{U} \tag{7.9}$$

More generally, the strongest link of the *i*-th unit can be calculated as the maximum of the truth values of any of the $1 \leq j \leq K$ disjuncts, as in (7.10):

$$\bigvee A_{ij} = \max\left(A_{i1}, \ldots, A_{iK}\right) \tag{7.10}$$

Therefore, in a universe of N units, the cardinality of the union of the K literals of interest corresponds to the sum of the $1 \leq i \leq N$ units' strongest links, as in (7.11):

$$\bigcup A_j = \sum_{i=1}^{N} \max\left(A_{i1}, \ldots, A_{iK}\right) \tag{7.11}$$

Necessity and Sufficiency

The reliance of QCA on the assumptions of *inus* causation gives center stage to the concepts of necessity and sufficiency.

Mackie (1974) illustrates them with the different behavior of coin-operated vending machines. A "sufficiency machine" always drops a snack for a coin, and sometimes it drops one without apparent reason, too. A "necessity machine" never drops a snack without a coin, and sometimes the coin fails. Last, one and only one snack for each coin is the behavior of the perfect "necessity-and-sufficiency machine." These intuitions capture both set-theoretical and logical relationships between an observed input, or antecedent (the coin), and an observed output, or consequent (the snack), connected by an unobserved—but possibly observable—mechanism.

As for notation, QCA indicates necessity with an arrow running from the outcome to the cause and sufficiency with an arrow running from the cause to the outcome. Thus, $A \rightarrow B$ reads ⌜A is sufficient to B⌝; $\overline{A} \leftarrow \overline{B}$ reads ⌜not-A is necessary to not-B⌝.

Set-theoretically, the *necessity* of A to B corresponds to A being a *superset* of B, indicated as $B \subset A$. The relationship is satisfied when *all the B are also A* although there can be instances of A in the universe that do not display B. This corresponds to the logical situation in which being B *implies* being A or, more compactly, ⌜B, only if A⌝. The hallmark of necessity is the impossibility of the outcome in the absence of the factor, as in (7.12). Set-theoretically, it means that the proof of the necessity of A to B in the universe comes from the empty intersection in (7.13).

$$\overline{A}_i \wedge B_i = 0 \tag{7.12}$$

$$\overline{A} \cap B = \varnothing \tag{7.13}$$

Set-theoretically, the *sufficiency* of A to B corresponds to A being a *subset* of B, indicated as $A \subset B$. The relationship is satisfied when *all the A are also B*. In short, sufficiency renders the intuition of A as the constant antecedent condition of

B. Logically speaking, it corresponds to saying that, for any u_i, ⌜B, if A⌝ without exceptions. The hallmark of sufficiency coincides with the impossibility that the outcome fails when the factor is present, summarized by requirement (7.14) and its set-theoretical translation (7.15):

$$\overline{B}_i \wedge A_i = 0 \tag{7.14}$$

$$\overline{B} \cap A = \varnothing \tag{7.15}$$

7.3.1.3 Truth Tables

Stipulations and rules construe valid logical inferences as the calculus of truth values, visualized with the aid of a *truth table*. These tables clarify the possibilities that the selected literals make available ahead of observation. Logic sees it as the exhaustive catalog of the combinations of the literals' truth-values (Wittgenstein, 1922). Probabilistic theories dub such a structure "*sample space*" and understand it as the list of the potential events from random trials (e.g., Clarke, 2020). In any case, this structure reports the maximum diversity that units can display given specific literals and gauges.

The truth table entails a fundamental sense-making operation (Quine, 1982); thus, in it, each combination of the literals' truth values can be dubbed a *primitive*. The number of primitives depends on the number of literals and truth values under consideration; K bivalent literals yield 2^K unique primitives. In the remaining, a truth table will be indicated as Ω and its primitives as ω.

The shape of truth tables follows conventional rules. The primitives are listed as rows: ω_1 displays all true literals; ω_{2^K}, all false ones (cfr. Duşa, 2019). Each of the remaining columns in the classical truth table is for the *truth function* of a connective, i.e., the truth values that each primitive returns when the connective's rule is applied to the states of its literals.

Table 7.1 displays a truth table of two literals (A, B) and five operators to indicate as many relationships—respectively, of conjunction (*and*), disjunction (*or*), necessity (*only if*), sufficiency (*if*), plus necessity and sufficiency (*iff*).

The values in the truth functions of each operator indicate the type of units that will (1) and will not (0) be observed if the relationship holds in the universe of reference (Sprenger, 2011). These expectations inform the discourse on the threats to the validity of inferences that are currently addressed by either design (e.g., Chap. 3) or model (e.g., Chaps. 6 and 8, Sect. 7.3.2 below).

- The *and* truth function follows from the application of the weakest link rule as in Eqs. (7.6) and (7.7) and returns a single true point in correspondence with the matching primitive (ω_1 in Table 7.1). Thus, evidence of a conjunction is only provided by the units displaying every conjunct in the right state.

Table 7.1 Truth table of two literals and five operators

Ω	A	B	*A and B*	*A or B*	*B, only if A*	*B, if A*	*B, iff A*
ω_1	*1*	*1*	1	1	1	1	1
ω_2	*1*	*0*	0	1	1	0	0
ω_3	*0*	*1*	0	1	0	$1^{(*)}$	0
ω_4	*0*	*0*	0	0	1	1	1

Note: $^{(*)}$ observing this primitive makes the statement of sufficiency vacuously true

- The *or* truth function follows from the strongest link rule as in Eqs. (7.10) and (7.11) and always returns a single false point, corresponding to the primitive with no matching values (ω_4 in Table 7.1). It conveys that any unit displaying at least one disjunct in the right state provides evidence of a disjunction.
- The *only if* truth function has a single false point corresponding to the impossible primitive established by Eqs. (7.12) and (7.13). It shows that the relationship of necessity is only inconsistent with evidence of the consequent B occurring in some units where the antecedent A is missing (ω_3 in Table 7.1). Therefore, the logical relationship of necessity assumes the antecedent A is not substitutable, as is oxygen to fire.
- The *if* truth function has a single false point in the impossible primitive defined by Eqs. (7.14) and (7.15). It shows that the claim of sufficiency is only inconsistent with evidence that the consequent fails under the antecedent in some units (ω_2 in Table 7.1). The logical relationship of sufficiency is the regular connection of antecedent and consequent. When the actual cause is composite, the requirement can only be satisfied by the antecedent that comprises all the components of a compound—including the factors that shield the causal process from obstructions. Section 7.4.2 will suggest a strategy for construing suitable shielding factors.

A further note is due about the starred value of ω_3 in Table 7.1. The instances of this primitive do not contradict the claim of sufficiency after the principle that *ex falso quodlibet*—meaning that anything can follow in the units where the antecedent is missing or otherwise false. However, units of this type provide *vacuous* evidence about the relationship (e.g., Salmon, 2020), as they may

(a) point to its nonsensical nature. The evidence that Socrates is not a triangle yet is a philosopher makes the claim vacuous that "if Socrates is a triangle, then he is a philosopher."
(b) divert attention from the conditionality of interest. Evidence about salt that is not put in water is irrelevant to establish the claim that "if salt is put in water, then it dissolves."
(c) unveil some spurious relationship or incomplete explanation. The evidence that the barometer reads "storm" during a sunny day makes the claim vacuous that "if the barometer reads 'fair,' then it is a sunny day."

Although the exact meaning of a vacuous observation depends on the interpretability of the relationship of interest, it nevertheless makes the problem visible as a formal issue of validity.

- The *iff* relationship arises from the conjunction of the truth functions of necessity and of sufficiency. It indicates the identity of the two literals and the overlapping of the respective classes of units in the universe. Thus, the truth function has two false points. In Table 7.1, these correspond to ω_2 and ω_3. In short, evidence of any inconsistency in the covariation of the two states challenges the validity of the identity.

QCA does not deploy logic, truth tables, and truth functions normatively. Instead, it relies on them as modeling tools and heuristics for the analysis.

7.3.2 Identifying Valid Inus *Hypotheses*

Logic provides scaffolding and criteria to render an *inus* hypothesis first, then decide whether it is rightly specified to the universe under analysis.

7.3.2.1 Rendering Hypotheses

Logic renders an *inus* hypothesis as a theoretically meaningful yet unwarranted claim about the sufficiency of a conjunction of K conditions to the occurrence of the outcome Y, as in (7.16)

$$\bigcap_{j=1}^{K} A_j \rightarrow Y \tag{7.16}$$

The formula means that ⌜were it the case that these K conditions together make an *inus* machine, then the outcome should certainly occur in an ideal instance displaying them all in the right state, and fail otherwise⌝. For it to hold, the starting hypothesis should contain the sufficient bundle to the positive and the negative outcome, which may have different specifications. QCA acknowledges this fact and addresses the positive and the negative outcomes in separate analyses. Nevertheless, the two sets of findings are related as long as both follow from the same truth table in which primitives are exclusively assigned to one outcome, and no contradiction is detected.

The value of an explanatory QCA lies in identifying the *plausible* bundle beneath the success and failure of an outcome in the population of interest, to define the tenability of the starting hypothesis and its underlying theory. Its identification procedure addresses validity issues as the underspecification or the overspecification of the starting hypothesis.

7.3.2.2 Tackling Underspecification

QCA deploys truth tables as a diagnostic device for detecting underspecification. Therefore, QCA's truth tables are partially different from those of logic.

A QCA's truth table contains as many columns as *inus* conditions in the hypothesis, plus one for the outcome and at least three additional columns for as many parameters of fit. The truth value of the outcome is the last column to be filled, depending on the researcher's decisions about the parameters, as follows:

Decision 1: Frequency Cut-Off

This parameter establishes whether a primitive is observed or realized in the universe of reference based on the minimum number of its "best instances" (Ragin, 2008). A unit is the best instance of the primitive in which it gets a membership scorehigher than 0.5 according to the weakest link rule (7.6).

Units' classification yields two kinds of primitives: *observed* or *realized*, and *unobserved* or *unrealized*. The unrealized ones are also known as *logical remainders* and constitute a common occurrence. Although the ratio of units to conditions inevitably plays a role in raising them (Marx & Duşa, 2011), their number is relatively independent of the richness of the hypothesis or the size of the universe. Instead, the logical remainders expose the *limited diversity* of the units under analysis and serve as a source of counterfactual reasoning (Ragin, 2008; see below).

The researcher's decision regarding the frequency cut-off may also increase the number of unrealized primitives. Conventionally, one best instance is enough to declare a primitive realized albeit rare. However, the frequency cut-off can be raised if the numerosity of the population and the gauging strategy suggest a risk of errors in units' classification.

Decision 2: The Consistency Threshold

The second of the researcher's decisions on the truth table for a QCA concerns the assignment of the realized primitives to either the positive or the negative outcome. In Standard QCA, the decision mainly follows considerations on consistency.

In line with consolidated axiomatizations (Hájek, 2011), QCA captures the *consistency of the sufficiency* of each primitive to an outcome (*S.cons* for short, also known as *incl* for "inclusion": Ragin, 2008; Schneider & Wagemann, 2012; Duşa, 2019) as an extensional gauge that checks for empirical violations of the impossibility requirement in (7.15) through the ratio in Eq. (7.17):

$$S.cons_{\omega_* \to Y} = \frac{\left|\omega_* \bigcap Y\right|}{\left|\omega_*\right|} \tag{7.17}$$

The vertical bars indicate the size of a partition. The denominator of the ratio is for any antecedent of interest—otherwise understood as the number of trials—and here corresponds to the primitive of interest. The numerator is for the number of successful trials, that is, the intersection of the primitive with the outcome. When none of the N units under analysis qualifies as an instance of the inconsistent intersection $\omega_* \overline{Y}$, the numerator overlaps the denominator, and the *S.cons* gets its highest value of 1.00, which supports the claim that ω_* is sufficient to Y. The lower the overlapping, the lower the *S.cons* parameter and the credibility of the claim of sufficiency.

The detection of critical inconsistencies justifies the dismissal of the hypothesis in the current shape as incomplete or otherwise misspecified (e.g., Rihoux & De Meur, 2009; Rohlfing, 2020). The textbook illustration comes from a configurational model applying Lipset's socioeconomic theory of democratization to account for the breakdown of democracy in Europe between the two World Wars. The model yielded a straightforward truth table with a single remarkable contradiction: the German case displayed all the socioeconomic conditions for a thriving democracy, but it experienced a clear regime breakdown. The contradiction disappeared after adding institutional conditions of government stability to the model.

The researcher's decision concerns the value of the *S.cons* below which the inconsistency is severe enough to preclude the assignment of the primitive to the outcome. An established convention suggests setting it at 0.85, although the range of *S.cons* values in the table may justify a different choice. An additional criterion considers "natural gaps"—that is, steep falls in the ordered series of the primitives' *S.cons* values. These gaps may suggest setting the consistency threshold in between clusters of primitives.

The primitives not assigned to Y cannot be automatically assigned to $\overline{Y}$. Instead, the consistency of each primitive has to be tested with both states of the outcome separately. Nevertheless, meaningful solutions can be expected when the realized primitives below the consistency cut-off to Y return high *S.cons* values to $\overline{Y}$. This suggests that the starting hypothesis can account for both the occurrence and the non-occurrence of the outcome consistently.

Decision 3: The Coverage Cut-Off

The least common and last of the possible researcher's decisions concerns the empirical import of the claim of sufficiency—how relevant the primitive is to the set of instances of the outcome of interest. The related parameter, dubbed *coverage of sufficiency* (*S.cov* for short) is calculated as in (7.18)

$$S.cov_{\omega_* \to Y} = \frac{\left|\omega_* \bigcap Y\right|}{\left|Y\right|} \tag{7.18}$$

When all the instances of a primitive ω_* display the outcome, the numerator in (7.18) equals the denominator, and the parameter takes its highest value of 1.00 supporting the claim that the primitive accounts for any unit with the positive outcome. But the empirical relevance of a factor to an outcome is the extensional gauge of its necessity in the cases at hand. Hence, the *S.cov* of ω_* to Y gauges the *consistency of necessity* (*N.cons* for short) of the primitive to the outcome. Specularly, the *S.cons* of ω_* to Y gauges the empirical relevance of the primitive as a necessary compound to the outcome—and hence counts as the *N.cov* of ω_* to Y.

A primitive's *S.cov* value decreases with the increase in the evidence that the outcome can occur without the primitive. Coverage cut-offs may be established to ensure the analysis is based on sufficient primitives that also are empirically relevant. However, decisions driven by empirical relevance may prove unwise, as even rare primitives may contribute to specify the composition of *inus* machines.

7.3.2.3 Tackling Overspecification

Overspecification depends on having included factors in the starting hypothesis that prove irrelevant to account for the units' diversity.

The issue arises as mistaking some features for an *inus* component entrenches solutions in very specific contexts and unnecessarily reduces their portability (e.g., Craver & Kaplan, 2020; Salmon, 2020; cfr. Álamos-Concha et al., 2021; Chap. 10).

The acknowledged sources of overspecification are twofold: irrelevant components, and trivial factors.

Irrelevant Components

Quine-McCluskey's *minimizations* provide the standard approach to irrelevant conditions (Ragin, 1987/2014, 2000, 2008). These minimizations identify irrelevant components in the single varying conjunct of two otherwise identical primitives. To witness, the minimization is possible of the primitives *ABCD* and $ABC\overline{D}$ if both display high *S.cons* values to the same outcome. The formal reason is that the two allow the factorization $ABC\left(D \cup \overline{D}\right)$, where $D \cup \overline{D} := \mathbb{U}$ by Eq. (7.9). The operation highlights that the *implicant ABC* is sufficient to Y regardless of D, which can be dismissed as not *inus* a factor.

The adjudication of the *inus* nature of single components may change depending on how minimizations deal with the logical remainders. The Standard Analysis affords three alternative *counterfactual assumptions*, each leading to "solutions" at different degrees of specification, as follows:

- *Conservative or complex solutions*. These are returned under the assumption that unrealized logical remainders would have proven ambiguous had they been realized. Hence, minimizations only operate on observed primitives. With high lim-

ited diversity, the solutions could be as rich as the disjunction of any realized primitive.

- *Parsimonious solutions*. A superset—and hence, more general in scope—of the conservative solutions, the parsimonious solutions are returned under the assumption that any logical remainder could prove sufficient if matching a realized primitive except for one literal.
  The surviving factors are the *inus* components in the hypothesis that are essential to account for the difference between the instances of the successful outcome and the instance of the failed one.
  However, parsimonious minimizations can yield gappy explanations. Like the treatment variable in the Potential Outcome Framework (see Chap. 3) or the mediators in Path Analysis (see Chap. 6), the solutions from the parsimonious minimization may capture a causal channel, but certainly dismiss the information about the covariates needed to account for the effect (Damonte, 2021b). The reason is that the parsimonious minimizations drop factors regardless of the plausibility of the logical remainders that they employ.
- *Intermediate or plausible solutions.* These are returned under the assumption that only those logical remainders qualifying as *easy counterfactuals* would have proven sufficient if realized.
  To understand the difference between an easy and a hard counterfactual, imagine the following. At the outset, we include condition *A* in the starting hypothesis under theoretical and empirical reasons to assume that it is an *inus* factor. More specifically, we assume that the condition makes an unknown causal compound Φ sufficient to the outcome *Y* when given in a state, say *A*, while in the opposite state, say $\overline{A}$, it turns Φ into a failure machine. In short, we add *A* under the *directional expectations* that

  (i) $A\Phi \subset Y$; and

  (ii) $\overline{A}\Phi \subset \overline{Y}$,

  where $\subset$ indicates a subset.

After we build and populate the truth table, we find the primitive $\omega_1 = ABCD$ is observed with an *S.cons* of 1.00 to *Y*, while we do not observe (hence we star) the primitive $\omega_9^* = \overline{A}BCD$. According to the single difference rule, ω_1 and ω_9^* can be minimized to BCD. However, the minimization entails that ω_9^* is consistent with *Y*, and hence that $\overline{A}\Phi$ would yield *Y* if observed. This goes against our directional expectation (*ii*) and makes a *hard* or *implausible counterfactual* of ω_9^*.

Now imagine the primitive $\omega_{13} = \overline{A}BCD$ is realized with an *S.cons* of 1.00 to *Y*, while the primitive $\omega_5^* = A\overline{B}CD$ is a logical remainder. Again, according to the single difference rule, ω_{13} and ω_5^* can be minimized to $\overline{B}CD$. The minimization entails that ω_5^* is consistent with *Y* and that $A\Phi$ would yield the outcome if observed. This agrees with our directional expectation (*i*); hence, ω_5^* qualifies as an *easy* or *plausible counterfactual*.

Intermediate minimizations return solutions from observed primitives and easy counterfactuals only. The factors added to the parsimonious solution terms may not

be essential to preserve the non-contradictoriness of the compounds. As they improve the sufficiency of the implicant, they offer a more complete account of why the outcome failed in specific units while succeeding in others (Ragin, 2008; Fiss et al., 2013; Duşa, 2019; Oana & Schneider, 2018; Damonte, 2021a; cfr. Baumgartner, 2015; Baumgartner & Thiem, 2020).

A Note on Ambiguity in Solutions

Regardless of the usage of the logical remainders, it has been emphasized that solutions in Standard QCA may encounter problems of ambiguity as the same primitives to an outcome may yield different prime implicants. To witness, the primitives $ABC, AB\overline{C}, A\overline{B}C$ can legitimately be minimized as $AB \cup A\overline{B}C$ or $AC \cup AB\overline{C}$. The information is displayed in a *Prime Implicant Chart* that shows which prime implicant covers which primitive, as displayed in Table 7.2.

Originally, the PI Chart was devised to allow the researchers making a decision on which implicants could be retained in solutions in light of their theoretical import. The practice has been deprecated, as cherry-picking implicants may build a confirmation bias into solutions (e.g., Baumgartner & Thiem, 2020; Baumgartner, 2015), and the current good practices require that alternative implicants are reported, too. Besides, the alternative minimizations may contain information of interest for discussion. For instance, in the example above, the two solutions indicate that A is always required—it can be an enabling condition—but, in the cases at hand, it obtains in team with B or C—which can play as triggering conditions. The richer implicants $A\overline{B}C, AB\overline{C}$ add that the one trigger can compensate for the absence of the other. These two richer implicants are currently left implicit by the reporting conventions that reward lean solutions. Under these rules, privileged prime implicants are those terms that, together, maximize the coverage of primitives—as are AB, AC in Table 7.2. Indeed, the conclusion that the union $AB \cup AC$ obtains the outcome does justice to alternative minimizations while logically entailing the richer implicants. Still, the information in the PI Chart deserves some attention, for it may suggest more accurate causal interpretations.

Table 7.2 Example of Prime Implicant Chart

Primitives *Implicants*	ABC	$A\overline{B}C$	$AB\overline{C}$
AB	x		x
AC	x	x	
$AB\overline{C}$			x
$A\overline{B}C$		x	

Dealing with Trivial Factors

Trivial factors are degenerate necessary conditions, that is, limiting cases of supersets. These arise when all or almost all the units in the universe of reference make the same state of the condition true—in short, when their distribution is skewed or constant.

Trivial factors can be detected by plugging the size of one condition in the place of the primitive in the formulas of the *N.cons* as in (7.18). When all the instances of the tested condition display the outcome, the numerator equals the denominator, and the parameter takes its highest value of 1.00, supporting the claim that the condition is necessary to the outcome. Conditions with a score of *N.cons* higher than 0.95 can be tested for skewness through a further parameter dubbed *Relevance of Necessity* (*RoN*: Schneider & Wagemann, 2012) and calculated as in (7.19) below:

$$RoN_{A \leftarrow Y} = \frac{|1 - A|}{|1 - A \cap Y|} \tag{7.19}$$

The parameter takes its lowest scores when the distribution of the condition by the outcome of reference proves trivial—when the size of $1 - A$ is remarkably smaller than the size of $1 - A \cap Y$, indicating the instances of the negative outcome raise independently of the absence of the condition. The standard recommendation is to consider dropping the factors with *N.cons* close to 1.00 and low *RoN* from the hypothesis. Thus, such "analysis of necessity" is a recommended step to be performed ahead of constructing the truth table (Schneider & Wagemann, 2012).

The original expected advantage was of pinpointing those constant conditions that double the number of primitives in the truth table while leaving almost half of them unobserved and lowering the consistency of every solution. However, the dismissal of a quasi-constant may prove unwise if the model requires it to prevent contradictory primitives (Rohlfing, 2020). The essentiality of the contribution can be easily ascertained by verifying whether a change in the consistencies of the primitives occurs after the seemingly trivial condition is dropped from the hypothesis (Damonte, 2021a). Nevertheless, the calculation of the parameters of fit on individual conditions remains a crucial source of information, as their values can support directional expectations or suggest reconsidering them.

7.4 Soundness

The actual link between sets, predicates, and the real world is decided by how truth values are assigned to literals—that is, by gauging.

The standard assumption in representation measurement theory maintains real-world properties depend on some units' deep structure that we can know indirectly only as meaningful variations in related observable attributes. This theory assumes

we can represent these attributes through *numerical images* and capture their variation through adequate scales. Scales warrant that for any manifestation p_i of the property P in the unit u_i there is a measure q_i of the image Q such that the functional relationship between measures preserves some fundamental relationship in the variation of the attribute.

The seminal work of Stevens (1946) pinpointed four such fundamental relationships: sameness, rank, distance, and proportion, preserved by nominal, ordinal, interval, and ratio scales, respectively. Conventional textbooks have long taught that a hierarchy of scope exists among measurements with the ratio scale at the top as the most "robust" one—i.e., abstracted from actual entities and their contexts. Intended as a prudential rule for naive statisticians (e.g., Luce, 1959), the hierarchy has turned into a canon and, as such, has been disputed since its introduction. Indeed, any measurement entails a *loss function*, and the loss is admissible that allows retaining crucial information (e.g., Guttman, 1977). Thus, prominent comparatists contend that ratio scales prove robust for detecting fine-grained changes, but sacrifice the information on "critical points." The qualitative change that occurs in the state of a unit when the measure of a crucial attribute reaches a special value is better conveyed by nominal scales (e.g., Sartori, 1984, 1991; Collier & Mahon, 1993; Ragin, 2000; Goertz, 2020).

In short, scales entail a trade-off between *precision* and *meaning*. However, the trade-off can weaken when metric variables are remapped as *fuzzy sets*.

7.4.1 Gauging for QCA: The Theoretical Side

7.4.1.1 The Starting Point

Zadeh (1968, 1978) introduced fuzzy sets to widen the scope of algorithmic problem-solving. He noted how machines could deliver precise solutions, but limited to trivial problems, while the human brain tackles complex issues through linguistic structures with hazy *hedges* such as ⌜*very*⌝, ⌜*somewhat*⌝, or ⌜*almost*⌝.

Fuzzy scores translate hedges into weights (μ) ranging from 0.00 to 1.00 to convey the degrees of membership of u_i to the set of A instances. They, too, understand the membership in a set and its opposite as complements, calculated as in (7.20):

$$\mu_{i \in \bar{A}} = 1.00 - \mu_{i \in A} \tag{7.20}$$

where $\in$ reads ⌜*in*⌝.

The meaning of the relationship between complements is established by a third relevant value, the *crossover*. Conventionally weighing 0.50, the crossover is the point of neutrality and signals a membership neither in the set nor in its complement.

Logically, fuzzy scores capture the *possibility* that the statement ⌜is A⌝ is true for the actual unit u_i: 1.00 indicates the statement is *certainly* true; 0.00 indicates the statement is *certainly not* true; 0.50 indicates that the positioning of u_i is *highly*

ambiguous given the observation. Therefore, original fuzzy scores defy a strictly bivalent logic. The advantage is that the three points allow alignment of linguistic hedges, sets, and metric variables through a triangular, trapezoidal, or bell-shaped function. This *filter function* maps the raw values ν_A—e.g., age in years—into fuzzy scores μ_A—e.g., membership in the set ‹YOUNG›—so that it conveys the certainty that a 16-year-old is in the set and a 36-year-old is almost so.

To map meanings onto fuzzy scores, then, the researcher needs to establish

- The raw value of the *inclusion* threshold, α. The threshold truncates any variation above α as irrelevant: for any value higher than α, the unit u_i does qualify as an instance of the set and takes 1.00 as its fuzzy score.
- The raw value of the *exclusion* threshold, β. The threshold truncates any variation below β as irrelevant: for any lower values, the unit u_i does not qualify as an instance of the set and takes the fuzzy score of 0.00.
- The raw value of the *crossover* γ, which makes the classification of u_i uncertain and corresponds to the fuzzy score of 0.50. In Zadeh's original system, the raw value of the crossover is the arithmetic mean of α and β.

7.4.1.2 Ragin's Reinvention

For QCA, Zadeh's original proposal is affected by a twofold ambiguity. First, linguistic hedges are seldom clearly ordered, and a straightforward correspondence with particular fuzzy scores can prove idiosyncratic. Second, triangular, trapezoidal, or bell-shaped relations can make each fuzzy score μ_A correspond to more than one raw scores on ν_A, which makes it hard to retrieve the raw value from the fuzzy score.

Ragin's fuzzy sets avoid these issues with a gauge that, before rendering natural language, includes both pieces of information of interest to comparatists—those of "differences in degree," and of "differences in kind" (Ragin, 2000). His filter functions are monotonic non-decreasing, which re-establishes the isomorphism of raw values, fuzzy membership scores, and selected hedges—as in Table 7.3.

The remapping of raw variables into fuzzy scores is especially illuminating of Ragin's rationale of conversion. He portrays it as an operation of *calibration*—defined as the fine-tuning of an instrument to improve the validity of its measurements. Although the concept best applies to continuous variables, the calibration rationale also informs the transformation of qualitative data into fuzzy scores (e.g., De Block & Vis, 2019). Indeed, the instrument to be fine-tuned is the filter function, whose shape can be decided using different methods (Ragin, 2000, 2007, 2008:96; Duşa, 2019).

The *indirect method of calibration* assigns the same "qualitative score" from a scale such as (*c*) or (*f*) in Table 7.3 to groups of cases with similar raw values. Then, the cases' raw scores may or may not be filtered into predicted fuzzy scores through the qualitative scores by fractional polynomial regression.

Table 7.3 Possible positions of u_i to A, and corresponding membership values μ_A

Position	μ_A (a)	(b)	(c)	(d)	(e)	(f)
Fully in	1	1	1	1	1	1
Mostly in					$^4/_5$	$^5/_6$
More in than out			$^2/_3$	$^3/_4$		$^4/_6$
More or less in					$^3/_5$	
Neither in nor out		$^1/_2$		$^2/_4$		$^3/_6$
More or less out			$^1/_3$		$^2/_5$	$^2/_6$
More out than in				$^1/_4$	$^1/_5$	
Mostly out						$^1/_6$
Fully out	0	0	0	0	0	0

Source: Ragin (2000:156, 2009)

The *direct method of calibration*, on the other hand, stipulates that the filter function is a growth curve of odds. The smoothness of the slopes is decided every time by suitable raw values for α_A, γ_A, β_A. These chosen raw scores are pegged to conventional fuzzy values, fixed at 0.953, 0.500, 0.047, respectively. The log-odds of $\mu\alpha$ are $\ln\left(\frac{0.953}{1-0.953}\right)=3$, while those of $\mu\alpha$ are $\ln\left(\frac{0.047}{1-0.047}\right)=-3$; thus, the fuzzy membership of the i-th unit with raw value ν_i is calculated as in (21) below:

$$\mu_i = \begin{cases} \dfrac{e^{3\frac{\nu_i-\gamma}{\alpha}}}{1+e^{3\frac{\nu_i-\gamma}{\alpha}}}, & \nu_i > \gamma \\ 0.5, & \nu_i = \gamma \\ \dfrac{e^{-3\frac{\nu_i-\gamma}{\beta}}}{1+e^{-3\frac{\nu_i-\gamma}{\beta}}}, & \nu_i < \gamma \end{cases} \tag{7.21}$$

Ragin's fuzzy sets can be conceived of as crisp sets weighted by a *classification error*. As such, they convey both qualitative and quantitative information, circumventing the trade-off between scales. Indeed, the crisp classification still holds with fuzzy scores, following the rule of conversion in (7.22):

$$A_i = \begin{cases} 1, & \mu_{i\in A} > 0.50 \\ 0, & \mu_{i\in A} < 0.50 \end{cases} \tag{7.22}$$

where A_i is the crisp membership of the i-th unit in the set A, while $\mu_{i\in A}$ is the fuzzy membership of the same i-th unit in the same set.

The preservation of crisp sets' qualitative information by QCA's fuzzy scores is further ensured by the convention that the crossover shall not be assigned to any

actual unit of analysis—or of dropping the 0.5-instances under the argument that they cannot bring helpful information in the analysis (Ragin, 2008; Duşa, 2019).

Furthermore, the basic rules for calculating intersection and union as in (7.6) and in (7.10) also apply to fuzzy sets. However, fuzzy scores cannot meet the axiom of strong identity (7.1); instead, they follow the more common version (7.23) below, meaning that sameness is preserved for units with the same score.

$$A_i := A_i \tag{7.23}$$

The principles of non-contradiction and excluded middle again hold with fuzzy scores in a crisp understanding, as clarified by (7.24) and (7.25):

$$\mu_{i\in\left(A\cap\bar{A}\right)} < 0.5 \tag{7.24}$$

$$\mu_{i\in\left(A\cup\bar{A}\right)} > 0.5 \tag{7.25}$$

It is worth noting that the size of a fuzzy union calculated by (7.6) is usually smaller than its crisp versions, while the size of a fuzzy intersection calculated by (7.10) is usually larger than its crisp version due to the *residuals* that fuzzy scores leave in the partition.

7.4.1.3 Fuzzy Sufficiency and Necessity

With fuzzy scores, subset relationships are established as the *containment* (Ragin, 2000; cfr. Zadeh, 1978) of membership functions.

Therefore, fuzzy-set sufficiency is captured by Eq. (7.26):

$$\mu_{i\in\omega_{\cdot}} < \mu_{i\in Y} \tag{7.26}$$

Equation (7.26) entails that, if we plot our units on a Cartesian plane defined by the membership scores in $\omega_{\cdot}$ as the x-axis and the membership scores in Y as the y-axis, if $\omega_{\cdot}$ is sufficient to Y, it distributes the units *above* the bisector in an *upper-triangular* shape.

Instead, fuzzy-set necessity corresponds to (7.27):

$$\mu_{i\in\omega_{\cdot}} > \mu_{i\in Y} \tag{7.27}$$

Equation (7.27) means that the antecedent $\omega_{\cdot}$ that is necessary to Y distributes the units *below* the bisector in a *lower-triangular* shape.

By extension, the relationship of necessity and sufficiency arises when the units' membership scores in a primitive (or implicant, or condition) equal those in the outcome, distributing the units *along* the bisector in a *linear* shape.

The *S.cons* parameter preserves its meaning with fuzzy scores, although they can blur the *recognition* of violations as the residuals $\mu_{i\in(Y\cap\overline{Y})}$ inflate their values. The *Proportional Reduction of Inconsistency* (*PRI*: Ragin, 2008; Schneider & Wagemann, 2012) has been introduced to deflate and complement the information from the *S.cons* calculated with fuzzy scores. The parameter builds on the rationale of the proportional reduction of error commonly employed to determine whether the information about *A* improves our prediction of *Y* (e.g., Menard, 1995). It reads as in (7.28):

$$PRI_{\omega_*\rightarrow Y}=\frac{\left|\omega_*\cap Y\right|-\left|\omega_*\cap Y\cap\overline{Y}\right|}{\left|\omega_*\right|-\left|\omega_*\cap Y\cap\overline{Y}\right|} \tag{7.28}$$

where the vertical bars again indicate the size of the fuzzy partition as the sum of the units' fuzzy membership scores in the partition—such that, for instance, $\left|\omega_*\right|:=\sum_{i=1}^{N}\mu_{i\in\omega_*}$.

The set-theoretical task of the *PRI* is to establish whether the conditional relationship holds, net of fuzzy residuals. It takes the same value as the *S.cons* when the size of the residuals is null $\left|Y\cap\overline{Y}\right|=0.00$. It degenerates when the units systematically display higher residuals than membership in the primitive: $\mu_{i\in(Y\cap\overline{Y})}>\mu_{i\in\omega_*}$. Last, it takes lower values than the *S.cons* when the units' residuals are non-null and lower than the membership in the primitive: $0<\mu_{i\in(Y\cap\overline{Y})}<\mu_{i\in\omega_*}$.

A *PRI* value sensibly lower than the corresponding *S.cons* points to inconsistencies that may justify the exclusion of the primitive from minimizations—or the reconsideration of gauges, conditions, or the starting hypothesis.

7.4.2 *Gauging for QCA: The Empirical Side*

Whether fine-grained membership scores properly render an *inus* factor only depends on how we construe our gauge—here, on how we set the thresholds. Thresholds elicit a solution to the problem of aligning the extension and the intension of an attribute (Quine, 1982; Sartori, 1984; Goertz, 2020).

A theory-driven approach to the problem clarifies the intension first to prevent the risk of stretching attributes beyond their meaning, which would introduce more hidden heterogeneity than would be desirable for the analysis (see Chap. 10). At the same time, thresholds may spoil the analysis when they enforce some ideal yardstick that none of the units can meet. In short, theoretical thresholds can become useless when decisions are not fine-tuned to actual diversity.

QCA scholars have developed several recommendations to balance these opposite risks. The recommendations assist the researcher in tackling three intertwined problems—namely, unit selection, the operationalization of causal properties, and

the identification of thresholds that align meanings and empirics. In actual research, the point of attack may change; however, the resulting membership scores provide a single solution to all three issues—likely, after some iteration.

7.4.2.1 Establishing the Universe of Reference

As in any technique, units of observation provide as solid an empirical ground to the analysis as the criteria for their selection. Such criteria should prevent or minimize the later rise of threats to credible results (e.g., Geddes, 1990; Goertz, 2020).

In explanatory QCA, case selection has to ensure enough diversity to capture the causal facts of interest. Thus, the criterion cannot exclusively focus on the dependent or the independent. Units selected on the outcome of interest would artificially prevent inconsistencies—thus making the validity of results undecidable. On the other hand, units selected on the factor of interest would turn it into a constant background feature and make its causal contribution undecidable. Hence, the first criterion that unit selection shall meet is the *variability* in realized states and combinations of factors.

The broadest variability follows from open universes, but open universes may endanger the preservation of meaning (i.e., Ragin, 2008). Geographical, historical, and cultural boundaries provide the closure of the units' heterogeneity required for making interpretable decisions about thresholds. Indeed, different α, β, γ may be needed to establish whether a country qualifies as <RICH>, <DEMOCRATIC>, or <EQUAL> in different world regions and time frames. Therefore, the second and related criterion for unit selection consists of finding the meaningful *scope condition* that encloses the universe of reference and ensures interpretable membership scores. In short, the correspondence of meaning and numbers comes at the cost of a restriction in the scope of the analysis—and in the generalizability of results (e.g., Goertz, 2017; Walker & Cohen, 1985; Verweij & Vis, 2021; Findley et al., 2021). The limitation, however, might not apply to the starting explanatory hypothesis, which may travel farther than its operational specifications.

7.4.2.2 Operationalizing Intension

The operation of connecting gauges and attributes meaningfully is seldom straightforward. Again, it opens to two opposite risks of providing too a specific or generic definition of an attribute (e.g., Sartori, 1984; Ragin, 2008).

Hyper-Specificity

The fallacy of composition occurs when we recognize each "token" empirical manifestation as a different property and build a plethora of conditions with too narrow an extension (e.g., Menzies, 2004; Craver & Kaplan, 2020; cfr. Chap. 10). The

problem can be solved by recognizing functional equivalences, climbing the ladder of abstraction, and gathering functionally equivalent manifestations under a single label.

Verba (1967) elaborates on the point by discussing how case-based evidence can be turned into a causal factor. From the historical report on how the eruption of Mount Vesuvius had a significant impact on the stability of the Pompeiian political system, we may identify either ‹ERUPTION› or ‹CALAMITY› as a relevant *inus* factor; however, the latter includes the former and accommodates a broader number of functionally alternative sources of disruptions, thus widening the scope of comparisons.

According to Verba, an even better operationalization shifts the attention from contextual conditions to the properties of the unit of analysis. Instead of gauging the sources of disruption, the operationalization can narrow on those resources and arrangements that make the system respond to disruption effectively. From this viewpoint, ‹RESILIENT› better contributes to an explanatory theory of political systems' stability than ‹CALAMITY›. The system attribute can apply to the Pompeiian case, but travel farther across contexts.

Hyper-Generality

The second and opposite problem arises when the properties are encompassing to the point of losing their analytic capacity.

The problem often arises when the available measure of a concept is a composite of predictors, enabling factors, proxies, outputs, and outcomes. Such assorted content can make these composites apply "*everywhere*, as any universal should" but also "to *everything*." As a result, we incur "theoretically, a 'nullification of the problem' and, empirically, what may be called 'empirical vaporization'" (Sartori, 1991; Chap. 9; cfr. Collier & Mahon, 1993).

QCA detects these composites as trivial conditions and suggests they can be dismissed. However, composites may contain relevant explanatory information. The *inus* standing of selected components can be decided by their consistency to the outcome and by minimizations. In addition or as an alternative, suitable rules of composition by disjunction and conjunction may be devised to compress sub-properties into "superconditions" (Elman, 2005; Berg Schlosser & De Meur, 2009; Goertz, 2017; Damonte & Negri, 2019).

The Problem of Missing Values

Often, available raw measures are plagued with missing values. QCA's algorithm technique cannot handle them clearly, as the units for which the value is missing would belong to two primitives. This ambiguity can be tackled by running parallel analyses to verify whether the different classifications result in different solutions. If not, the unit and its partial information would prove irrelevant. When different

classifications affect solutions—for instance, because they decide whether a primitive is realized or not—the information proves relevant, but the problem arises of how to decide between the two solutions.

Missing raw values require some credible criterion of adjudication. Alternatively, the measure can be substituted with a complete gauge of the same intension, if any. Last, the unit can be dropped from the analysis (Ragin, 2008; Basurto & Speer, 2012; Duşa, 2019). The move may increase the number of logical remainders, but remainders can be more adequately addressed with counterfactual rules in minimization.

7.4.2.3 Identifying Membership Thresholds

Thresholds explicate the rule that establishes a unit to be an instance of the set given its raw value. The default recommendation is to anchor these decisions on external theories and conventions (Ragin, 2000, 2007, 2008).

Special values of national and international policy indicators—for instance, household income to establish the risk of poverty; the share of people in an age cohort in education or training to expect a certain quality of society; the share of debt to revenue to establish the credibility of a borrower—may offer accepted anchorages to calibration decisions. However, conventional knowledge may evolve at a slower pace than actual phenomena. Under particular contingencies or within special areas, its usage for calibration may return skewed membership scores that would not survive the *RoN* test. Besides, a conventional tipping point may coincide with some units in the population, making them uninformative.

To avoid these issues, conventional knowledge can be adjusted in light of distributional considerations (Ragin, 2008). Although descriptive statistics lack qualitative meaning, considerations about quintiles seem unavoidable in large-N studies or whenever previous knowledge is wanting (e.g., Ragin & Fiss, 2017). A supplementary strategy—and consistent with the concern for non-contradictory partitions—prescribes cluster analysis to identify the raw values to be used as thresholds. The underlying rationale maintains that units close to each other belong to the same partition—and hence, that thresholds lie in the "natural gaps" between clusters.

Although long offered as a standard function for threshold setting by many software packages (e.g., Duşa, 2019), cluster analysis has driven concerns that its application might convey a deceiving sense of certitude about calibration and solutions. The risk of overconfidence can also increase when the membership scores are assigned directly following one of the scales in Table 7.3. Indeed, the researcher's classification error can always affect scoring operations in unknown directions.

To keep the risk at bay, zooming into the units around a threshold can help to support decisions with empirical knowledge when the number of cases allows it (Ragin, 2000; De Block & Vis, 2019). Frontier literature has also developed on false negatives and false positives in solutions (Braumoeller, 2015; Rohlfing, 2018) and on alternative filtering functions (Thiem, 2010). A further strategy suggests ascertaining the "robustness" of the solutions by running parallel analyses under different

perturbations of units and thresholds (Marx & Duşa, 2011; Maggetti & Levi-Faur, 2013; Duşa, 2019; Oana & Schneider, 2018).

Many of these considerations are more justified in exploratory than in explanatory applications of QCA. When the driving concern is the preservation of particular meanings, seldom different gauges can render it equally well. To witness, Ostrom's theory of corruption maintains that people's perception of ineffective monitors and sanctions drives the belief of diffused wrongdoing that invites resorting to corruption along the lines of a self-fulfilling prophecy. In testing the tenability of this theory, the indexes of inefficiency in administration often used as a proxy of corruption are less suitable gauges of the phenomenon to be explained than the measures of perceived corruption.

In explanatory usages, however, coder's biases are possible, and this possibility can be explored by simulating some systematic tendencies toward strictness, generosity, confidence, or coyness in assigning membership scores. These tendencies can be rendered by calculating the *concentration* (7.29), *dilation* (7.30), *intensification* (7.31), or *moderation* (7.32) of the original fuzzy scores (Smithson & Verkuilen, 2006):

$$\mu_{i\in A}^{Conc} = \mu_{i\in A}^{2.0} \tag{7.29}$$

$$\mu_{i\in A}^{Dil} = \mu_{i\in A}^{0.5} \tag{7.30}$$

$$\mu_{i\in A}^{Int} = \begin{cases} \mu_{i\in A}^{0.5}, & \mu_{i\in A} > 0.5 \\ \mu_{i\in A}^{2.0}, & \mu_{i\in A} < 0.5 \end{cases} \tag{7.31}$$

$$\mu_{i\in A}^{\text{mod}} = \begin{cases} \mu_{i\in A}^{2.0}, & \mu_{i\in A} > 0.5 \\ \mu_{i\in A}^{0.5}, & \mu_{i\in A} < 0.5 \end{cases} \tag{7.32}$$

These transformations expose the worsening or the improvement that coders' biases can impart to solutions. They prove that truth tables and solutions inevitably change with scoring strategies—and the intensification, by bringing the fuzzy truth table closer to its crisp version, inevitably enhances the consistency and symmetry of observed primitives. In the end, the relative fragility of findings mirrors the specificity of our operationalization —but also its local value. It counts less as a problem of the technique or the algorithm than an issue in our knowledge, models, and gauging strategies.

7.5 Summing Up

To run a credible explanatory QCA, a researcher may want to

1. *Define the outcome of interest, the causal stories about its generative process, and the conditions that make it "certain."* This step implies reviewing the theoretical and empirical literature to find testable definitions of the outcome, and identifying a convincing (type of) data-generation mechanism beneath it. Based on the mechanism, triggering, enabling, and shielding *inus* conditions can be hypothesized that, jointly given in an ideal unit, would compel the generation process and ensure it unfolds unimpeded. This bundle provides the starting *inus* hypothesis.
2. *Identify the universe of reference and the raw variables that render the hypothesis, then declare the directional expectations about each factor*. Define a scope condition for a population ensuring meaningful units' diversity. Choose the raw measures at the proper level of abstraction to render each factor as faithfully as possible. Estimate the missing values, or discard the corresponding unit. Then, declare the directional expectations about the contribution of each factor to the occurrence and failure of the outcome.
3. *Turn raw data into membership scores*. Explore the variation in the raw measures; identify thresholds; assign membership scores to instances with proper operations. Different scaling may affect the assessment of set-relationships; consider applying the same scaling. Consider whether the specification of the hypothesis may benefit from the compression of some factors; in that case, add the new superconditions to the dataset. Calculate different datasets with diluted, concentrated, moderated, and intensified scores to run parallel analyses for robustness.
4. *Assess the claim of individual consistency.* Calculate the necessity parameters for single conditions against the outcome and its negation. Identify those conditions from the starting hypothesis with *N.cons* above 0.95 and low *RoN*, and fork the analysis by running the next steps with and without them. If compressed conditions obtain better *N.cons* and *N.cov* values than the original ones, consider dropping the latter. *N.cons* and *N.cov* values can also be used to establish whether the directional expectations stand in the population.
5. *Assess the claims of sufficiency.* Build the truth tables of the positive and negative outcome, assign instances to primitives, and calculate the *S.cons* and the *PRI* of the realized primitives. Check for inconsistent instances in configurations; if found, re-run the calibration. Be it of no help, add a further condition in line with the starting hypothesis to improve the consistency of each primitive to one outcome.
6. *Minimize*. Establish the cut-off in the values of *S.cons* below which the observed primitives will not be deemed consistent with the claim of sufficiency—in case, with the help of *PRI* values—to both the positive and the negative outcome. Find the conservative, parsimonious, and plausible solutions. Consider the difference in the composition of each prime implicant from the parsimonious and the plausible solution. If new conditions appear in the latter, check whether the *S.cons* values of the plausible solution are higher than the parsimonious. Higher consistency values indicate the addition is detectably meaningful, and the plausible solution is more credible than the parsimonious. If the additional conditions in

the plausible solution do not improve the *S.cons* values on the parsimonious, consider re-running the analysis from step 5 without these additional conditions to verify the robustness of minimizations.

7. *Plot the solutions to the outcome and its negation*. Check the fitting of the instances to the upper triangular shape, assuming the shape is met when instances fall above the $y = x + 0.1$ line (Ragin, 2000). Discuss which implicants explain which instances of the outcome. Consider the unexplained instances.
8. *Return to theory*. Consider the logical relationship between the solutions to the outcome and its negation. Identify the strategies that a negative instance can adopt to reach the closer positive group.
9. *Run re-analyses and extensions for robustness*. Run the analysis with different calibrations and scope conditions, and compare the raising of contradictory configurations, the change in necessity, the differences in solutions.

You can find the example here https://doi.org/10.5281/zenodo.7117973.
Enjoy your explanatory QCA!

Suggested Readings

The full-fledged version of the original proposal remains Charles C. Ragin, 2008. *Redesigning social inquiry: Fuzzy sets and beyond.* University of Chicago Press. An updated version and close to the original proposal is Patrick A. Mello's *Qualitative Comparative Analysis: An Introduction to Research Design and Application* (Georgetown University Press, 2021). A more case-oriented version is Ioana-Elena Oana, Carsten Q. Schneider, and Eva Thomann's *Qualitative Comparative Analysis Using R: A Beginner's Guide* (Cambridge University Press, 2021).

The detailed documentation of the R functions for QCA is in Adrian Duşa's *QCA with R: A comprehensive resource* (Springer, 2019). Additional functions are in Ioana-Elena Oana and Carsten Q. Schneider's *SetMethods: an Add-on R Package for Advanced QCA* (The R Journal https://doi.org/10.32614/RJ-2018-031).

The standards of transparency in reporting QCA are detailed in Schneider, Carsten Q., Vis, Barbara and Koivu, Kendra, 2019. *Set-Analytic Approaches, Especially Qualitative Comparative Analysis (QCA)*, https://doi.org/10.2139/ssrn.3333474

Review Questions

Section 7.2

(a) What is *inus* causation?
(b) What is an *inus* machine?
(c) How are the two concepts related to directional expectations?

Section 7.3

(a) What is a literal?
(b) What is a set?

(c) What is the relationship between the membership in a set and the truth value of a proposition?
(d) What is a truth table?
(e) How many primitives has a truth table of seven literals?
(f) Construe the truth table of literal A and the ⌜not⌝ connective.
(g) What does the principle of non-contradiction say?
(h) What does the weakest link rule say?
(i) How do you calculate the membership of a unit in an intersection?
(j) Construe the truth table of literals A, B, C, D and compute the truth function of the ⌜*and*⌝ connective.
(k) What does the principle of the excluded middle say?
(l) What does the strongest link rule say?
(m) Construe the truth table of literals A, B, C, D and compute the truth function of the ⌜*or*⌝ operator.
(n) What is the consistency of sufficiency?
(o) How can the consistency of sufficiency support the assessment of underspecification?
(p) What is the consistency of necessity?
(q) How can the consistency of necessity support the assessment of overspecification?
(r) What is in a parsimonious solution?
(s) What is a hard counterfactual, and what is an easy one? In which round of minimizations are they employed?

Section 7.4

(a) How do fuzzy scores accommodate qualitative and quantitative information?
(b) What are the shapes of the filter function in Zadeh's fuzzy sets, and how do they differ from Ragin's?
(c) What is the meaning of the inclusion and exclusion points in terms of relevant and irrelevant variation?
(d) What is the rule for turning fuzzy into crisp scores? Can we reverse the transformation?
(e) The membership score of u_1 in set A is 0.3. Calculate the value of its membership in the intersection $A \cap \overline{A}$.
(f) Do fuzzy scores violate the principle of non-contradiction?
(g) The membership score of u_1 in set A is 0.3. Calculate the value of its membership in the union $A \cup \overline{A}$.
(h) Do fuzzy scores stretch the principle of the excluded middle?
(i) What is the *PRI* for?
(j) How can you ascertain the robustness of configurational solutions?
(k) Calculate the concentrated, dilated, intensified, and moderated scores of unit u_i with original membership in Y of 0.9 and in A of 0.8.
(l) Calculate the *S.cons* of each transformation from exercise 11, and order them from the strongest to the weaker. Which fares better, and which worse?

References

Álamos-Concha, P., Pattyn, V., Rihoux, B., Schalembier, B., Beach, D., & Cambré, B. (2021). Conservative solutions for progress: On solution types when combining QCA with in-depth process-tracing. *Quality and Quantity*. https://doi.org/10.1007/s11135-021-01191-x

Amenta, E., & Poulsen, J. D. (1994). Where to begin. A survey of five approaches to selecting independent variables for qualitative comparative analysis. *Sociological Methods and Research, 23*(1), 22–53. https://doi.org/10.1177/0049124194023001002

Ansell, C., Doberstein, C., Henderson, H., Siddiki, S., 't Hart, P.: Understanding inclusion in collaborative governance: A mixed methods approach. Policy and Society 39(4), 570–591 (2020). https://doi.org/10.1080/14494035.2020.1785726.

Basurto, X., & Speer, J. (2012). Structuring the calibration of qualitative data as sets for qualitative comparative analysis (QCA). *Field Methods, 24*(2), 155–174. https://doi.org/10.1177/1525822X11433998

Baumgartner, M. (2015). Parsimony and causality. *Quality & Quantity, 49*(2), 839–856. https://doi.org/10.1007/s11135-014-0026-7

Baumgartner, M., & Thiem, A. (2020). Often trusted but never (properly) tested: Evaluating qualitative comparative analysis. *Sociological Methods & Research, 49*(2), 279–311. https://doi.org/10.1177/0049124117701487

Befani, B. (2013). Between complexity and generalization: Addressing evaluation challenges with QCA. *Evaluation, 19*(3), 269–283. https://doi.org/10.1177/1474022213493839

Berg Schlosser, D., & De Meur, G. (2009). Comparative research design: Case and variable selection. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques* (pp. 19–32). London. https://doi.org/10.4135/9781452226569.n2

Boole, G. (1853). *An investigation of the Laws of thought on which are founded the mathematical theories of logic and probabilities*. Walton and Maberly.

Braumoeller, B. F. (2015). Guarding against false positives in qualitative comparative analysis. *Political Analysis, 23*(4), 471–487. https://doi.org/10.1093/pan/mpv017

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Cartwright, N. (2017). Causal powers. Why Humeans can't even be instrumentalist. In J. D. Jacobs (Ed.), *Causal powers* (pp. 9–23). Oxford University Press.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

Clarke, K. A. (2020). Logical constraints: The limitations of QCA in social science research. *Political Analysis, 28*(4), 552–568. https://doi.org/10.1017/pan.2020.7

Colby, M. E. (1991). Environmental management in development: The evolution of paradigms. *Ecological Economics, 3*(3), 193–213. https://doi.org/10.1016/0921-8009(91)90032-A

Collier, D., & Mahon, J. E. (1993). Conceptual stretching revisited. *American Political Science Review, 87*(4), 845–855. https://doi.org/10.2307/2938818

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science., 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Damonte, A. (2013). Policy tools for green growth in the EU15: A qualitative comparative analysis. *Environmental Politics, 23*(1), 18–40. https://doi.org/10.1080/09644016.2013.817759

Damonte, A. (2021a). Gauging the import and essentiality of single conditions in standard configurational solutions. *Sociological Methods & Research, 50*(2), 683–707. https://doi.org/10.1177/0049124118794678

Damonte, A. (2021b). Modeling configurational explanations. *Italian Political Science Review, 51*(2), 18U 2–18U97. https://doi.org/10.1017/ipo.2021.2

Damonte, A., & Negri, F. (2019). Gauging fiscal worlds: How the EU countries balanced equality and wealth between 2007 and 2016. *Quality and Quantity, 53*(4), 1675–1692. https://doi.org/10.1007/s11135-018-00833-x

De Block, D., & Vis, B. (2019). Addressing the challenges related to transforming qualitative into quantitative data in qualitative comparative analysis. *Journal of Mixed Methods Research, 13*(4), 503–535. https://doi.org/10.1177/1558689818770061

De Meur, G., & Berg-Schlosser, B. (1994). Comparing political systems: Establishing similarities and dissimilarities. *European Journal of Political Research, 26*(2), 193–219. https://doi.org/10.1111/j.1475-6765.1994.tb00440.x

De Morgan, A. (1847). *Formal logic: Or, the calculus of inference, necessary and probable*. Taylor and Walton.

Duşa, A. (2019). *QCA with R: A comprehensive resource*. Springer.

Elman, C. (2005). Explanatory typologies in qualitative studies of international politics. *International Organization, 59*(2), 293–326. https://doi.org/10.1017/S0020818305050101

Findley, M. G., Kikuta, K., & Denly, M. (2021). External Validity. *Annual Review of Political Science, 24*(1), 365–393. https://doi.org/10.1146/annurev-polisci-041719-102556

Fiss, P. C., Sharapov, D., & Cronqvist, L. (2013). Opposites attract? Opportunities and challenges for integrating large-N QCA and econometric analysis. *Political Research Quarterly, 66*(1), 191–198. jstor.org/stable/23563602

Geddes, B. (1990). How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political Analysis, 2*, 131–150. https://doi.org/10.1093/pan/2.1.131

Goertz, G. (2017). *Multimethod research, causal mechanisms, and case studies: An integrated approach*. Princeton University Press.

Goertz, G. (2020). *Social science concepts and measurement: New and completely. Revised Edition*. Princeton University Press.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*(2), 81–107. jstor.org/stable/2987957

Hájek, A. (2011). Conditional probability. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of philosophy of science* (*Philosophy of statistics*) (Vol. 7, pp. 99–135). North-Holland.

Hinterleintner, M., Sager, F., & Thomann, E. (2016). The politics of external approval: Explaining the IMFs evaluation of austerity programmes. *European Journal of Political Research, 55*(3), 549–567.

Huntjens, P., Pahl-Wostl, C., Rihoux, B., Schlüter, M., Flachner, Z., Neto, S., Koskova, R., Dickens, C., & Kiti, I. N. (2011). Adaptive water management and policy learning in a changing climate: A formal comparative analysis of eight water management regimes in Europe, Africa and Asia. *Environmental Policy and Governance, 21*(3), 145–163. https://doi.org/10.1002/eet.571

Kogut, B. K., & Ragin, C. C. (2006). Exploring complexity when diversity is limited: Institutional complementarity in theories of rule of law and national systems revisited. *European Management Review, 3*(1), 44–59. https://doi.org/10.1057/palgrave.emr.1500048

Krogslund, C., Choi, D. D., & Poertner, M. (2015). Fuzzy sets on shaky ground: Parameter sensitivity and confirmation bias in fsQCA. *Political Analysis, 23*(1), 21–41. https://doi.org/10.1093/pan/mpu016

Lauri, T., Põder, K., & Ciccia, R. (2020). Pathways to gender equality: A configurational analysis of childcare instruments and outcomes in 21 European countries. *Social Policy & Administration, 54*(5), 615–863. https://doi.org/10.1111/spol.12562

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95. https://doi.org/10.1037/h0043178

Mackie, J. L. (1965). Causes as conditions. *American Philosophical Quarterly, 2*(4), 245–264. jstor.org/stable/20009173

Mackie, J. L. (1966). The direction of causation. *The Philosophical Review, 75*(4), 441–466. jstor.org/stable/2183223

Mackie, J. L. (1974). *The cement of the universe. A study of causation*. Clarendon Press.

Maggetti, M., & Levi-Faur, D. (2013). Dealing with errors in QCA. *Political Research Quarterly, 66*(1), 198–204. jstor.org/stable/23563603

Mahoney, J. (2021). *The logic of social science*. Princeton University Press.

Marx, A., & Duşa, A. (2011). Crisp-set qualitative comparative analysis (csQCA), contradictions and consistency benchmarks for model specification. *Methodological Innovations, 6*(2), 103–148. https://doi.org/10.4256/mio.2010.0037

Mello, P. A. (2021). *Qualitative comparative analysis: An introduction to research design and application*. Georgetown University Press.

Menard, S. (1995). *Applied logistic regression analysis*. Sage.

Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science, 71*(5), 820–832. https://doi.org/10.1086/425057

Most, B. A., & Starr, H. (2015). *Inquiry, logic, and international politics: With a new preface by Harvey Starr*. University of South Carolina Press.

Oana, I.-E., & Schneider, C. Q. (2018). SetMethods: An add-on R package for advanced QCA. *The R Journal, 10*(1), 507–533. https://doi.org/10.32614/RJ-2018-031

Oana, I.-E., Schneider, C. Q., & Thomann, E. (2021). *Qualitative comparative analysis using R: A Beginner's guide*. Cambridge University Press.

Pahl-Wostl, C. (2008). Requirements for adaptive water management. In C. Pahl-Wostl, P. Kabat, & J. Möltgen (Eds.), *Adaptive and integrated water management* (pp. 1–22). Springer. https://doi.org/10.1007/978-3-540-75941-6_1

Quine, W. V. O. (1982). *Methods of logic*. Harvard University Press.

Ragin, C. C. (1987/2014). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press.

Ragin, C. C. (2000). *Fuzzy-Set Social Science*. University of Chicago Press.

Ragin, C. C. (2007). Measurement versus calibration: A set-theoretic approach. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 174–198). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0008

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. University of Chicago Press Chicago.

Ragin, C. C., & Fiss, P. (2017). *Intersectional inequality: Race, class, test scores, and poverty*. University of Chicago Press.

Rihoux, B., & De Meur, G. (2009). Crisp-set qualitative comparative analysis (csQCA). In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques* (pp. 33–68). Sage.

Rohlfing, I. (2018). Power and false negatives in qualitative comparative analysis: Foundations, simulation and estimation for empirical studies. *Political Analysis, 26*(1), 72–89. https://doi.org/10.1017/pan.2017.30

Rohlfing, I. (2020). The choice between crisp and fuzzy sets in qualitative comparative analysis and the ambiguous consequences for finding consistent set relations. *Field Methods, 32*(1), 75–88. https://doi.org/10.1177/1525822X19896258

Rosenberg, A. S., Knuppe, A. J., & Braumoeller, B. F. (2017). Unifying the study of asymmetric hypotheses. *Political Analysis, 25*(3), 381–401. https://doi.org/10.1017/pan.2017.16

Sabatier, P., & Mazmanian, D. (1980). The implementation of public policy: A framework of analysis. *Policy Studies Journal, 8*(4), 538–560. https://doi.org/10.1111/j.1541-0072.1980.tb01266.x

Salmon, W. C. (2020). *Scientific explanation and the causal structure of the world*. Princeton University Press. https://doi.org/10.1515/9780691221489

Sartori, G. (1984). *Social science concepts: A systematic analysis*. Sage.

Sartori, G. (1991). Comparing and Miscomparing. *Journal of Theoretical Politics, 3*(3), 243–257. https://doi.org/10.1177/0951692891003003001

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge University Press.

Smithson, M., & Verkuilen, J. (2006). *Fuzzy set theory: Applications in the social sciences*. Sage.
Sprenger, J. (2011). Hempel and the paradoxes of confirmation. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic* (Vol. 10, pp. 235–263). North-Holland. https://doi.org/10.1016/B978-0-444-52936-7.50007-0
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. jstor.org/stable/1671815
Stiller, S. (2017). The interplay of actor-related strategies and political context: A fuzzy-set QCA analysis of structural reforms in continental welfare states. *Journal of European Public Policy, 24*, 81–99. https://doi.org/10.1080/13501763.2015.1118146
Stone, M. H. (1936). The theory of representations of Boolean algebras. *Transactions of the American Mathematical Society, 40*(1), 37–111. https://doi.org/10.1090/S0002-9947-1936-1501865-8
Thiem, A. (2010). Set-relational fit and the formulation of transformational rules. fsQCA. *COMPASSS WP Series, 2010*(61) http://www.compasss.org/wpseries/Thiem2010.pdf
Verba, S. (1967). Some dilemmas in comparative research. *World Politics, 20*(1), 111–127. jstor.org/stable/2009730
Verweij, S., & Vis, B. (2021). Three strategies to track configurations over time with Qualitative Comparative Analysis. *European Political Science Review, 13*(1), 95–111. Cambridge University Press. https://doi.org/10.1017/S1755773920000375
Walker, H. A., & Cohen, B. P. (1985). Scope statement: imperatives for evaluating theory. *American Sociological Review, 50*(3), 288–301. jstor.org/stable/2095540
Wittgenstein, L. (1922). *Tractatus Logicus Philosophicus*. London, UK.
Zadeh, L. A. (1968). Fuzzy algorithms. *Information and Control, 12*(3), 94–102. https://doi.org/10.1016/S0019-9958(68)90211-8
Zadeh, L. A. (1978). PRUF a meaning representation language for natural languages. *International Journal of Man-Machine Studies, 10*(4), 395–460. https://doi.org/10.1016/S0020-7373(78)80003-0

Chapter 8
Causal Inference and Policy Evaluation from Case Studies Using Bayesian Process Tracing

Andrew Bennett

Abstract Case studies enable policy-relevant causal inferences when experimental and quasi-experimental methods are not possible. Even when other methods are possible, case studies can strengthen inferences either as a standalone method or as part of a multimethod research design. The chapter outlines the case study method of process tracing (PT), which is a within-case mode of analysis that builds upon Bayesian logic to make inferences to the best explanation of the outcomes of single cases. The chapter locates the epistemological basis of PT in the development and testing of theories about the ways in which causal mechanisms operate to generate outcomes. It then defines PT and outlines best practices on how to do it, illustrating these with examples of case study research on the COVID pandemic. The chapter then outlines the comparative advantages of PT vis-à-vis other methods, and identifies the kinds of research questions and research contexts for which PT is most useful. This leads to a brief discussion of two methodological innovations: formal Bayesian PT and the use of causal models in the form of Directed Acyclic Graphs to assist PT and integrate qualitative and quantitative evidence. The chapter concludes with the strengths and limits of PT.

Learning Objectives
After reading this chapter, you should be able to:

- Explain the epistemological basis of PT and its focus on theories about causal mechanisms.
- Carry out PT on a case study and use evidence from that case study to update your initial estimates of the likelihoods that alternative explanations of the outcomes of the case are true.
- Follow best practices of PT.
- Identify the kinds of research questions and contexts in which PT is most useful.
- Understand the Bayesian logic that underlies PT inferences.
- Understand the strengths and limits of PT as a method of causal inference.

A. Bennett (✉)
Georgetown University, Washington, DC, USA
e-mail: BennettA@Georgetown.edu

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_8

8.1 Introduction

Policymakers often need to assess the likely outcomes of alternative policies. To do so, they frequently need to develop causal understandings of past outcomes in situations where few cases exist and experiments are not possible for ethical or financial reasons. Process tracing (PT), a technique of within-case analysis analogous to detective work or medical diagnosis, is a key method of causal inference in individual cases. The goal is to explain the outcome of a single case, and as in detective work, the researcher can build upon both "suspects" (theories that provide potential alternative explanations for the outcome of a case) and "clues" (evidence or diagnostic tests).

Case studies have a long history—implicitly, they have been the primary method for historians and political observers since the Greek historian Thucydides wrote his chronicles in the fifth century BC. Many case studies have been done without much methodological rigor, however, which has given case study methods a bad reputation in some fields of research. In the past two decades, methodologists in political science and sociology have greatly improved and systematized case study methods, particularly the method of PT. This includes efforts to both refine case study methods and disseminate them to researchers through organizations, such as the American Political Science Association's section on Qualitative and Multimethod work, and training programs, including those sponsored by the Institute for Qualitative and Multimethod Research (IQMR) at Syracuse University, the European Consortium for Political Research (ECPR), the Global School on Empirical Research Methods (GSERM) at the University of St. Gallen, summer schools at the University of Oslo and the University of Essex, and MethodsNet.

The present chapter gives an overview of PT and recent innovations in this method. It begins with a discussion of the epistemic assumptions of PT, building on Daniel Little's Chap. 2 in this volume. It then defines PT and outlines best practices on how to do it, illustrating these with examples of case study research on the COVID pandemic. Next, the chapter assesses the comparative advantages of PT vis-à-vis other methods, including some of those addressed in the other chapters in this volume. This section also identifies the kinds of research questions and research contexts for which PT is most useful. The chapter then outlines two new developments in PT methods: formal Bayesian PT, and the use of causal models in the form of Directed Acyclic Graphs to assist in PT and to integrate qualitative and quantitative evidence. The chapter concludes with the strengths and limits of the method.

8.2 The Epistemic Foundations of Process Tracing

For policy purposes as well as academic theoretical progress, we need causal knowledge: what will be the outcome if we try policy X or if X happens in the world? Yet all research methods confront what has been called the "fundamental problem of

causal inference": we cannot rerun history after trying policy X, or after X happens in the world, and observe the outcome in the absence of X, while holding all other variables and historical developments constant.

Although no method can fully surmount this problem, scholars have outlined four general approaches to causation and associated methodological approaches to causal inference: regularity, counterfactual analysis, manipulation/experiments, and the causal mechanism account (Brady, 2008; see Chap. 2). The regularity approach, which Henry Brady calls "neo-Humean" after the philosopher David Hume, focuses on what Hume called "constant conjunction," or what we now call correlation as the key to scientific explanation (Brady, 2008). The well-known limitation of this approach is that correlation does not equal causation. Even when observational data is plentiful, and robust correlations convince us that some causal relationship probably exists, the nature of the process that generates the correlations may be unknown, and the direction of causation—does A cause B, or does B or the expectation of B cause A—is not always certain. Statistical analyses also face the "ecological inference problem": even if a correlation is causal, it does not necessarily explain any individual case in the population under study. A medicine could be helpful on average, for example, and at the same time be lethal to those who have an allergy to that medicine.

The counterfactual approach, and associated "potential outcomes" methods, posit that something is a cause if it satisfies the following: "if A then B, if not A then not B" (or, if not A then B does not happen in the same way, at the same time, or with the same magnitude). This definition of causation is intuitively appealing as a kind of common-sense understanding of causation, but it is more a thought experiment than a method of inference because we cannot observe counterfactual outcomes. In addition, while counterfactuals offer an intuitively appealing account of causation, they are also intuitively unsatisfying, and a weaker guide to policy choices in other cases, if they lack some account of the process through which the observed outcome arose (and that through which the unobserved counterfactual could have arisen).

The manipulation or experimental approach works to get as close as possible to observing the counterfactual outcome. It does so by selecting a "control" case or unit (or many randomly selected control cases or units) on which no manipulation is performed, and comparing the outcome to that of a case or unit to the outcome of a case that is as similar as possible to the control unit except that it has been subject to some treatment (or if there are many randomly selected cases, a comparison is made to a randomly selected treatment group).

This gets around some of the limitations of observational statistical analyses, but experiments have many demanding requirements or assumptions that must be met to be internally and externally valid. By one account, 26 requirements must be met for an experiment to allow a valid causal inference, including that random assignment has been properly done, that the proper statistical test is applied, that the sample size is sufficiently large, that there is no "compensatory rivalry" (which can happen if experimental subjects find out which group they have been assigned to and try harder to achieve a favorable outcome), and that there are no treatments that

occur apart from the specific one under study (Cook, 2018). Even when these assumptions are met, an experiment may or may not get us much closer to understanding the processes that generate the observed outcome(s), which limits our ability to anticipate the scope conditions under which the causal relationship holds. In addition, for many important policy challenges, experiments are impractical, a point elaborated below. Even when field experiments are possible or historical processes provide "natural" experiments with nearly random assignment of individuals to some "treatment," experiments outside of a controlled laboratory setting introduce many potential confounding variables that make it difficult to satisfy the assumptions necessary for causal inference.

The fourth approach, focusing on causal mechanisms and their capacities, provides the epistemological basis for PT (see Chap. 2, herein). In one much-cited definition, causal mechanisms can be thought of as "entities and activities, organized such that they are productive of regular changes" (Machamer et al., 2000). Causal mechanisms are the ontological entities in the world that generate the outcomes we observe, and we attempt to model these mechanisms with theories. This approach is consistent with and, in some sense, more fundamental than the others outlined above, as it includes a focus on the activities or processes that create correlations, that make experiments work, and that explain both actual and, if we could observe them, counterfactual outcomes. It is the regularity of causal mechanisms, or what some have called "invariance," that gives them explanatory power.[1] Put another way, causal mechanisms cannot be "turned off" when the conditions that enable their operation exist.

Unlike some approaches to explanation, the causal mechanisms view rejects "as if" theoretical assumptions, or assertions that theories need not be consistent with more micro-level processes as long as these theories are predictively accurate "as if" their stated or implicit micro-mechanisms were true. In a causal mechanisms approach to explanation, theories must be consistent with the evidence at lower levels of analysis or smaller slices of space and time. We may, for pragmatic reasons, consider a simplified theory adequate for some policy purposes even if it does not give details on micro-level processes, but we do so knowing that a theory that is more consistent with the details at the next level down has greater accuracy and might lead to more nuanced policy prescriptions. The 1960s theory that "smoking can cause cancer," for example, was sufficient for the public health policy advice "don't smoke," even though the detailed processes relating smoking to cancer were unknown at the time. We now have a more detailed theory about smoking and cancer that allows more precise policy prescriptions, such as "people with a mutation at a specific region on chromosome 15 are at a particularly high risk of cancer if they smoke." Theories on macro-level social processes and outcomes can be useful, and for some purposes, it may be more efficient to do PT at the macro level, but if macro-level theories work through lower levels of analysis like individuals' choices,

[1] "Invariance," as used here, does not exclude probabilistic causal relations; it can include probabilistic relations that are in some way bounded (Waldner 2012, 2016).

they must still be consistent with the processes through which those choices are made to be considered as accurate as possible.

PT exploits this aspect of mechanistic explanations by generating and assessing evidence, sometimes in detailed slices of space and time, on the explicit or implicit processes hypothesized by alternative explanations for the outcomes of individual cases. It thus takes advantage of two sources of evidence and inference that Hume did not include as core features of his constant conjunction account: *contiguity* and *sequencing*. Contiguity gets at entities in spatial proximity, bumping into each other or exchanging information—in social phenomena, who said or did what to whom. Sequencing uses the order in which things happened to help make inferences to the best explanation of the outcomes of cases—although it can be empirically hard to tell which of two parties escalated a confrontation, for example, the order in which it happened matters in explaining the outcome.

The focus on evidence on hypothesized processes raises three challenges for PT: how far down must we go into the details of processes? when should we stop gathering evidence? and how far back in time should we go to provide adequate explanations? Unfortunately, while Bayesian logic, outlined below, provides answers to these questions, they are rather general: we stop pushing into more detailed observations, gathering additional evidence, or probing earlier points in time when we think it is unlikely that doing so will change our confidence in the likelihood of alternative explanations sufficiently to be worth the effort it would entail. Put another way, process tracers balance two risks:

1. Of stopping the collection and analysis of evidence too soon, when just a little more effort would have provided evidence that would convince us of a different explanation, and
2. Of stopping too late, expending effort that does not change our confidence in alternative explanations of the outcome.

On a more pragmatic level, at some point, social scientists leave the study of more detailed social and psychological processes to other fields of study that have the skills and equipment to gather and assess evidence on these processes: cognitive psychology, neuroscience, microbiology, and so on. But we should—and do—pay at least some attention to the research at these lower levels of analysis because findings inconsistent with our theories indicate that we need to modify those theories. In the fields of economics and political science, for example, numerous theories build on research[2] that demonstrates how human decision-making often involves cognitive biases that depart from the assumptions of earlier rational choice models.

[2] Studies of the biological basis of emotions, and the effect of emotions on decision-making, are at an earlier stage of development, but are starting to gain notice in the social sciences as well.

8.3 Process Tracing Best Practices and Examples from COVID Research

8.3.1 Definition of Process Tracing

PT is the gathering and "analysis of evidence on processes, sequences, and conjunctures of events within a case for the purposes of either developing or testing hypotheses about causal mechanisms that might causally explain the case" (Bennett & Checkel, 2015:7).

Bayesian logic is the underlying foundation of PT. Bayesianism in PT treats probabilities as degrees of belief in alternative explanations.[3] In this approach, we use our existing background knowledge to form initial degrees of belief in alternative explanations of the outcome of a case (called the "priors"), and then analyze evidence to form updated degrees of belief, now conditioned on the evidence (called the "posteriors"). The relative probability of evidence under the explanations is called the "likelihood" (or, when comparing two explanations, the "likelihood ratio"). Bayesianism uses the laws of probability to convert the likelihood of the evidence conditioned on the explanations to the posteriors, or the likelihood of the explanations conditioned on the evidence.

In mathematical symbols, Bayes Theorem outlining this process of updating can be expressed as in Eq. (8.1):

$$Pr\left(P \mid k\right) = \frac{Pr(P)Pr\left(k \mid P\right)}{Pr(P)Pr\left(k \mid P\right) + Pr\left(\sim P\right)Pr\left(k \mid \sim P\right)} \tag{8.1}$$

where

- $Pr(P|\, k)$ is the posterior or updated probability of proposition P given (or conditional on) evidence k.
- $Pr(P)$ is the prior probability that proposition P is true.
- $Pr(k|\, P)$ is the likelihood of evidence k if P is true (or conditional on P).
- $Pr(\sim P)$ is the prior probability that proposition P is false.
- $Pr(k|\, \sim P)$ is the likelihood of evidence k if proposition P is false (or conditional on $\sim P$).

A mathematically equivalent equation, known as the "odds," form Bayes Theorem, which in some ways in easier to work with, is as follows:

$$Posterior\,Odds\,Ratio = Likelihood\,Ratio \bullet Prior\,Odds\,Ratio$$

[3] In frequentist statistics, by contrast, probability represents the limit of an event's relative frequency in many trials.

where the Likelihood Ratio is the probability of finding evidence k conditional on P being true divided by the likelihood of k conditional on P being false. In the notation of probability, the equivalent equation reads as in (8.2):

$$\frac{Pr\left(P \mid k\right)}{Pr\left(\sim P \mid k\right)} = \frac{Pr\left(k \mid P\right)}{Pr\left(k \mid \sim P\right)} \bullet \frac{Pr(P)}{Pr\left(\sim P\right)} \tag{8.2}$$

An intuitive way to understand Bayesian logic is to think of the strength of evidence, or the relative likelihood of finding a particular piece of evidence under alternative explanations. Evidence that is much more likely under one explanation than under another has high probative value. We already have a colloquial language for the strength of evidence (Van Evera, 1997: 31–32): evidence can constitute "smoking gun" tests, "hoop" tests, "doubly decisive" tests, or "straw in the wind" tests.

- A *smoking gun* piece of evidence is information that strongly affirms an explanation if the evidence proves to exist, but only weakly undermines that explanation if the evidence is not found. The metaphor here is that if a smoking gun is found in the hand of a murder suspect immediately after a shot is heard and the victim's body falls, then that suspect is very likely to be the murderer. The failure to find a smoking gun in the hand of a suspect, however, does not exonerate that suspect.
- *Hoop tests* involve strong evidence that is asymmetric in the other direction. Passing a hoop test means an explanation is still a viable candidate, but it only slightly increases the probability that the explanation is true. Failing a hoop test, on the other hand, greatly undermines our confidence in an explanation. If a murder suspect was in a different city from the victim at the time of the murder, for example, the suspect is exonerated, as the "guilty" hypothesis has failed a hoop test. But finding that the suspect was in the same city as the victim does not greatly incriminate the suspect, as many people were in the city at the time.
- *Doubly decisive tests* are symmetrical: they are strong at both affirming one explanation and casting doubt on others. An example here is a bank video camera that catches the face of a robber, incriminating them and exonerating others at the same time.
- *Straw in the wind tests* are symmetrical but weak—in court cases, we refer to them as "circumstantial evidence." The labels and descriptions of these four kinds of evidence are useful for teaching and understanding Bayesian logic, but it is also important to note that they are points on a continuum: the relative probability of evidence under alternative explanations can range from zero to one, and evidence can have different degrees of (a)symmetry.

8.3.2 How to Do Process Tracing

A brief outline of how to do PT is as follows:

- First, identify the dependent variable or outcome to be explained and develop some candidate theories that might explain the outcome of interest, together with their associated independent variables.
- Second, after gaining at least preliminary knowledge of the values of the independent and dependent variables of cases in the population of interest, select the case or cases on which to do PT.[4] There are many rationales for different types of case selection in small-n research, depending on the research objective. While a full discussion of case selection is beyond the scope of the present chapter (see Gerring & Seawright, 2008), as one example, if the goal is to try to identify processes or variables omitted from extant theories or models, it can be useful to study an outlier or deviant case that does not fit existing theories or statistical models.
- Third, after selecting the case or cases for PT, revisit the initial candidate theories and develop a more precise set of mutually exclusive and exhaustive potential explanations of the outcomes of the particular cases to be studied. This might include some potentially causal features of the individual cases that were not initially considered among the general candidate theories.
- Fourth, make a preliminary estimate of the likelihood that each explanation is true (the "prior" in the Bayesian logic that underlies PT).
- Fifth, derive the observable implications of each alternative theory for each case, asking: "what specific and concrete processes must have operated, in what sequence, if this theory explains the case, and what kind of potentially accessible evidence would those processes leave behind? What evidence would be true if each theory is not a valid explanation of the outcome of the case?"
- Sixth, gather the evidence and weigh its likelihood under the alternative explanations. When evidence is more likely to be true under one explanation than under the others, it increases our confidence that the first explanation is true. The most powerful kind of evidence is that which is far more likely under one theory or explanation than the others. Such evidence allows the researcher to strongly

[4] In contrast to statistical methods, random selection of cases is inadvisable in small-n research, and it is best to select cases for study with at least preliminary knowledge of the values of their independent and dependent variables. Cases that are positive on an independent variable of interest and positive on the outcome of interest (positive-positive cases) present potential opportunities to examine whether and how/through what processes or mechanisms the independent variable generates the outcome. Positive-negative cases are cases in which a hypothesized variable does not lead to a positive outcome can clarify the scope conditions of that variable. Negative-positive cases show paths to the outcome that do not involve the independent variable whose value is negative. Negative-negative cases provide less useful information. One should not study nuclear weapons proliferation, for example, by looking at countries that have neither a nuclear power program nor a close ally that might share nuclear technology and that (unsurprisingly) do not have nuclear weapons.

update their degrees of confidence in alternative possible explanations for the outcome.
- Finally, weigh the totality of the evidence, including both strong and weak evidence, and update the prior estimate of each explanation's likelihood of being true to produce a new posterior estimate.

Thus far, this account outlines the deductive side of PT. In addition, PT has an inductive side. Any unanticipated evidence that appears to perhaps play a causal role but does not fit any of the candidate explanations might provide the basis for a new explanation of the case. When a researcher adds a new alternative explanation, it is necessary to re-estimate the priors of the revised set of explanations, re-estimate the likelihood of evidence under each explanation relative to the others, and re-weigh the totality of the evidence to update the likelihood that each of the alternative explanations is true.

Bayesian logic in PT helps dispel a common misconception about the validity of different kinds of iterations between theories and evidence. Methodologists often argue that a researcher cannot develop a theory from a case and then test it against that same case. There is a good rationale for this injunction in frequentist statistical methods, as a theory derived from correlations found in a population sample cannot legitimately be tested against that same population sample, as the probability of disproving the new theory is zero. Using Bayesian logic in PT, however, makes it possible to derive a theory from a piece of evidence and then test that theory in the same case (Fairfield & Charman, 2018). There are two reasons for this, one incontrovertible and one more contestable. The incontrovertible reason is that it is often possible to develop a theory from a case and then to test it against different, independent, and heretofore unexamined evidence from the case that could still prove the new theory to be wrong. Detectives and doctors do this all the time—a doctor might find one piece of diagnostic evidence that suggests a patient might be afflicted by a disease the doctor had not previously considered, and this insight can lead to additional diagnostic tests on the same patient. If the new tests are based on biological relationships that are independent of the first test, they can either affirm or disconfirm the new candidate diagnosis. It would be nonsensical to argue that the new diagnosis should be tested on a different patient to find out why the first patient is ill.

The second rationale for developing and testing a theory in the same case is more ambitious and contestable—it argues that it is legitimate to derive a theory from a piece of evidence in a case and to claim that this *same evidence* can be a severe test of the theory. In Bayesianism, it does not matter whether one first identifies an explanation and then assesses the likelihood of evidence under that explanation relative to rival explanations, or first derives a theory from evidence and then assess the relative likelihood of that evidence vis-à-vis the new explanation and its rivals. Evidence that is consistent with one explanation and inconsistent with its rivals is strong evidence in favor of the explanation, no matter when or how the explanation was derived (Fairfield and Charman, 2022). To use an analogy, if a detective thought an aggrieved business associate was the most likely suspect in a robbery, but then found a video recording of the crime scene showing a neighbor whom she had not

previously suspected carrying out the crime, the very evidence that turned attention to the new suspect would also be powerful evidence for a conviction. The counterpoint to the unqualified application of this view is that humans are subject to potential confirmation bias, and it may be harder to objectively assess the likelihood of less definitive evidence under alternative explanations once the evidence is known to be true. Either way, Bayesian logic dictates that when we develop a new explanation or theory, we have to go back and re-evaluate all the evidence we gathered earlier, assessing its likelihood under the new theory in comparison to its likelihood under the theoretical explanations we had already considered.

8.3.3 Best Practices in Process Tracing

This chapter outlines, in the section below, on new and future developments, more recent and formal Bayesian ways of carrying out PT. Here, it turns to pragmatic advice about best practices in both informal and formal Bayesian PT. These practices are summarized in Table 8.1 (from Bennett & Checkel, 2015:21), and briefly elaborated below.

8.3.3.1 Cast the Net Widely for Alternative Explanations

It is important to consider a wide range of alternative explanations. Considering a few additional explanations that may quickly prove to be weak and deserving only of a footnote risks spending additional time and effort, but leaving out a viable explanation skews the analysis of the likelihood of the evidence and jeopardizes inferences from a case study. How do we know whether we have considered a

Table 8.1 Best practices in PT

1. Cast the net widely for alternative explanations
2. Be equally tough on the alternative explanations
3. Consider the potential biases of evidentiary sources
4. Take into account whether the case is most or least likely for alternative explanations
5. Make a justifiable decision on when to start
6. Be relentless in gathering diverse and relevant evidence, but make a justifiable decision on when to stop
7. Combine process tracing with case comparisons when useful for the research goal and feasible
8. Be open to inductive insights
9. Use deduction to ask: 'If the explanation is true, what will be the specific process leading to the outcome?'
10. Remember that conclusive process tracing is good, but not all good process tracing is conclusive

Source: Bennett and Checkel (2015)

sufficiently wide range of alternative explanations? I present here several "checklists" of common sources of potential social explanations as a pragmatic guide.

First, we can look to "off-the-shelf" theories academics have applied to similar questions, participants' and stakeholders' explanations for events and outcomes, historians' and area and functional experts' explanations, and the implicit or explicit explanations offered by news reporters (Bennett & Checkel, 2015: 23).

Second, the literature on quasi-experiments and program evaluation identifies many general explanations to consider. These include the following[5]:

- *Theory of change*: the implicit or explicit theory that is the basis for a policy that seeks a change in outcomes.
- *History*: exogenous events (events outside of the scope of the theories or explanations that a researcher is applying to a case) during the period under study that can affect outcomes (such as economic cycles, elections, natural disasters, wars, etc.).
- *Maturation*: individuals might go through aging processes that improve or degrade outcomes or policy effects over time.
- *Instrumentation*: changes in measurement instruments or technologies can affect the assessment of outcomes.
- *Testing*: exposure to testing or assessment can change the way stakeholders respond to events or policies.
- *Mortality*: there may be selection bias regarding which stakeholders or recipients drop out of a population being studied.
- *Sequencing*: the order in which events happen or program treatments are implemented may affect outcomes.
- *Selection*: if acceptance into a program or population is not random—for example, if the program chooses to address the easiest cases first (low-hanging fruit) or the hardest cases first (triage), there can be selection bias.
- *Diffusion*: if stakeholders interact with each other, this can affect results of a policy or program.
- *Design contamination*: competition among stakeholders can affect outcomes; those not selected as beneficiaries of a policy might try harder to improve their own outcomes, or they might become demoralized and not try as hard to succeed.
- *Multiple treatments*: if governments or other organizations are administering programs at the same time, or if a program being evaluated includes multiple treatments this can affect outcomes.

A third checklist of explanations to consider includes four kinds of agent–structure relations: (1) agents affecting structures; (2) structures enabling or constraining agents; (3) agent to agent interactions; and (4) structure to structure relationships (like demographic change). These four kinds of agent–structure relations intersect with three broad families of social and political theories focused on (1) ideas/

[5] Many of these are discussed in Shadish et al. (2002); this same list is included, nearly *verbatim*, in Bennett, forthcoming.

identities/social relations; (2) material resources and incentives; and (3) institutional transactions costs/functional efficiency. The resulting matrix encompasses 12 common kinds of theories. For example, the functional efficiency family of theories includes agents emulating other agents whom they view as successful, structures selecting out efficient agents as in evolutionary selection, functional competition among agents creating market or balance of power structures, and structure to structure processes like adverse selection (see Bennett, 2013; Bennett & Mishkin, 2023, for elaboration).

It is important to note that the requirement for mutual exclusivity among candidate explanations is often misunderstood (Bennett et al., 2021, cfr. Zaks, 2020). Mutual exclusivity can always be set up by explanations that point to different independent variables as the primary or most important variable in determining the outcome—only one variable can be the main one. It can also take the form of explanations that draw on different variables, but this does not have to be the case. Mutual exclusivity does not require that explanations be monocausal, and it does not prohibit explanations that draw on some or even all of the same variables. Explanations can involve as many variables as a researcher wants, in any functional forms or relationships the researcher wants to specify. They can also use exactly the same variables but just pose different possible functional relations among them. For example, an internal combustion engine needs four things to function: fuel, oxygen, a spark, and compression. These same four things could produce failure to function in different combinations or functional relationships. It may be that an engine does not turn over because the spark plug and piston rings are both a bit worn, the fuel is low octane or has some contaminants, and the air intake is a bit clogged, in such a way that improving any one of these would be enough to get the engine to turn over. Or maybe, two of these components are fine and two are just faulty enough that together they prevent the engine from turning over.

In addition, the aspiration or claim to have achieved an exhaustive set of alternative explanations is always provisional. We can never be sure that the candidate explanations are exhaustive because it is always possible that the true explanation is one we have not considered or discovered. We cannot include an explanation we have not conceived. This is one reason that Bayesians are never 100% confident that they have identified the correct explanation for an outcome.

8.3.3.2 Be Equally Tough on the Alternative Explanations

It is tempting to pick a "favorite" explanation early in a research project, but it is important to resist this temptation, as it can lead to confirmation bias. The alternative explanations should be plausible—if they are not plausible, they need to be reformulated or other explanations need to be considered. One of the ways that rigorous methods work is that they help us, or even force us, to guard against our own confirmation biases.

In PT, this takes the form of thinking through the observable implications for *all* of the hypotheses. This includes asking for each explanation "what would be the

observable implications about the process and sequence in the case if this explanation is true"—a question that comes naturally due to the way our brains work. It also includes asking "what would be true if this explanation is false"—a question we might overlook if PT methods did not require us to address it.

It is also important to do PT in relatively equal depth on each of the alternative hypotheses. Otherwise, there is an inclination to favor one hypothesis or another and to keep looking for confirming evidence for that explanation until you find it, and to stop looking for PT evidence on the alternative explanations after finding one or a few pieces of evidence that make them less likely.

8.3.3.3 Consider the Potential Biases of Evidentiary Sources

Documentary records can be biased by the preferences or instrumental goals of the people who made them regarding what they want to record, keep, and make available. Interviewees can have instrumental goals or motivated biases as well. They can also have unmotivated biases—recalled memories can be accurate, and the interviewee may have had access to some information streams and not others at the time of the events being studied. One way to take such potential biases into account is to discount the weight of evidence that could be subject to these biases.

8.3.3.4 Consider Whether the Case Is Most or Least Likely for Alternative Explanations

This recommended practice relates to the estimation of the case-specific priors on the alternative explanations.

When an explanation has a high prior (a most-likely case), but there is strong evidence in the case that the explanation is not correct, this might not only affect our explanation of the case at hand—it might lead us to narrow the scope conditions of the failed explanation and lower its prior for similar cases. Conversely, if the evidence from a case strongly supports an explanation that had a low prior, this might lead us to widen the scope conditions of this explanation and increase its prior for similar cases.

It is also useful at times to pick cases in which some of the explanations usually offered for the kind of case being studied simply cannot apply because their key variables or enabling scope conditions were not present. This can simplify the PT on such cases as it reduces the number of explanations on which PT is necessary.

8.3.3.5 Make a Justifiable Decision on When to Start

As discussed above in the section on epistemology, there is no general rule for selecting the temporal starting point for a case study. Often, it is useful to start at a critical juncture at which a key choice was made among alternative policies or at

which a strong exogenous shock occurred. But the choice of a temporal starting point also depends on whether we want to study deep, structural, and often, slow-moving causes or shorter-term, proximate causes that often relate more to agency than to structures.

Either way, the researcher must balance the costs and risks of going too far back in time, which increases the time and effort required for the PT, versus those of not going sufficiently far into the past, which risks overlooking important earlier causes that set in motion later mediating causes that explain less of the variation in outcomes across cases.

8.3.3.6 Be Relentless in Getting Diverse Evidence, but Make a Justifiable Decision on When to Stop

Here again there is no precise general rule: the researcher must balance the costs and risks of stopping the collection of evidence too soon, when a little more evidence could have greatly changed our confidence in the explanations, versus those of stopping too late, which leads to wasted time and effort and little additional updating on the alternative explanations.

Bayesian logic adds a little more specificity to this broad advice, as it indicates that after you have examined a lot of the same kind of evidence, each additional piece of that kind of evidence has a low probability of surprising you or pushing you to update your beliefs on the likelihoods that alternative explanations are true. This is because similar evidence has already been taken into account or used for updating. However, different kinds of evidence that have not been so exhaustively examined are more likely to lead to significant updating on the alternative explanations.

8.3.3.7 Combine PT with Case Comparisons if Relevant

While PT is a within-case method, it can be fruitfully combined with comparative case studies to strengthen causal inferences and clarify the scope conditions of explanations. A particularly powerful combination is the use of PT on "most-similar" and "most-different" cases.

Most-similar cases are the same (or at least roughly the same)[6] in the values of all but one of the independent variables and they have different values on the dependent variable. This provides some evidence that the difference on the one independent may cause the difference on the dependent variable, but this inference is provisional, since there may be other potentially causal factors that differ between the two cases and that are not included among the independent variables. It is thus useful to apply PT both to assess whether there is a pathway through which the

[6] Fully similar comparisons (comparisons between cases with roughly similar values on all the independent variables and on the dependent variable) are analogous to the "coarsened exact matching" that some quantitative methods use. See the Chap. 4 herein.

value on the independent variable that differs leads to the outcomes of the two cases and to assess whether the other potentially causal factors that differ do not lead to or cause the outcomes.

Conversely, a least similar case comparison involves two cases with the same value on the dependent variable and only one independent variable that has the same value. Here, PT can assess whether the common independent variable leads to the outcomes and whether other shared potentially causal factors do not.

8.3.3.8 Be Open to Inductive Insights

PT is most efficient when the researcher first develops a set of candidate explanations as described in (1) above and identifies their observable implications and the associated evidence to gather. The deductive effort this requires is quick and inexpensive compared to the field, interview, or archival work of actually gathering of the evidence. At the same time, it is important to remain alert for evidence that suggests possible causal processes not included in the initial set of explanations.

The feeling of puzzlement or surprise at an unexpected or unanticipated piece of evidence can lead to the development of a new explanation of a case for which the researcher can identify new observable implications on which to seek evidence. For this reason, it is often useful to do some initial open-ended research on a case—a process that some have called "soaking and poking"—as researchers immerse themselves in a case.

This is not the same as trying to approach a case without preconceptions, as some suggest in the grounded-theory or other traditions[7]: soaking and poking is still preceded by developing a set of theories and unexpected evidence emerges against the background of those theories. In other words, we recognize it as puzzling because it does not fit any of our candidate explanations well. In practice, there can be many iterations between the explanations and the evidence (Fairfield & Charman, 2018).

8.3.3.9 Use Deduction to Infer What Must Be True if a Hypothesis Is True

While deductively deriving the observable implications of a theory is fast and easy compared to gathering evidence, it is still challenging and contestable. Theories are usually not sufficiently detailed to immediately identify their observable implications in a particular case. This means that researchers and their readers or critics will not always agree on what the observable implications are for an explanation.

[7] While scholars in the grounded theory approach recognize that approaching a case without preconceptions is impossible, as our minds are pre-ordered by all kinds of theories and experiences, they nonetheless urge trying to do so as much as possible. The standard advice in the process tracing approach is to instead develop and be explicit about candidate explanations, drawing on the sources identified above, and use them to decide which evidence to look for.

The best that a researcher can do here is to be clear and explicit about the implications they derived from a theoretical explanation and the logic through which they derived them. It is also possible to entertain alternative readings of the implications of a theory, and to factor into the conclusions whether some or all of these proved true. If the evidence was consistent with both of two possible interpretations of a theory, for example, then the theory is likely to be true regardless of which interpretation one uses.

To identify observable implications, it is necessary to mentally inhabit the hypothetical world in which the explanation is true and imagine very concretely the specific steps, sequences, and processes through which the explanation's independent variable(s) could have generated the outcome.[8] Often, researchers are not sufficiently concrete and specific in thinking about who should have said or done what to whom when if an explanation were true. There can also be functionally equivalent substitutable steps at different points in the hypothesized process. If possession of a gun was necessary for a suspect to have committed a crime, for example, evidence that the suspect had purchased a gun is equally informative no matter whether the gun was paid for by check or credit card.

8.3.3.10 Remember Not All PT Is Conclusive

A final injunction is to remember that not all PT is conclusive. Whether it is highly conclusive depends on whether the evidence is much more likely under one explanation than under the others, and this cannot be known beforehand. In addition, even when the evidence does greatly raise the likelihood that one explanation is true, there is always some possibility that an even more accurate explanation never occurred to the researcher.

For these reasons, process tracers can never be 100% certain, and it is important to be clear about any uncertainty that remains after analyzing the evidence. In the formal Bayesian PT approach described below, this takes the form of specifying the posterior on each hypothesis in terms of an explicit probability or range of probabilities.

8.3.4 Examples from COVID Case Studies

While laboratory studies on the COVID-19 coronavirus have led to a rapid accumulation of knowledge about its biochemistry, case studies using a PT logic have been vitally important in learning about its transmission in real-world settings, where experiments are not possible. When COVID-19 first emerged as a public health

[8] Fairfield and Charman (2017) suggest this practice of mentally inhabiting the world of a hypothesis to help assess the likelihood of evidence under that hypothesis; it is also useful in deciding what evidence to look for in the first place.

concern, doctors, scientists, and government officials had limited knowledge of how the disease spread. It is easy in this instance to construct mutually exclusive and exhaustive means of transmission: (1) airborne inside only; (2) airborne inside and outside; (3) airborne inside plus transmission via common contact surfaces; or (3) airborne inside and outside plus infection through contact surfaces.[9] Epidemiologists had a range of views on what prior likelihood they should assign to each hypothesis, but in the end the priors did not matter much because powerful evidence emerged that was much more likely under explanation 1, rather than under explanations 2–4, as by far the most common means of transmission.

A key early case study came from a restaurant in Guangzhou, China, where one patron who had COVID dined on January 24, 2020 with three family members. Two other families dined at adjacent tables. Within 5 days, nine members of the three families developed COVID, with no other known exposures apart from the restaurant and subsequent within-family transmission. Close study of the restaurant seating revealed that, outside of the index patent's family, only those in the airflow path of the air conditioner that blew air across the table of the index patient developed COVID, while none of the other 83 restaurant patrons or eight staff developed COVID. The authors of a study on this case concluded that droplet transmission in the air-conditioner airflow was likely the key transmission mechanism, and recommended improved ventilation and greater table distancing in restaurants. The absence of any cases among the restaurant staff who handled the index patient's dirty dishes can be considered a failed smoking-gun test: it slightly reduces the likelihood of transmission of coronavirus through contact with surfaces of objects (Lu et al., 2020).

A later case study of a superspreader event at a choir practice in March 2020 underscored the danger of air transmission inside. Of the 61 people who attended the 2.5-hour practice, including one symptomatic index patient, 32 confirmed and 20 probable secondary COVID-19 cases occurred. The study concluded that close proximity and the act of singing led to high rates of transmission (Hamner et al., 2020).

The most definitive case study of COVID transmission, however, came from an event that provided a strong natural experiment (Shen et al., 2020). In January 2020, 128 people took two separate buses with recirculating cooling units (60 people in the first bus and 68 in the second, including a symptomatic index patient in the second bus) on a 100-minute round trip ride to a 150-minute event. Another 172 individuals attended the event but did not travel on either bus. None of the attendees wore masks. At the event, participants attended a morning service outdoors, followed by a brief lunch inside. They then returned to the same bus that had brought them, and took the same seats. Within days, 23 people on the second bus developed COVID, none of the passengers of the first bus developed COVID, and another

[9] While some lung diseases, like Legionnaire's disease, can grow in bodies of water and then most commonly infect people through inhalation of contaminated aerosols, and other diseases like Ebola are transmitted by direct contact with bodily fluids, early cases of COVID and its similarity to other coronaviruses strongly suggested transmission by air and possibly also by contact surfaces.

seven individuals who were in close contact with the index patient at the ceremony or lunch but who had not ridden by bus developed COVID. Passengers seven rows behind the index patient on the bus developed COVID, while passengers next to windows that could be opened had lower rates of infection. This case provided further smoking gun evidence of air transmission in long exposure indoors, including transmission by small and relatively far-traveling aerosol droplets as well as heavier droplets. Later studies concluded that while transmission through surface contacts could not be ruled out, and that cases of such transmission have been reported when individuals touched an object that had been sneezed or coughed upon by a COVID patient, the odds of catching COVID were approximately one case for every 10,000 surface contacts (CDC, 2021). Similarly, while the bus study did not discuss outdoor transmission and such transmission could not be ruled out due to the seven individuals who developed COVID without riding a bus, the rarity of confirmed cases of outdoor transmission has reportedly led many experts to conclude that such cases constitute only 1% of total cases and perhaps as low as 0.1% (Leonhardt, 2021).

A fourth case study indicates the high efficacy of mask-wearing to prevent COVID transmission. This study focuses on two hair stylists in Missouri who contracted COVID in 2020. While these individuals were symptomatic, they were in proximity to 139 patrons indoors. All wore masks, and none of the patrons developed COVID (Hendrix et al., 2020).

Although these four studies use the logic of PT implicitly rather than explicitly, their conclusions follow Bayesian logic. The authors intuitively used the likelihood of evidence under alternative explanations, together with the laws of probability, to update views of the likelihood of alternative COVID transmission paths in light of the evidence.

The chapter turns in the penultimate section to new methodological developments and the question of whether using the Bayesian logic of PT more formally and explicitly improves inference to the best explanation.

8.4 The "Replication Crisis" and the Comparative Advantages of Process Tracing Case Studies

8.4.1 The Replication Crisis

In the last 15 years, concerns over a "replication crisis" have swept through the social and medical sciences and the policy analysis and program evaluation communities. The crisis centers on the concern over high rates of failure in attempts to replicate peer-reviewed research findings in medicine and the social sciences, including those based on experiments as well as observational statistical studies. This does not necessarily mean that studies whose findings cannot be replicated are wrong—there are many reasons it may not be possible to replicate a study or its findings, including changes in the historical context that make it impossible to

recreate the same sample as that in the original study. Yet there is also evidence that such sample differences do not account for much of the variation in results found in replication failures (Klein et al., 2018). In addition, there are well-known methodological problems that can lead to false or overly confident conclusions that could account for the high rate of replication failures of published research. These problems include publication bias (papers supporting their hypotheses are published at a higher rate than those that do not and a higher rate than studies with null findings), "*p*-hacking" (manipulation of experimental and analysis methods, possibly unwitting, that artificially produces statistically significant results [see Chap. 4 herein, especially Sect. 4.2.3, on the model dependence of statistical analyses]),[10] "*p*-fishing" (seeking statistically significant results beyond the original hypothesis), and "HARKing" (Hypothesizing After the Results are Known, or *post-hoc* reframing of experimental intentions to fit known data).

One result of the replication crisis has been renewed emphasis on lab experiments, field experiments, natural experiments, regression discontinuity designs, and other research designs that attempt to allow causal identification. Even though experiments are among the methods that have experienced replication problems, and even though they have very demanding requirements and assumptions (especially field experiments: Cook, 2018), properly done experiments are less subject to some of the methodological limits of observational statistical studies. "Natural experiments," or real world situations in which samples of a population are assigned to or end up in two different contexts or "treatment" conditions in a way that is random or close to random, can also be powerful. Another approach that has generated increased attention is regression discontinuity designs, in which the investigator compares samples of a population just above and just below a threshold that is a cutoff at which a treatment, such as class size in public schools, is assigned (see Chap. 3 herein).

These experimental and quasi-experimental methods all have important roles to play in policy-relevant causal inferences. Researchers and journal editors have also taken steps to address the problems associated with the replication crisis. Preregistration of research designs, for example, limits the risk that researchers might unintentionally make so many modifications to their models that one model will produce a high degree of fit just by chance. Public repositories for data and replication materials are making research more transparent. Researchers have become more transparent about the assumptions behind instrumental variable and regression discontinuity designs and the conditions under which these achieve internal,

[10] The *p*-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. In other words, it tells you the probability of obtaining a test statistic as extreme or more extreme than the one calculated by your statistical test under the assumption that the null hypothesis is correct. It gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis. A *p* level of 5% has by convention been considered in many journals to be the threshold for publishing results: this means, however, that there is still a 5% chance to see a test statistic at least as extreme as the one you found if the null hypothesis was correct.

statistical, and external validity (see Chap. 3 herein, especially Sect. 3.3). Some journals are carrying out replications before publication. Matching techniques (see Chap. 4 herein) and out-of-sample testing have become more common, and some journals have de-emphasized p-values in favor of a broader range of measures of the robustness of quantitative results, or moved to *p*-values of 1% rather than 5% as the standard for publication.

Still, even with improved practices, experimental and quasi-experimental methods have limits that are different from those of PT. For many problems of interest to both scholars and policymakers—wars, epidemics, economic crashes, etc.—these methods can be subject to practical and ethical constraints and problems of internal or external validity. Lab experiments are quite different from real world conditions. Field experiments on large-scale phenomena that involve potential harm are unethical, and other kinds of field experiments may be prohibitively costly or operationally impossible. Natural experiments require a level of "as-if random" assignment to "treatment" and "control" groups that is rarely fully met except in studies of lottery winnings (Dunning, 2015). Regression discontinuity designs, as well as field and natural experiments, have the challenge of assessing potential confounding variables. In addition, all population-level analyses face the ecological inference problem.

Because case studies using PT have a different set of comparative advantages from those of experimental and quasi-experimental research designs, they are useful as both a standalone method and as a complement to these other methods in multimethod designs. Most obviously, PT is useful when policymakers are interested in understanding causation in individual cases. PT can be especially useful in studying deviant cases, or cases that do not fit existing theories, and inductively deriving and then assessing new potential explanations. But PT case studies are not just for situations in which we want to explain outcomes in one or a few cases, or when only a small number of cases exist. Even when there is a large and relatively homogenous population available for statistical or experimental study, case studies can help get closer to causal mechanisms, examining how they work down to small slices of space and time.

8.4.2 Process Tracing on Complex Phenomena

In addition, PT is useful for assessing various kinds of complexity. These include the following:

- *Endogeneity*. Endogeneity arises when there are feedback loops between the dependent and independent variables and when the direction of causation ($X \rightarrow Y$ versus $Y \rightarrow X$) is unclear. In this regard, PT helps untangle the direction of causation by focusing on the sequence of events. This helps with the assessment of which events or pieces of information came first, and what events actors may have anticipated when they took action.

- *Multiple treatments*. PT can assess multiple treatments or explanations by considering the likelihood of evidence under each of them.
- *Path dependence*. PT can untangle path dependence by examining the sequencing of events and the observable implications of theories about path-dependent mechanisms like positive returns to scale, learning by doing, first mover advantages, complementary institutions, and so on (Bennett & Elman, 2006). Most, if not all, of the research on path-dependency uses PT case studies rather than quantitative analysis.
- *Equifinality*. Equifinality is the existence of alternative paths to the same outcome. These paths may have many or no independent variables in common. Case studies using PT can chart out different paths to the outcome one case at a time.
- *Non-independence of cases*. PT can assess the evidence on mechanisms that create dependences among cases, such as learning or emulation from one case to another.
- *Potential confounders*. PT can assess whether any potential confounders identified in the course of research have a causal path to the outcome.

8.4.3 Process Tracing in Multimethod Research

PT can also be combined with other methods. One useful approach is to carry out a statistical analysis on observational data and then process trace one or a few cases to see if the hypothesized mechanisms that might explain population level correlations are evident in individual cases (Lieberman, 2005; Small, 2011). Statistical analysis can help identify outlier or deviant case, and PT on these cases may help identify omitted variables (Bennett & Braumoeller, 2022). In natural experiments, PT, on the ways in which different individuals or groups are "assigned" to or end up in the "treatment" and "control" groups, can help assess the validity of the assumptions of "as-if random assignment," unbiased dropout rates, and no unmeasured confounders (Dunning, 2015). PT can be combined with Qualitative Comparative Analysis as well, helping to identify the potentially causal processes that generate the outcomes of individual cases (Schneider & Rohlfing, 2013).

8.4.4 Process Tracing and Generalizing from Case Studies

One alleged limitation of PT case studies is their supposed inability to generalize from their results, or to achieve external validity. This issue has often been misunderstood, however (George and Bennett, 2005; Bennett, 2022). "Average treatment effects" are not the only way to conceptualize generalization, and they are not always the most useful ones. The "average treatment effect" of being born, for example, is having 1.5 X chromosomes and 0.5 Y chromosomes, an outcome that does not exist for any single person. Sometimes it is useful instead to have narrow

but strong "contingent generalizations," or generalizations that apply to only a few cases or to a specified subset of a population, such as cases that share similar values on the independent variables and the dependent variable.

Single and comparative case studies using PT may or may not allow contingent generalizations. It is impossible for a researcher to know whether and to what population or scope conditions the findings of a case study will generalize before they have developed, perhaps partly inductively, a satisfactory explanation of the case. The understanding of the causal process that emerges from PT in a case study, together with theoretical intuitions on the scope conditions in which it operates and background knowledge on the frequency with which those conditions arise, is what determines whether, where, and how a case study's findings might generalize. Charles Darwin, for example, studied several bird species on remote islands and came away with the theory of evolution, whose scope conditions include all living things. Conversely, imagine discovering that a voter favored a candidate not because of party affiliation, ideology, or any of the usual reasons, but because the candidate was the voter's sister-in-law. This would only generalize to the relatives of candidates, or perhaps more loosely to social relations not ordinarily considered to be important to voting decisions (and some voters might vote against their in-laws despite sharing their party affiliations and policy views!).

In addition, the understanding of causal mechanisms that emerges from PT on a case, to the extent that this understanding is accurate, may generalize not only to similar cases or populations but to populations and contexts different from those of the case study at hand. As noted above, Darwin's theory of evolution applied not only to birds but to all living creatures. This is different from testing or applying a theory to an out-of-sample subset of a population, as is sometimes done in statistical analyses; it is applying a theory to an out-of-population case or sample.

8.4.5 Limitations of Process Tracing

The limitations of PT correspond with the strengths of experimental, quasi-experimental methods and studies using statistical analyses of observational data. PT does not produce estimates of average effects, or correlation coefficients of independent variables. PT can shed light on how or through what mechanisms independent variables generated outcomes, but its inferences are more provisional and do not necessarily produce as confident an answer as randomized controlled experiments on whether a variable has an effect on the outcome.

8.5 New Developments in Process Tracing

Two new methodological developments are pushing the frontiers of process tracing. Both developments are outlined in forthcoming books, and both are rather technical and complex, so this chapter provides only a brief overview of each.

8.5.1 Formal Bayesian Process Tracing

Tasha Fairfield and Andrew Charman have worked out several methodological challenges to develop procedures for formal Bayesian PT (Fairfield & Charman, 2017; Fairfield & Charman, 2022). In formal Bayesian PT, researchers develop explicit numerical priors, between 0 and 1 or 0% and 100%, on the likelihood that alternative explanations are true (these could be ranges between high and low bounds, rather than point estimates). They also identify explicit numerical likelihood ratios for evidence conditioned on the alternative theories (which, again, need not be point estimates), and use these, together with Bayesian analysis of the collected evidence, to arrive at numerically explicit posterior estimates on the likelihood that alternative theories are true. Estimates of priors can be based on background information, on crowd-sourcing, or on a principle of indifference that assigns equal prior probability to all explanations. Estimates of likelihood ratios of evidence come from the theoretical logic of the alternative explanations. Researchers can check on the robustness of the posterior estimates by trying different distributions or ranges of priors and likelihood ratios.

One useful innovation that Fairfield and Charman introduce is the use of a logarithmic scale for the likelihood ratios of evidence. This simplifies the math, as logarithms allow adding the weight of different pieces of evidence rather than using multiplication. In addition, logarithmic scales, such as the decibel (*db*) scale, reflect the ways in which humans experience stimuli such as light or sound. It is intuitively easy to ask if a piece of evidence is "whispering" (30 *db*), "talking" (60 *db*), "shouting" (70–80 *db*), or "screaming" or above (90+ *db*) in favor of one explanation or another. After assigning logarithmic weights to how much each piece of evidence argues in favor of one explanation vis-à-vis another, the researcher can simply add up all of the weights to arrive at posterior estimates, just as if adding weights on a scale.

A common misunderstanding here is that the number of necessary comparisons of theories vis-à-vis the evidence becomes combinatorially large as the number of explanations grows (Bennett et al., 2021; cfr Zaks, 2021). This assumes that the likelihood for each piece of evidence under every hypothesis must be compared directly to that of every other hypothesis. In fact, it is necessary only to compare the likelihood of each piece of evidence for one explanation to that of each of the other explanations, and this implicitly compares the likelihood of the evidence under all the explanations to each other. By way of analogy, one could weigh a watermelon in terms of strawberries, and then weigh all the other fruits in a store in terms of strawberries, and this would provide the relative weight of every fruit in terms of either watermelons or strawberries.

Formal Bayesian PT has the advantage of making explicit all the judgements that are made implicitly in informal PT. This clarifies where and why an author and their readers or critics might disagree: they could disagree on the priors, on the likelihood of evidence, or on the reading of the evidence itself (one person may think a person interviewed in a research project is untruthful, for example, and another may not). Despite the advantages of formal Bayesian PT, however, its advocates do not

recommend doing it fully on every piece of evidence for every hypothesis. Doing so requires an unrealistically long and tedious write-up of research results. Researchers may find it useful, however, to carry out full formal Bayesian analysis on a small number of pieces of evidence that they consider to be the most powerful in discriminating among the hypotheses. In addition, even though it is inadvisable to fully carry out and write up formal Bayesian PT, the demonstration that it is in principle possible to do so, and the explication of the logic of doing so, help guide the reasoning of informal or partially formal Bayesian PT.

8.5.2 New Modes of Multimethod Research

A second innovation, in an article and a forthcoming book by Macartan Humphreys and Alan Jacobs, also builds on Bayesian logic and moves in a compatible but different direction. Humphreys and Jacobs use formal causal models, in the form of Directed Acyclic Graphs (DAGs), to help identify the hypothesized probabilistic dependencies among variables that enter into PT (Humphreys & Jacobs, 2015; Humphreys and Jacobs 2023; on DAGs, see also Chap. 6 herein). These authors argue, as the present chapter has, that design-based inferential approaches like experimental and quasi-experimental methods cannot be carried out on many questions that interest both policymakers and scholars, and that these methods can sometimes provide information on effect sizes without clarifying the underlying models or mechanisms. Consequently, Humphreys and Jacobs focus on model-based inference rather than design-based inference.

DAGs are models that formally represent theories in ways that make these theories' assumptions about mediating, moderating, and potential confounding variables clear and precise. Put another way, DAGs are graphical representations of Bayesian networks. Mediators are variables along the hypothesized causal path between an independent and dependent variable, so they help explain how the independent variable affects the dependent variable. Moderators are variables that affect the relationship between an independent variable and the dependent variable—they can strengthen, weaken, or negate that relationship. Confounders are variables that affect both the value of an independent variable and that of the dependent variable in a causal model, making it hard to estimate the true effect of the independent variable.

Humphreys and Jacobs argue that the core logic of their approach is most closely connected to PT and Bayesian inference, and they maintain that formally representing theories as DAGs helps guide methodological choices in both PT and quantitative analysis in ways that modify some traditional advice about how to carry out PT. Contrary to some earlier advice on case selection, for example, they argue that model-based inference demonstrates that for many inferential purposes "on the regression line" cases, or cases in which the outcome of interest occurred, are not necessarily the most informative. Optimal case selection, in their view, depends on the population distribution of different kinds of cases and the probative value of the

available evidence. They also argue that the focus on intervening causal chains (mediators) in PT can sometimes be less productive than examining moderating conditions (moderators). Finally, DAGs can inform choices in multimethod work between breadth (how many cases to study) and depth (how intensively to study individual cases).

More generally, Humphreys and Jacobs argue that their approach dissolves the usual distinctions between qualitative and quantitative research, and that it can address and integrate case level and population level queries.

8.6 Conclusions

PT methods have many uses and comparative advantages. Unlike experimental and quasi-experimental and statistical methods, they can develop inferences on alternative explanations of individual cases. As PT is always on observational evidence in single cases, its scope is not as limited by cost, ethical concerns, or availability as experiments or quasi-experiments (although, to the extent that PT involves human subjects research such as interviews, it can raise ethical issues that require approval from an Institutional review board). PT brings causal inference close to the operation of causal mechanisms, sometimes in relatively small slices of space and time. While it is the only method (other than ethnographic methods) that is possible when one or a few cases exist, it is still useful for illuminating the operation of causal mechanisms and assessing the assumptions behind other methods even when large or randomly assigned populations are available for study. It can therefore contribute to multimethod projects involving statistical, experimental, and quasi-experimental methods.

At the same time, PT has several limitations and poses a number of research challenges. Collecting the necessary evidence can be laborious and time-consuming, and the conclusions can only be as strong as the evidence allows. Identifying the observable implications of alternative explanations requires careful thought, and scholars might not agree on what rather general theories imply about such implications in particular cases. PT case studies may allow strong contingent generalizations, or they may not. More broadly, just as the strengths of PT arise in areas where quantitative methods are weak, PT is weak where these other methods are strong. PT does not produce estimates of average effects, or correlation coefficients of independent variables. It can shed light on *how* or through what mechanisms independent variables generated outcomes, but its inferences do not necessarily produce as confident an answer as randomized controlled experiments on *whether* a variable actually had any effect on the outcome. Yet precisely because the strengths and weaknesses of PT and quantitative methods offset each other, there is great value in combining these approaches in multimethod research.

Recent innovations by Fairfield, Charman, Humphreys, and Jacobs hold great promise for continuing the recent and rapid improvement of PT methods and practices.

These authors' ambitious innovations are at the cutting edge of PT techniques. As such, they have thus far been of interest mostly to methodologists and have not yet had a chance to be taken up by the much larger community of case study researchers. In short, although PT methods and practices are in some senses thousands of years old, they will continue to develop.

Review Questions

1. What are the differences among the neo-Humean regularity, counterfactual, manipulation/experiments, and the causal mechanism accounts of causation and causal inference?
2. What is a 'prior' in Bayesian terms? What is a 'posterior?' What is the 'likelihood of evidence' and how does it help us 'update' our prior to form our posterior? What kind of evidence allows the most updating?
3. Why is it important to 'cast the net widely' when formulating potential alternative explanations for the outcome of a case? How can you combine process tracing with case comparisons? Why are Bayesians never 100% sure they have the true explanation for an outcome?
4. What does it mean for alternative explanations to be 'mutually exclusive and exhaustive?' Does mutual exclusivity require that the explanations use completely different independent variables?
5. What does each of the following terms mean in the context of process tracing: Theory of change, History, Maturation, Instrumentation, Testing, Mortality, Sequencing, Selection, Diffusion, Design contamination, Multiple treatments.
6. Why is it important to pay attention to surprising or unexpected evidence from a case?
7. How can process tracing be combined with comparisons between cases?
8. What kind of conclusions can be drawn from the following case studies, and how does process tracing logic lead to these conclusions: (1) the transmission of COVID in an air-conditioned restaurant; (2) the spread of COVID at a choir practice; (3) the spread of COVID in one bus attending a ceremony and lunch but not the other bus; (4) the lack of transmission of COVID at a hair-dressing shop where two hairdressers had symptomatic COVID?
9. What are the meanings of the following terms for kinds of complexity: indigeneity, path dependence, equifinality, multiple treatments, non-independence of cases, potential confounders? How can process tracing help untangle each of these kinds of complexity?
10. How can process tracing be combined with statistical analysis of observational data? With quasi-experiments?
11. Under what conditions is it possible to generalize the results of a case study, and under what conditions is it not possible to do so?
12. How does formal Bayesian process tracing differ from less formal methods of process tracing? Is it advisable to do and write up formal Bayesian process tracing on every piece of evidence in a case study? Why or why not?
13. What is a Directed Acyclic Graph and how can it assist in process tracing and the integration of qualitative and quantitative evidence?
14. What are the limits and costs of process tracing?

References

Bennett, A. (2013). The mother of all isms: Causal mechanisms and structured pluralism in international relations theory. *European Journal of International Relations., 19*(3), 459–481. https://doi.org/10.1177/1354066113495484

Bennett, A. (2022). Drawing contingent generalizations from case studies, and process tracing for program evaluation. In M. Woolcock, J. Widner, & D. Ortega-Nieto (Eds.), *The case for case studies*. Cambridge University Press.

Bennett, A., & Braumoeller, B. (2022). Where the model frequently meets the road: Combining statistical, formal, and case study. *Methods*. arxiv.org/abs/2202.08062

Bennett, A., & Checkel, J. (Eds.). (2015). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.

Bennett, A., & Elman, C. (2006). Complex causal relations and case study methods: The example of path dependence. *Political Analysis, 14*(3), 250–267. http://www.jstor.org/stable/25791852

Bennett, A., & Mishkin, B. (2023). Nineteen kinds of theories about mechanisms that every social science graduate student should know. In H. Kinkaid & J. Van Bouwel (Eds.), *Oxford handbook of philosophy of social science*. Oxford University Press.

Bennett, A., Fairfield, T., & Charman, A. (2021). Understanding Bayesianism: Fundamentals for process tracers. *Political Analysis*. https://doi.org/10.1017/pan.2021.23

Brady, H. (2008). Causation and explanation in social science. In J. Box-Steffensmeier, H. Brady, & C. Collier (Eds.), *Oxford handbook of political methodology* (pp. 217–270). https://doi.org/10.1093/oxfordhb/9780199286546.003.0010

Centers for Disease Control, Science Brief. (2021). SARS-CoV-2 and Surface (Fomite) Transmission for Indoor Community Environments. *CDC COVID-19 Science Briefs*. https://www.ncbi.nlm.nih.gov/books/NBK570437/

Cook, T. (2018). Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: A commentary on Deaton and Cartwright. *Social Science & Medicine, 210*, 37–40. https://doi.org/10.1016/j.socscimed.2018.04.031

Dunning, T. (2015). Improving process tracing: The case of multi-method research. In A. Bennett & J. Checkel (Eds.), *Process tracing: From metaphor to analytic tool*. Cambridge University Press.

Fairfield, T., & Charman, A. (2017). Explicit Bayesian analysis for process tracing: Guidelines, opportunities, and caveats. *Political Analysis, 25*(3), 363–380. https://doi.org/10.1017/pan.2017.14

Fairfield, T., & Charman, A. (2018). The Bayesian foundations of iterative research in qualitative social science: A dialogue with the data. *Perspectives on Politics, 17*(1), 154–167. https://doi.org/10.1017/S1537592718002177

Fairfield, T., & Charman, A. (2022). *Social inquiry and Bayesian inference: Rethinking qualitative research*. Cambridge University Press.

George, A., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

Gerring, J., & Seawright, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294–308. https://doi.org/10.1177/1065912907313077

Hamner, L., Dubbel, P., Capron, I., Ross, A., Jordan, A., Lee, J., Lynn, J., Ball, A., Narwal, S., Russell, S., Patrick, D., & Leibrand, H. (2020). High SARS-CoV-2 attack rate following exposure at a choir practice – Skagit County, Washington. *Centers for Disease Control Morbidity and Mortality Weekly Report, 69*(19), 606–610. https://doi.org/10.15585/mmwr.mm6919e6

Hendrix, J., Walde, C., Findley, K., & Trotman, R. (2020). Absence of apparent transmission of SARS-CoV-2 from two stylists after exposure at a hair salon with a universal face covering policy — Springfield, Missouri. *CDC Morbidity and Mortality Weekly Report, 69*(28), 930–932. https://doi.org/10.15585/mmwr.mm6928e2

Humphreys, M., & Jacobs, A. (2015). Mixing methods: A Bayesian approach. *American Political Science Review, 109*(4), 653–673. https://doi.org/10.1017/S0003055415000453

Humphreys, M., & Jacobs, A. (2023). *Integrated inferences*. Cambridge University Press..
Klein, R., Vianello, M., Hasselman, F., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practice in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225
Leonhardt, D. (2021, May 25). A misleading C.D.C. number. *The New York Times*. https://www.nytimes.com/2021/05/11/briefing/outdoor-covid-transmission-cdc-number.html
Lieberman, E. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review, 99*(3), 435–452. https://doi.org/10.1017/S0003055405051762
Lu, J., Gu, J., Li, K., Xu, C. S. W., Lai, Z., Zhou, D., Yu, C., Xu, B., & Yang, Z. (2020). COVID-19 outbreak associated with air conditioning in restaurant, Guangzhou, China, 2020. *Emerging Infectious Diseases, 26*(7). https://doi.org/10.3201/eid2607.200764
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25. https://www.jstor.org/stable/188611
Schneider, C., & Rohlfing, I. (2013). Combining QCA and process tracing in set-theoretic multi-method research. *Sociological Methods and Research.* https://doi.org/10.1177/0049124113481341
Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin.
Shen, Y., Li, C., Dong, H., et al. (2020). Community outbreak investigation of SARS-COV-2 transmission among bus riders in Eastern China. *JAMA Internal Medicine, 180*(12), 1665–1671. https://doi.org/10.1001/jamainternmed.2020.5225
Small, M. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology, 37*, 57–86. https://doi.org/10.1146/annurev.soc.012809.102657
Van Evera, S. (1997). *Guide to methods for students of political science*. Cornell University Press.
Waldner, D. (2012). Process tracing and causal mechanisms. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science* (pp. 65–84). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195392753.013.0004
Waldner, D. (2016). Invariant causal mechanisms. *Qualitative & Multi-Method Research, Spring/Fall*, 28–33. https://doi.org/10.5281/zenodo.823264
Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.
Zaks, S. (2020). Updating Bayesian(s): A critical evaluation of Bayesian process tracing. *Political Analysis, 29*(1), 58–74. https://doi.org/10.1017/pan.2020.10
Zaks, S. (2021). Return to the scene of the crime: Revisiting process tracing, Bayesianism, and murder. *Political Analysis*. https://doi.org/10.1017/pan.2021.24

Suggested Reading

Bennett, A., & Checkel, J. (Eds.). (2015). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.
Bennett, A. (2022). Drawing contingent generalizations from case studies, and process tracing for program evaluation. In M. Woolcock, J. Widner, & D. Ortega-Nieto (Eds.), *The case for case studies*. Cambridge University Press.
Bennett, A., Fairfield, T., & Charman, A. (2021). Understanding Bayesianism: Fundamentals for process tracers. *Political Analysis*. https://doi.org/10.1017/pan.2021.23
Fairfield, T., & Charman, A. (2022). *Social inquiry and Bayesian inference: Rethinking qualitative research*. Cambridge University Press.

George, A., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

Humphreys, M., & Jacobs, A. (2015). Mixing methods: A Bayesian approach. *American Political Science Review, 109*(4), 653–673. https://doi.org/10.1017/S0003055415000453

Humphreys, M., & Jacobs, A. (2023). *Integrated inferences*. Cambridge University Press..

Waldner, D. (2012). Process tracing and causal mechanisms. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science* (pp. 65–84). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195392753.013.0004

Chapter 9
Exploring Interventions on Social Outcomes with In Silico, Agent-Based Experiments

Flaminio Squazzoni and Federico Bianchi

Abstract Agent-Based Modeling (ABM) is a computational method used to examine social outcomes emerging from interaction between heterogeneous agents by computer simulation. It can be used to understand the effect of initial conditions on complex outcomes by exploring fine-grained (multiple-scale, spatial/temporal) observations on the aggregate consequences of agent interaction. By performing *in silico* experimental tests on policy interventions where *ex ante* predictions of outcomes are difficult, it can also reduce costs, explore assumptions and boundary conditions, as well as overcome ethical constraints associated with the use of randomized controlled trials in behavioral policy. Here, we introduce the essential elements of ABM and present two simple examples where we assess the hypothetical impact of certain policy interventions while considering different possible reactions of individuals involved in the context. Although highly abstract, these examples suggest that ABM can be either a complement or an alternative to behavioral policy methods, especially when understanding social processes and exploring direct and indirect effects of interventions are important. Prospects and critical problems of these *in silico* policy experiments are then discussed.

Learning Objectives
By studying this chapter, you will:

- Learn the basic concepts and methodological principles of agent-based modeling.
- Understand the advantages of agent-based modeling compared to other research methods when examining social dynamics.
- Understand how to design agent-based modeling for *in silico* experiments.
- Understand the importance of agent-based modeling for policy appraisal.
- Practice with two examples of agent-based modeling for policy experiments.

F. Squazzoni (✉) · F. Bianchi
Department of Social and Political Sciences, University of Milan,
Milan, Italy, Via Conservatorio 7, 20122
e-mail: flaminio.squazzoni@unimi.it; federico.bianchi1@unimi.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_9

9.1 Introduction

Behavioral science methodology, including randomized controlled trials (RCTs), is increasingly being used in public policy as a gold standard to estimate causal relationships between interventions and outcomes (e.g., Shafir, 2012; Straßheim & Beck, 2019). Examples of behavioral policies, from public health to education, have shown the malleability of individual preferences and decisions, as well as the sensitivity of targeted individuals to cognitive frames in responding to policy interventions (Galizzi & Wiesen, 2018). The profound non-linear relationships between policy stimuli and observable and measurable people's responses, which impinge the mantra of 'big stimuli vs. big outcomes' of conventional policy (Squazzoni, 2014), has suggested that if well-conjectured and 'incentive compatible', even minimal interventions could cause large-scale outcomes (Dolan & Galizzi, 2014).

The reason why RCTs are considered the "gold standard" in behavioral policy is that random assignment of a representative, targeted population to control and treatment groups, differing only in their manipulated conditions and the identification of any controllable, salient confounding factors by *ex ante* design, are instrumental to estimate causal effects. However, besides fundamental criticism on the often neglected influence of implicit assumptions on unobservable processes in research design (e.g., Imai et al., 2008), the use of experimental methods for public policy has also important pragmatic limitations.

On the one hand, whenever feasible, RCTs for public policy purposes could have a negative benefit-cost ratio. Indeed, ethical obstacles can prevent group selection or the exploration of conditions that would introduce inequality and negative externalities for certain groups. Secondly, economic costs are often severe even for small-scale pilots. Furthermore, the intrusive, 'outside-in' nature of experimental policies can affect real-life outcomes and people's behavior in other domains beyond any intended purpose. This is indeed a fundamental problem: not only do people often react unpredictably and adaptively to interventions (note that this has been a key argument for supporters of behavioral policies against the traditional policy framework based on positive/negative incentives and 'rational' response), individuals are also embedded in social contexts so that their exposure to policy treatments can trigger positive and negative network externalities or knowledge spillovers, which might also affect outcome measurements (Dolan & Galizzi, 2015; Squazzoni, 2017). Disentangling any established causal effect between interventions and outcomes in such situations is difficult.

Finally, as suggested by Battistin & Bertoni in Chap. 3, inferences on causal effects of policy interventions would require counterfactual procedures to assess what would have happened to the estimated outcomes had these interventions not taken place. Besides the difficulty of isolating a control group in social reality and introducing a placebo-like neutral information in behavioral policies, endogenous social forces and processes cannot be suspended during a policy experiment. Treating data in a quasi-experimental way by randomization, instrumental variation and discontinuity design can increase the robustness of estimates, thus improving

the internal and external validity of causal inferences. Here, we suggest a complementary strategy: the use of agent-based modeling (ABM) as *in silico* experiments accompanying, augmenting, or even substituting RCTs—whenever needed—in the traditional toolbox of the experimentalist policy analyst.

This policy function of ABM is key especially when: (a) there are no or insufficient empirical data on which to corroborate estimated causal relationships and perform *ex post*, counterfactual assessments; (b) the economic, social, or political costs of RCTs for policy appraisal or assessment are hardly sustainable; (c) 'social experimenters' are interested not only in estimating outcomes but also understanding generative processes; (d) there is added value in exploring extreme, boundary, or counterfactual conditions that either do not exist in reality or have not yet occurred but in principle could. In all these cases, we argue that ABM is the only alternative to *ex post* observational analysis to explore and quantify hypothesized relationships between policy interventions and social outcomes. What is lost in terms of empirical realism is gained in terms of understanding the possible generative processes.

Reviews on recent applications of ABMs in various fields, from public health (Giabbanelli et al., 2021; Tracy et al., 2018) to agriculture (Kremmydas et al., 2018) and energy consumption (Klein et al., 2019), have shown that ABM is particularly suitable for providing insights into causal mechanisms, potentially linking interventions to outcomes. By generating "artificial data" via computer simulation, models can help to: (a) explore cases of multiple realizability (i.e., the same effect generated by different social causes and paths), (b) build 'what-if' scenario analysis that supports inferences about interventions-outcomes without impacting the targeted population; (c) estimate 'interference', network effects and spillovers of policy interventions (e.g., the situation in which one individual's exposure affects other individuals' outcomes); and (d) measure possibly multiple direct and indirect outcomes of the same intervention (Chalabi & Lorenc, 2013; Murray et al., 2021; Powell et al., 2017).

While most research has outlined the differences between ABM and more conventional policy approaches and methods, e.g., RCTs (e.g., Gilbert et al., 2018), here we would like to discuss complementarities and potential synergies between various experimental approaches. Indeed, as exemplified by Bravo et al. (2012), by using the computer as an 'artificial experimental environment', model parameters can be calibrated on existing individual (experimental) data to perform *in silico* counterfactual tests on any established causal relationship by quantifying the effect of varying initial conditions, especially those that could not be estimated empirically. What could happen to the observed causal relationship between *A* (intervention) and *B* (outcome), if certain hypothesized conditions *C* (either observable or not) were different? Why would *A* necessarily lead to *B* given that *C* may include adaptive, unpredictable individual behavior? As suggested by Manzo (2022), this is not only a problem of internal vs. external validity of estimated relationships (the effect of *A* on *B* would be contingent to a specific empirical instance with all due problems of generalization). It implies a search for causal or dependence relationships of interest not only within data but also via formalized models of "generative mechanisms" that consider mediating behavior and processes on which we might

not have any data. Why and how, when exposed to *A* and under interaction effects that typically occur in social contexts, would individuals behave in such a way to 'cause' the emergence of *B*?

The rest of the chapter is organized as follows: In Sect. 9.2, we provide a brief introduction to ABM, by highlighting their specificity compared to other modeling approaches. In Sect. 9.3, we present some hypothetical policy cases on which the advantages of ABM can be understood. Model code is provided to help the reader to understand the potential of ABM for: (1) exploring the effect of parameter variations on the emergence of social outcomes; (2) building alternative scenarios to understand the effect of individual reactions on social outcomes. In Sect. 9.4, we summarize the main contributions of the chapter and discuss critical points and possible developments. Indeed, besides the (many) positive aspects, ABM has also certain weaknesses, including problems of model resolution, empirical validation, and external validity, which all require careful scrutiny.

9.2 Agent-Based Modeling

Agent-based modeling is a "computational method that enables a researcher to create, analyze, and experiment with models composed of agents that interact within an environment" (Gilbert, 2008). Agents may represent individuals, households, organizations, or any other entities, whose actions depend on conditional or stochastic decision-making rules (Bianchi & Squazzoni, 2015; de Marchi & Page, 2014; Macy & Willer, 2002; Tesfatsion & Judd, 2006). Agents can adapt their behavior in response to their own experience (e.g., learning), the interaction with other agents or in response to changes in the environment—e.g., policy interventions (Gilbert & Troitzsch, 2005; Squazzoni, 2012; Tracy et al., 2018).

As dynamic and process-based, ABMs are ideal to study the effects of complex interactions between micro- and macro-levels by exploring 'generative explanations' of social outcomes (Epstein, 2006; Hedström & Bearman, 2009; Macy & Flache, 2009). This is especially important in the case of complex adaptive social systems, whose stochastic, non-linear behavior can seldom be mathematically tractable and cannot be estimated deductively without computer simulation exploring various initial conditions and possible input/output paths (Miller & Page, 2009).

Unlike statistical models, which concentrate on relations between aggregate factors (Bianchi & Squazzoni, 2020), ABM starts from representing individual behavior and ends up exploring aggregate dynamics from agent interaction via computer simulation. Social regularities and patterns are neither derived by estimating the values of stochastic parameters that would maximize a model's fitness to observed data, nor obtained by assumptions on aggregate properties that do not consider individual-level differences (e.g., Hedström & Manzo, 2015; Hedström & Udehn, 2009). ABM parameters are not estimated a posteriori, they are manipulated a priori following an experimental rather than an observational research design (Squazzoni, 2012).

Indeed, instead of being inferred from (or tested against) empirical data, the model allows us to explore hypothesized micro-social processes according to this Coleman-like connection: (a) initial macro parameter conditions → (b) heterogeneous individual behavior → (c) interaction effects → (d) social outcomes (Coleman, 1990). In line with the so-called 'analytical sociology' agenda (Hedström & Bearman, 2009; Hedström & Manzo, 2015; Manzo, 2022), ABMs can be viewed as generative models ensuring a high degree of internal validity regarding the "generative sufficient conditions" leading from (a) to (d) via the manipulation of (b) and (c) (Epstein, 2006). Unlike statistical models, generative explanations via ABM does not require the independence of observations as they aim to explore systemic, interdependent social processes, i.e., specific configurations of (a), (b), and (c) that would determine (d). Furthermore, ABM allows us to explore various patterns of agent interaction directly within explicitly represented network structures (Macy & Flache, 2009).

While traditional equation-based models condense either a 'representative', collective agent or a homogenous population into stochastic parameters (e.g., think about the modeling tradition in either standard economics or demography), ABM explicitly considers a population of heterogeneous, autonomous agents with different features and decision-making rules who interact either directly or indirectly while being exposed to various environmental stimuli, typically manipulated by the model maker (Gilbert, 2008; Macy & Flache, 2009; Macy & Willer, 2002; Squazzoni, 2012). By running experiments with human subjects, experimentalists aim to test theoretically deduced hypotheses on cause–effect relationships by manipulating the occurrence of an *explanans* (i.e., the treatment) in a randomized sample of individuals and studying the control vs. treatment group differences in the *explanandum*. In a similar fashion, an experimenter can use ABM to run several instances of a model by manipulating the *explanans*—i.e., changing the related model parameters—and then studying any differences in the simulated outcome. Instances could be designed as 'group-treatment' policy correlates, artificial agents (whose behavior could be empirically inferred from experimental data, if the ABM exercise is combined with a behavioral experiment, or theoretically postulated if data is not available) would be the correlates of experimental subjects, and their group-level reactions would be the outcome measurement. As such, the computer is used as an artificial laboratory where theoretically derived hypotheses are tested in silico by comparing a baseline (control group) initialization with manipulated scenarios (treatments) where the only difference is the introduction of a possible *explanans* (Squazzoni, 2012).

However, this does not constrain ABM to 'thought experiments' (Axelrod, 1997). Quantitative (e.g., population size, resources, network positions) and qualitative parameters (e.g., rules of behavior) related to (a), (b), and (c) can be calibrated according to empirical data (i.e., *empirical calibration*), and aggregate artificial outcomes (d) can be compared to empirical time series or distributions to adjudicate among potential configurations of (a), (b), and (c) those with higher explanatory power (i.e., *empirical validation*) (Boero & Squazzoni, 2005).

9.3 Exploring Artificial Policy Scenarios

In this section, we provide some abstract examples from our own research to illustrate the ABM approach to policy scenarios. Although there are many examples of concrete applications of ABM for policy interventions or design (e.g., Gilbert et al., 2018), here we have summarized two recent contributions that describe our idea of in silico experiments.

9.3.1 Interventions to Increase Competition or Collaboration in Science

Today, academic life is characterized by a "publish or perish" ethos and growing competition for funds and academic career (Edwards & Siddhartha, 2017; Grimes et al., 2018). While competition is expected to stimulate the quality of publications, scientists must also collaborate especially in reviewing manuscripts before publication to defend robust academic standards of knowledge. This is the important function of "peer review": vetting scientific manuscripts submitted by authors for publication to a journal by voluntary collaboration of experts guided by journal editors. Unfortunately, research has shown that lack of material incentives or a weak system of symbolic rewards can undermine peer review, as scientists would reduce time and effort in reviewing (typically voluntary and not rewarded), to maximize their efforts in new publishable research which funds, prestige, and career depend on.

Suppose that you are a policymaker wanting to test certain possible interventions to increase cooperation among scientists, but who also want to ensure that this does not compromise the quality of publication. Here are two examples of possible research policy interventions. The first represents a policymaker wanting to increase quality signals of publication so to induce scientists to compete for excellence, e.g., promoting only those scientists who publish in top journals. The second wants to reward peer reviewing by introducing an open science policy that would induce journals to shift from confidential to open peer review so that the identity of any reviewer is public, regardless of the final decisions on manuscripts. This would permit reviewers to claim their review as a reward. Note that even if abstract, both policy interventions are 'realistic': scientists are increasingly exposed to competitive rewards under the dominant rhetoric of excellence and comprehensive evaluation in almost all institutional contexts (e.g., Forsberg et al., 2022). In the second case, scientific associations and certain publishers have started to introduce open peer review policies as a means to recognize and reward reviewers (Bravo et al., 2019). Therefore, these examples are abstract (i.e., there is no 'real policy maker' commissioning a computational test of such policies) but not completely unrealistic (i.e., these interventions have been explored more locally and by trial and error).

Suppose we prepare a model to test these possible interventions. Assume a population of n agents representing a community of scientists. Assume that scientists are

hired by academic organizations that periodically provide them with some minimal funding R_i (e.g., laboratory equipment, access to online resources, etc.), allocated from a fixed overall amount of resources, $R = \sum_i R_i$. Assume that scientists are required to publish manuscripts to get more funds, reputation, prestige, and career, but that journals are competitive and so accept only a fixed proportion (P) of submitted manuscripts depending on a quality ranking determined by reviewers. Scientists then update their resource share according to their publication record as follows:

$$R_i = \frac{p_i}{\sum_i p_i} R$$

Suppose that, at each time step (t), scientists are required to perform two tasks, i.e., submitting their manuscripts to journals and reviewing manuscripts submitted by others (for the sake of simplicity, let us assume that each manuscript is submitted by only one author and is reviewed by only one reviewer; for a similar model, where we varied the number of reviewers, see Bianchi & Squazzoni, 2016). Assume that time is a scarce resource and both tasks are costly in that scientists need to decide how to allocate their resources between these two tasks.

Assume that the quality of submitted manuscripts (Q_i^s) and review reports (Q_i^r) linearly depends on the amount of resources allocated by scientists to these two tasks, as in:

$$Q_i^s = e_i R_i$$

$$Q_i^r = R_i - Q_i^s = \left(1 - e_i\right) R_i,$$

where e_i determines how resources are allocated between submitting and reviewing.

Following Squazzoni and Gandelli (2012, 2013), we assume that reviews may be biased, so the actual quality of manuscripts could be only approximated by the reviewer depending on the level of resources individually invested by the scientist in reviewing (higher investment = more precise evaluation of the quality of manuscripts), as follows:

$$\hat{Q}_i^s = \alpha_j Q_i^s,$$

with α_j being drawn from a normal distribution $N(\mu = 1, \sigma = \min\left(T^*, Q_j^r\right)$, where j is the reviewer and T^* is a quality threshold which estimates the minimum amount of resources needed by each j to provide a fair review.

Suppose that the quality of manuscripts can be unequivocally quantified so that manuscripts can be compared and ranked by journals for publication. Suppose we do not consider the role of editors, the presence of multiple journals, the possibility of resubmitting rejected manuscripts and other 'realistic' conditions. Let us

Table 9.1 Pseudo-code of the model (for more detail, see Bianchi et al., 2018)

Input: time t, number of iterations m, set of n agents, R, e, τ, Δe, p **Output:** publication bias, average publication quality, top quality	
1	initialize $t = 0$
2	**while** $t < m$ **do**
3	**for all** agents i:
4	update R_i
5	$e_i \leftarrow e_i + \Delta e$
6	compute Q_i^s
7	**end for**
8	**while** # of reviewers < $n/2$ **do**:
9	select random agent i
10	assign i "reviewer" role
11	match i to random j with no "reviewer" role
12	compute Q_i^r
13	compute $\hat{Q}_i^s$
14	**end while**
15	rank agents by $\hat{Q}_i^s$
16	**for all** top Pn agents in $\hat{Q}_i^s$ ranking do:
17	published? $\leftarrow$ *true*
18	**end for**
19	**end while**

consider these factors as irrelevant here (see the pseudo-algorithm describing the model in Table 9.1).

Let us next run our simulations for a sufficient number of iterations (in our cases, $m = 1500$) to reach a stable outcome equilibrium (in our case, we repeated our simulations at least 100 times for each initialization) and measure the outcomes as follows: (1) publication bias (i.e., the proportion of incorrectly rejected submissions on the total amount of published articles); (2) the average quality of publications; (3) average quality of the ten top-quality articles. All measurements are in time steps and so can be averaged at the end of each simulation (see the model parameter in Table 9.2).

9.3.1.1 Example 1

Let us now suppose that we want to explore a set of potential interventions to stimulate scientists to increase their quality of publication ((2)) while at the same time, minimizing publication bias at the system level ((1)). For instance, the policymaker could set up rewards or prizes to this purpose but would like to estimate the

Table 9.2 Example 1: Model parameters

Parameter	Description	Value
n	Number of scientists	500
e	Resources allocated to manuscript production	Range: 0–100% (uniform distribution)
R	Distribution of initial resources	Uniform
T^*	Minimum-quality threshold (the expected amount of resources required by each scientist to perform a review)	6
Δe	Variation of resources allocated to manuscript production	5%
P	Proportion of manuscripts accepted by journals for publication in each time step	25%

Adapted from Bianchi et al. (2018)

mediating effect of scientists' possible reactions. You could create two 'treatment scenarios': one in which rewards point to strong competition and excellence, e.g., scientists are induced to compare their Q_i^s (regardless of whether their submission was published or rejected) in the top ten publications (we called it "high competition"), another one in which rewards point to the average quality (we called it "minimum expected quality"), e.g., scientists use the average quality of below-median published articles as a comparison. In both scenarios, suppose that these comparisons would determine an individual binary satisfaction value, which would make scientists revise their resource allocation decisions between investing more either in their own manuscripts or for reviewing other manuscripts.

Now, let us hypothesize three possible decisions made by scientists: (1) always selfishly investing in their own publication against peer reviewing, (2) investing more in reviewing when their manuscripts have been previously rejected, and (3) investing more in reviewing when their manuscripts have been previously published. Let us then add a control factor: a level of subjective overconfidence when scientists compare the quality of their own manuscripts with current publications by others. This can be done by re-running all the same simulation scenarios while differing for two further conditions: all scenarios initialized with 'objective' comparison vs. all scenarios with 'subjective' quality comparisons. This factorial design would imply measuring the same outcomes. Then, let us suppose that you create an artificial 'control group' where you remove any comparison where scientists would follow their allocation strategies without any intervention regarding 'excellent' or 'minimum expected quality' signals.

We calculated cumulative moving average values of our outcomes on the last 100 steps of each iteration and the mean value of outcome measurements for each scenario. Table 9.3 shows the first outcome ((1)), i.e., publication bias, when scientists were induced to compete for excellent or looked at minimum expected quality adapt their allocation strategies accordingly. Confront the outcomes with the control group. Adding rewards for excellence determined high publication bias than 'minimum expected quality' signals. However, outcomes vary greatly depending on the scientists' adaptive reactions. Note that reviewing only after being published, e.g., a reciprocal behavior, without considering any comparison of quality was detrimental

Table 9.3 Evaluation bias (%) in different scenarios. (Mobile mean values over 100 repetitions)

	Rewards				
	Control ↓	*Minimum expected quality*		*High competition*	
Scientist behavior	Comparison bias →	Objective	Overconfidence	Objective	Overconfidence
Investing only in publication	57.61				
Reviewing after rejection	32.71	40.56	29.47	62.79	58.01
Reviewing after being published	66.91	27.86	28.05	30.66	27.04

Adapted from Bianchi et al. (2018)

Table 9.4 Average published quality in different scenarios. (Mobile mean values over 100 repetitions, then normalized 0–1)

	Rewards				
	Control ↓	*Minimum expected quality*		*High competition*	
Scientist behavior	Comparison bias →	Objective	Overconfidence	Objective	Overconfidence
Investing only in publication	0.60				
Reviewing after rejection	0.98	0.71	0.85	0.44	0.49
Reviewing after being published	0.41	0.00	0.01	1.00	0.36

Adapted from Bianchi et al. (2018)

to the publication bias. Furthermore, counterintuitively, overconfidence had a positive effect in both scenarios, especially in the high competition scenario (29.47%), where publication bias decreased even below the outcome of the 'control group' scenario (32.71%). Therefore, results suggest that publication bias was higher under stronger competition but precise effects depended on various behavioral factors.

If we were to consider the second outcome of interest, however, ((2), i.e., the average quality of publications), results did not vary similarly to the first outcome, i.e., publication bias. The highest value was achieved when scientists were induced to compete for excellence and reciprocated higher investment in reviewing whenever previously published (see Table 9.4). This was confirmed when considering the quality of the top ten published articles across different scenarios (see Table 9.5). In conclusion: (a) policy interventions that increase competitive spirits of scientists towards publications could backfire if norms of peer reviewing cannot be enforced; (3) even a minimal level of overconfidence can determine positive or negative outcomes compared to more objective self-evaluation (for detail, see Bianchi et al., 2018).

Table 9.5 Average publication quality of top ten published papers across different institutional settings and behavioral strategies. (Mobile mean values over 100 repetitions, then normalized 0–1)

	Rewards				
	Control ↓	*Minimum expected quality*		*High competition*	
Scientist behavior	Comparison bias →	Objective	Overconfidence	Objective	Overconfidence
Investing only in publication	0.51				
Reviewing after rejection	0.91	0.94	1.00	0.75	0.83
Reviewing after being published	0.36	0.01	0.00	0.93	0.34

Adapted from Bianchi et al. (2018)

9.3.1.2 Example 2

Now let us suppose that we would like to manipulate the peer-review policy adopted by journals testing the effect of shifting from confidential to open peer review in situations in which scientists would be sensitive to competition and status when reviewing others' manuscripts. Under confidential peer review, authors and reviewers do not know each other's identity and so they could just react to their own rejections by reducing their effort e_i in reviewing to punish the system which did not favor them. Under open-peer review, author and reviewer identities are disclosed and so scientists could reciprocate positive or negative editorial decisions by adapting $\hat{Q}_i^s$ once they are later matched by the journal. Note that the sensitivity of scientists to this shift of the peer review model has been found in some recent 'quasi-experimental' analysis (e.g., Bravo et al., 2019). Do the positive benefits of open peer review come at the price of increasing publication bias, if scientists can react to status and competition and use peer review to either help favorable or punish unfavorable authors who previously reviewed their own manuscripts? Can we ideally quantify how much that price would be?

Table 9.6 shows the initial parameters of this model. We tested various possible behaviors with a focus on reviewing (e.g., always being fair, being randomly reliable, deciding how much to invest in reviewing depending on previous rejection or acceptance of their manuscript). Here, we concentrated on comparing different reviewers' reactions to previous experience as authors in two journal settings: (1) journals following confidential peer review, in which reviewers invest in reviewing whenever previously published or otherwise disinvest, so providing unreliable reports; (2) journals following open peer review, in which reviewers and authors' identities are revealed and reviewers reciprocate positive reviews to authors who previously favored them when reviewers, and negative reviews to previously unfavorable reviewers.

Figure 9.1 shows the first outcome of interest ((1)), i.e., publication bias, when journals follow confidential peer review and reviewers are either always fair, always unreliable, or sensitive to previous experiences as authors (e.g., being fair when

Table 9.6 Example 2: Model parameters

Parameter	Value
Number of scientists	240
Scientists' initial resources	0
Fixed resource gain (initial endowment of resources for each scientist in each time step)	1
Author bias factor (noise coefficient in the conversion of scientists' resources into quality of manuscript)	0.1
Velocity of best quality approximation (fixed rate at which the quality of submitted manuscripts varies according to the increase of the author's resources)	0.1
Discount factor on resources for unreliable reviews (discount rate on resources when scientists perform unreliable reviews)	0.5
Proportion of accepted manuscripts at the end of each time step	25%

Adapted from Bianchi and Squazzoni (2022)

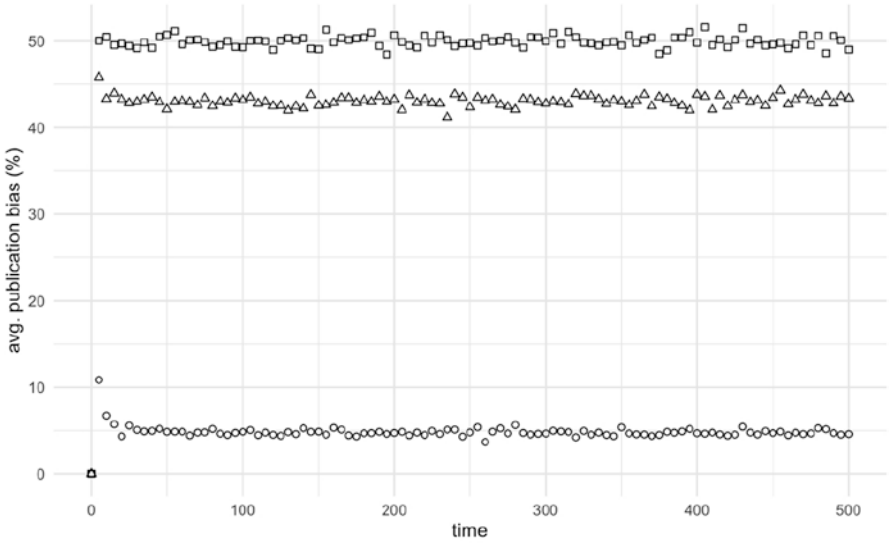

Fig. 9.1 The impact of reviewer behavior on publication bias in confidential peer review. Circles: fair; squares: unfair; triangles: reactive. Values averaged over 200 realizations. (Source: Bianchi & Squazzoni, 2022)

previously treated fairly, being unfair when previously being treated unfairly). If reviewers react to previous experience, the level of bias approximates a random situation in which the publication of manuscripts could be decided by editors tossing a coin. Let us use these outcomes as a baseline to compare the effect of reciprocity strategies in the two peer review settings.

Figure 9.2 shows the first outcome of interest ((1)), i.e., publication bias, when comparing reciprocal strategies in the two peer review settings. Publication bias increased more than 20% under open peer review and added an extra 20% of bias compared to a situation where editorial decisions would be random. This would

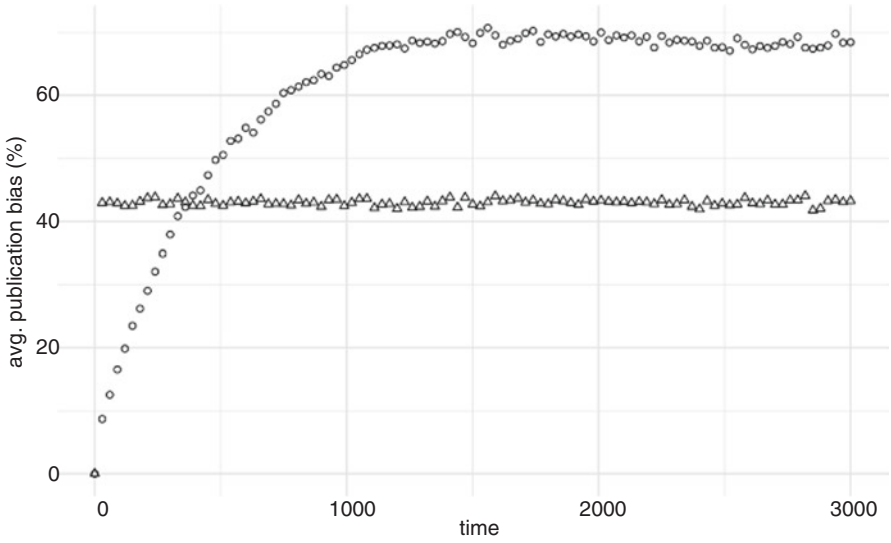

Fig. 9.2 The impact of scientists' reciprocity strategy on publication bias in confidential vs. open peer review. Triangles: indirect reciprocity (confidential peer review); circles: direct reciprocity (open peer review). Values averaged over 200 realizations. (Source: Bianchi & Squazzoni, 2022)

suggest that open peer review could be detrimental whenever we assume that reviewers are sensitive to cooperation signals. Further results (reported in Bianchi & Squazzoni, 2022) indicate that even if reviewers would retaliate only against previous reviewers of lower academic status (i.e., with lower resources compared to theirs) while being fair in case previous unfavorable reviewers were scientists of higher status, the effect on the outcome would differ only minimally (differences not higher than 5% on the level of publication bias).

Figure 9.3 shows the effect of reviewer behavior on the second outcome (((2)), i.e., the average quality of publications. Open peer review would determine the lowest quality of publications even when compared to random editorial decisions. Note that we tested the sensitivity of these outcomes to the variation of all initial parameters and findings were confirmed (see the Supplementary Material of Bianchi & Squazzoni, 2022). In conclusion, this exercise would suggest that if practices and norms exist that make scientists frame peer review as a signaling game, open peer review polices, once adopted globally, could increase publication bias by more than 20% compared to confidential peer review, thus compromising publication quality. Obviously, other computational tests could also be designed with the model by considering for example other factors, being more nuanced, and considering empirically grounded behavior. Although a more realistic and empirically calibrated parameterization of the model would be important, as suggested by Feliciani et al. (2019) in their overview of computer simulation research on peer review, these cases here were only aimed to exemplify a method to test policy interventions artificially.

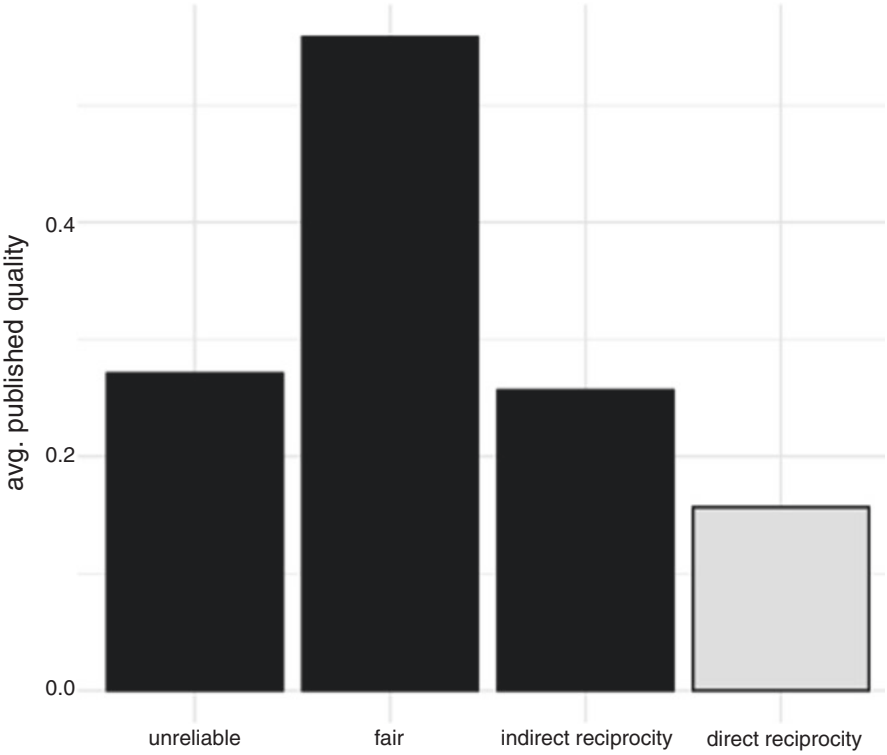

Fig. 9.3 The impact of reviewer behavior on the average quality of published papers under different peer review models. In the rectangle: comparison between reciprocity strategy in confidential (black) vs. open peer review (white). Values averaged over 200 realizations. (Source: Bianchi & Squazzoni, 2022)

9.4 Conclusions

In this chapter, we have presented ABM as a method to perform computational experimental tests on non-linear, complex effects of policy interventions as these can determine interaction effects and individual adaptations. This could enlarge the toolbox of experimental policy analysists, especially when RCTs cannot be designed due to various ethical, political, or economic constraints. *In silico* tests are also required before policy design to explore potential unintended consequences or when an understanding of social processes could provide relevant insights to enhance comprehensive policy appraisal. In our view, ABM can fruitfully complement, enrich, and even substitute—when necessary—more conventional behavioral methods for public policy.

However, the use of ABM also has important limitations. As discussed by Gilbert et al. (2018) in a comprehensive review of practices of computational modeling of public policy, deciding the appropriate model resolution requires critical decisions.

Besides the hypothetical exercises presented here, where we have proposed abstract examples, in concrete contexts, the optimal level of abstraction of a model depends on the purpose of modeling and the nature of the system being modeled (Edmonds et al., 2019). For instance, during the COVID-19 pandemic, epidemiologists have used ABMs to simulate a variety of anti-contagion policies to flatten the curve by reaching an appropriate level of resolution on certain parameters (e.g., population size). However, they followed empirically implausible assumptions on relevant others (e.g., social networks and externalities), which compromised a more comprehensive exploration of possible policy interventions while downplaying the fundamental role of uncertainty (see Squazzoni et al., 2020 for a critical overview; for an example of empirical calibration of networks in epidemiological models, see Manzo & van der Rijt, 2020).

This raises two interrelated challenges in the use of ABM for public policy, i.e., the use of empirical data to calibrate model parameters via existing or *ad hoc* data, and the heuristic value of model findings to inform policy interventions or policy evaluation (Tracy et al., 2018). In this regard, as suggested by Murray, Marshall & Buchanan (2021, 1655) in their proposed 'target trial framework', whenever combined with the usual experimental framework of behavioral policy, ABM could incorporate empirical data on the targeted population (e.g., calibrating salient characteristics of individuals from available data sources) and a detailed and explicit specification of the hypothetical trial, while using the in silico experimental nature of these models as an 'artificial world' "with no ethical, logistical, or financial constraints, and in which the exposure of interest is perfectly manipulable by study investigators, regardless of whether this is actually feasible or ethical in the real world." This would help to fill the gap between empirical data and unobservable variables and inform study design. Furthermore, following Bravo et al. (2012), calibrating ABM with results from small-scale pilots, RCTs or well-detailed observational studies or re-running existing trials in a model, while scaling the characteristics of the original target population to populations with other characteristics or testing other network structures compared to those originally reproduced in the previous study, could help us to increase generalization or perform counterfactual tests of policy findings. This would help to assess the dependence of outcomes from contextual details and help us understand how much causal inference exercises on complex social behavior require careful examination.

Suggested Readings

Epstein, J. M. (2006). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press.

Manzo, G. (2022). *Agent-Based Models and Causal Inference*. Hoboken, NJ: Wiley & Sons.

Page, S. E. (2018). *The Model Thinker*. New York, NY: Basic Books.

Review Questions

1. What are the limitations of RCTs for public policy?
2. What is agent-based modeling?
3. Which are the benefits of using ABM to examine social processes?

4. Can ABM be informed by empirical data?
5. What are the limitations of ABM as a method to inform policy interventions?

Replication Material
The models have been built in NetLogo. The code is available at the following links:

https://www.comses.net/codebases/6b77a08b-7e60-4f47-9ebb-6a8a2e87f486/releases/1.0.0/ (Example 1).
https://www.comses.net/codebases/3d99eb9f-ae4f-42d0-8c58-9d28757161c0/releases/1.0.0/ (Example 2).

References

Axelrod, R. (1997). *The complexity of cooperation. Agent-based model of competition and collaboration*. Princeton University Press.

Battistin, E., & Bertoni, M. (this volume). Counterfactuals with experimental and quasi-experimental variation. In A. Damonte & F. Negri (Eds.), *Causality in policy studies – A pluralist toolbox*. Springer.

Bianchi, F., Grimaldo, F., Bravo, G., & Squazzoni, F. (2018). The peer review game: An agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics, 116*(3), 1401–1420.

Bianchi, F., & Squazzoni, F. (2015). Agent-based models in sociology. *Wiley Interdisciplinary Reviews: Computational Statistics, 7*(4), 284–306.

Bianchi, F., & Squazzoni, F. (2016). Is three better than one? Simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.), *Proceedings of the 2015 winter simulation conference* (pp. 4081–4089). IEEE Press.

Bianchi, F., & Squazzoni, F. (2020). Modelling and social science. Problems and promises. In E. A. Moallemi & F. J. de Haan (Eds.), *Modelling transitions. Virtues, vices, visions of the future* (pp. 60–74). Routledge.

Bianchi, F., & Squazzoni, F. (2022). Can transparency undermine peer review? A simulation model of scientist behavior under open peer review. *Science and Public Policy*, scac027.

Boero, R., & Squazzoni, F. (2005). Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation, 8*(4), 6.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications, 10*, 322.

Bravo, G., Squazzoni, F., & Boero, R. (2012). Trust and partner selection in social networks: An experimentally-grounded model. *Social Networks, 34*(4), 481–492.

Chalabi, Z., & Lorenc, T. (2013). Using agent-based models to inform evaluation of complex interventions: Examples from the built environment. *Preventive Medicine, 57*(5), 434–435.

Coleman, J. S. (1990). *Foundations of social theory*. Belknap.

Dolan, P., & Galizzi, M. M. (2014). Getting policy-makers to listen to field experiments. *Oxford Review of Economic Policy, 30*(4), 725–752.

Dolan, P., & Galizzi, M. M. (2015). Like ripples on a pond: Behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology, 47*(4), 1–16.

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., & Squazzoni, F. (2019). Different modelling purposes. *Journal of Artificial Societies and Social Simulation, 22*(3), 6.

Edwards, M. A., & Siddhartha, R. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science, 34*(1), 51–61.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.

Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., Marušić, A., & Shankar, K. (2019). A scoping review of simulation models of peer review. *Scientometrics, 121*(1), 555–594.

Forsberg, E., Geschwind, L., Levander, S., & Wermke, W. (Eds.). (2022). *Peer review in an era of evaluation: Understanding the practice of gatekeeping in academia*. Edward Elgar.

Galizzi, M., & Wiesen, D. (2018). Behavioral experiments in health economics. In J. H. Hamilton, A. Dixit, S. Edwards, & K. Judd (Eds.), *Oxford research encyclopedia of economics and finance*. Oxford University Press.

Giabbanelli, P. J., Tison, B., & Keith, J. (2021). The application of modeling and simulation to public health: Assessing the quality of agent-based models for obesity. *Simulation Modelling Practice and Theory, 108*, 102268.

Gilbert, N. (2008). *Agent-based models*. Sage.

Gilbert, N., Ahrweiler, P., Barbrook-Johson, P., Narasimhan, P., & Wilkinson, H. (2018). Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation, 21*(1), 14.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). McGraw-Hill.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science, 5*(1), 171511.

Hedström, P., & Bearman, P. (2009). What is analytical sociology all about? An introductory essay. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 3–24). Oxford University Press.

Hedström, P., & Manzo, G. (2015). Recent trends in agent-based computational research: A brief introduction. *Sociological Methods & Research, 44*(2), 179–185.

Hedström, P., & Udehn, L. (2009). Analytical sociology and theories of the middle range. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology*. Oxford University Press.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society A, 171*(2), 481–502.

Klein, M., Frey, U. J., & Reeg, M. (2019). Models within models – Agent-based modelling and simulation in energy systems analysis. *Journal of Artificial Societies and Social Simulation, 22*(4), 6.

Kremmydas, E., Athanasiadis, I. N., & Rozakis, S. (2018). A review of agent-based modelling for agricultural policy evaluation. *Agricultural Systems, 164*, 95–106.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology, 28*, 143–166.

Macy, M. W., & Flache, A. (2009). Social dynamics from the bottom up: Agent-based models of social interaction. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 245–268). Oxford University Press.

Manzo, G. (2022). *Agent-based models and causal inference*. Wiley.

Manzo, G., & van der Rijt, A. (2020). Halting SARS-CoV-2 by targeting high-contacts individuals. *Journal of Artificial Societies and Social Simulation, 23*(4), 10.

de Marchi, S., & Page, S. E. (2014). Agent-based models. *Annual Review of Political Science, 17*, 1–20.

Miller, J. H., & Page, S. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton University Press.

Murray, E. J., Marshall, B. D. L., & Buchanan, A. L. (2021). Emulating target trials to improve causal inference from agent-based models. *American Journal of Epidemiology, 190*(8), 1652–1658.

Powell, K. E., Kibbe, D. L., Ferencik, R., Soderquist, C., Phillips, M. A., Vall, E. A., & Minyard, K. J. (2017). System thinking and simulation modelling to inform childhood obesity policy and practice. *Public Health Reports, 132*, 33–38.

Shafir, E. (Ed.). (2012). *The behavioral foundations of public policy*. Princeton University Press.

Squazzoni, F. (2012). *Agent-based computational sociology*. Wiley.

Squazzoni, F. (2014). A social-science inspired complexity policy: Beyond the mantra of incentivization. *Complexity, 19*(6), 5–13.

Squazzoni, F. (2017). Towards a complexity-friendly policy: Breaking the vicious circle of equilibrium thinking in economics and public policy. In J. Johnson, A. Nowak, P. Ormerod, B. Rosewell, & Y.-C. Zhang (Eds.), *Non-equilibrium social science and policy* (pp. 135–148). Springer.

Squazzoni, F., & Gandelli, C. (2012). Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics, 6*(2), 265–275.

Squazzoni, F., & Gandelli, C. (2013). Opening the black-box of peer review: An agent-based model of scientist behaviour. *Journal of Artificial Societies and Social Simulation, 16*(2), 3.

Squazzoni, F., Pohill, G. J., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, E., Borit, M., Verhagen, H., Giardini, F., & Gilbert, N. (2020). Computational models that matter during a global pandemic outbreak: A call to action. *Journal of Artificial Societies and Social Simulation, 23*(2), 10.

Straßheim, H., & Beck, S. (Eds.). (2019). *Handbook of Behavioural change and public policy*. Edward Elgar.

Tesfatsion, L., & Judd, K. L. (Eds.). (2006). *Handbook of computational economics. Volume 2: Agent-based computational economics*. North-Holland.

Tracy, M., Cerdá, M., & Keyes, K. M. (2018). Agent-based modelling in public health: Current applications and future directions. *Annual Review of Public Health, 39*, 77–94.

Chapter 10
The Many Threats from Mechanistic Heterogeneity That Can Spoil Multimethod Research

Markus B. Siewert and Derek Beach

Abstract The combination of cross-case and within-case analysis in Multi-Method Research (MMR) designs has gained considerable traction in the social sciences over the last decade. One reason for the popularity of MMR is grounded in the idea that different methods can complement each other, in the sense that the strengths of one method can compensate for the blind spots and weaknesses of another and vice versa. In this chapter, we critically address this core premise of MMR with an emphasis on the external validity of applying some cross-case method, like standard regression or Qualitative Comparative Analysis, in combination with case study analysis. After a brief overview of the rationale of MMR, we discuss in detail the problem of deriving generalizable claims about mechanisms in research contexts that likely exhibit mechanistic heterogeneity. In doing so, we clarify what we mean by mechanistic heterogeneity and where researchers should look for potential sources of mechanistic heterogeneity. Finally, we propose a strategy for progressively updating our confidence in the external validity of claims about causal mechanisms through the strategic selection of cases for within-case analysis based on the diversity of the population.

Learning Objectives
By studying this chapter, you should be able to:

- Understand the main rationale behind Multi-Method Research in the social sciences.
- Be aware of different ontological and epistemological assumptions and their consequences for conducting multimethod research.
- Grasp the concept of mechanistic heterogeneity analytically.

M. B. Siewert
Munich School of Public Policy, Technical University of Munich, Munich, Germany
e-mail: markus.siewert@hfp.tum.de

D. Beach (✉)
Aarhus University, Aarhus, Denmark
e-mail: derek@ps.au.dk

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_10

- Critically discuss different sources of causal heterogeneity at the level of mechanisms, and their repercussions for causal inference in multimethod research.
- More consciously generate generalization strategies for their own research projects, and critically examine the external validity of existing multimethod research.

10.1 Introduction

Over the last decades, multimethod research (MMR) has gained considerable popularity in the analysis of public policy (see Fielding, 2010; Hendren et al., 2018; Wolf, 2010 for overviews about MMR studies in public policy), echoing a general trend in the political and social sciences (seminally, Lieberman, 2005; for up-to-date discussions, see Beach and Kaas 2020; Goertz, 2017; Humphreys & Jacobs, 2015; Seawright, 2016). Many texts define MMR as any research design which uses two or more methods to analyze the same research topic, often involving cross-case analysis of patterns of association between causes and outcomes and within-case analysis of how the causal linkage(s) work (see Creswell & Plano Clark, 2018; Schoonenboom & Burke Johnson, 2017; Tashakkori & Teddlie, 2021 for various definitions).

The most common type of MMR in political science involves the combination of some form of cross-case analysis, e.g., using regression-based methods (see Chaps. 4 and 5), or some variant of mediation analysis (see Chap. 6) or Qualitative Comparative Analysis (see Chap. 7), and one or several within-case studies using methods like congruence analysis or process tracing (see Chap. 8).[1] The cross-case analysis enables the identification of the net causal effects or invariant association between X and Y, i.e., does X make a difference for Y? The within-case analysis, on the other hand, focuses on the causal linkage *aka* mechanism(s), i.e., how does X work to bring about Y? The core logic behind this variant of MMR, in a nutshell, is that combining methods that allow for different kinds of inferences bears the potential to use the particular strengths of one technique to cancel out the other's weaknesses, and vice versa (e.g., Beach, 2020, 163; Clarke et al., 2014, 341; Goertz, 2017, 5–6; Lieberman, 2005, 436; Weller & Barnes, 2016, 426–27). In doing so, the promise of MMR is that its design ultimately yields more robust inferences by shedding light on social phenomena or substantiating our understanding of policy problems from different analytical perspectives.

The question of whether MMR can deliver on this promise – whether different methods can efficiently complement each and strengthen overall causal inferences

[1] In this chapter, we deliberately leave aside the question of MMR using interpretative techniques. Irrespective of their many merits, interpretative techniques concentrate on research themes that fundamentally differ from the type of causal questions addressed in this chapter. Hence, we remain within the broad ontological assumption that causation exists in the form of causal effects/invariant associations and causal mechanisms and that they can be examined empirically (see Chap. 2) – a thread that connects all contributions in this volume. For recent developments in interpretative methods, see Schwartz-Shea and Yanow (2012).

"because taken on their own each sort of evidence has significant limitations" (Clarke et al., 2014, 341) – has not gone uncontested. In fact, there is a notable strand within the methodological literature reflecting upon the notion of mutual complementarity in MMR. The core of this debate deals with whether different methods that make different types of causal claims and use different types of evidence can really be merged as seamlessly as is frequently portrayed (Beach and Kaas, 2020). Among other things, it has been highlighted that MMR can involve the problem of conceptual stretching or might even introduce conceptual incongruity if specific causal properties are added/dropped from concepts when moving between the cross-case and the within-case level of an analysis (Ahmed & Sil, 2009; Ahram, 2013). Similarly, while case studies can be used to check for measurement errors or to develop context-sensitive indicators (e.g., Seawright, 2016, 50–53), it can be that translating within-case observations into comparable cross-case data, and the other way around, is neither intuitive nor straightforward (Ahram, 2013; Kuehn & Rohlfing, 2009). Finally, it has been frequently mentioned that case studies can be used in MMR to check for under- and/or overspecification of the explanatory model at the cross-case level (Lieberman, 2005; Seawright, 2016: 67–74). Yet, Rohlfing (2008) convincingly shows that model misspecifications can travel between different levels of analysis because residuals and effect sizes might point towards the wrong cases for further within-case study, hence aggravating the situation, since an incorrect model is corroborated by looking at the wrong cases. In short, numerous pitfalls can complicate the effective integration of different approaches and methods in MMR designs.

This chapter concentrates on another significant problem: How can insights about causal mechanisms gained by studying how they work in one case be generalized to cases that we have not studied using case studies but look similar at the cross-case level? The issue of generalization has so far largely been ignored in the political science literature on MMR. As we will show below, generalizing about mechanisms is particularly difficult in settings that exhibit mechanistic heterogeneity. We define *mechanistic heterogeneity* as a scenario where multiple different mechanisms link the same explanatory factor(s) X to the same outcome Y (Álamos-Concha et al., 2021; Beach et al., 2019). For instance, we might find out that epistemic authorities (*aka* experts) gained influence over a policy in one case through a mechanism involving a process where the experts gained access to decision-makers by joining the bureaucracy itself (Löblová, 2018). However, in another case, influence might have been achieved through other processes, such as experts or lobbies' framing of the debates from the outside.

This form of heterogeneity and complexity at the level of mechanisms is widely discussed in the literature on case-based methodology (Beach & Pedersen, 2016, 2019; Bennett & Checkel, 2015; Blatter & Haverland, 2012; Falleti & Lynch, 2009; George & Bennett, 2005; Rohlfing, 2012). However, it is largely neglected in most accounts that deal with the integration of cross-case and within-case analysis (but see Beach et al., 2019; Goertz, 2017; Weller & Barnes, 2016), which is why we do not yet have a good understanding of how to deal with the issue of making cross-case and within-case analysis communicate in MMR. To put it simply, the

cross-case analysis tells us about differences and similarities at the level of X's and Y's; in contrast, the within-case analysis tells us about linkages (if any) between X and Y. In fact, we are making different types of causal claims, using very different types of empirical material (Clarke et al., 2014).

Addressing this question in the context of a volume on causation in policy studies is important for several reasons. First, we can observe an apparent 'mechanistic turn' in the social sciences which gradually expands across its subfields, including the field of public policy analysis (e.g., Capano et al., 2019; Capano & Howlett, 2021; Fontaine, 2020; Kay & Baker, 2015; Lindquist & Wellstead, 2019; van der Heijden et al., 2019). For instance, Fontaine (2020, 274) stresses that there is an emerging consensus on the fact that producing evidence about mechanisms via process tracing bears a significant "potential contribution to comparative policy analysis." Capano and Howlett (2021, 142 italics in the original) go one step further, arguing that "[p]olicy-makers [..] need a realistic causal theory about what occurs when policy tools are deployed and how it occurs if they want to design something that will actually happen more often than not, and to escape the trap of poorly conceived and related tacit knowledge, experience, and heuristics." Yet, secondly, if we accept that producing comprehensive causal explanations requires both robust evidence that a probable cause X is correlated/associated with Y as well as sound evidence for the causal mechanisms linking X and Y, the ability to generalize mechanistic claims from one studied case to other cases belonging to the same population becomes a significant issue. In one case study, we might have found that the linkage worked in one way, but how would we know whether the linkage (if any) is similar in other cases if we have not also investigated them? For instance, can we assume that a particular strategy used by a political entrepreneur that worked during a crisis would work in other situations? Assuming that mechanisms work in similar ways in other, non-studied cases is in effect generalizing based on hope instead of evidence. If researchers and policymakers need to know what works, how, and under what conditions, a well-informed mapping of the underlying mechanisms operative within a population of cases is crucial to generalize how X and Y are linked in different cases within a population.

The chapter is structured as follows: Section 10.2 outlines the basic ideas behind MMR designs, introduces the main templates, and discusses key ontological and epistemological differences when combining cross-case and within-case analysis. Section 10.3 addresses the problem of mechanistic heterogeneity by illustrating what heterogeneity at the level of mechanisms means. After that, Sect. 10.4 presents a selected set of potential sources to which researchers should turn to check for mechanistic heterogeneity in MMR. In Sect. 10.5, we discuss a stepwise generalization strategy that is sensitive to mechanistic heterogeneity and whose primary goal is to progressively update the confidence in the external validity of mechanisms by gradually expanding the knowledge about how mechanisms work in different (sets of) cases. The chapter closes with some final remarks.

10.2 Basic Ideas Behind MMR

The main rationale behind combining cross-case and within-case methods in MMR is that it allows researchers to make different types of causal inferences (e.g., Beach, 2020; Beach & Rohlfing, 2018; Goertz, 2017; Lieberman, 2005; Rohlfing & Schneider, 2018; Seawright, 2016; Weller & Barnes, 2016). On the one hand, cross-case analyses are particularly good at identifying cause–effect relationships by examining regular associations in the form of controlled experiments, correlations, or set-relations across a sample of cases. On the other hand, within-case analyses can establish the causal linkages between one or several causes and the respective contributions by tracing the underlying causal mechanism(s). By integrating both analytic perspectives and using methods in combination to address a shared research theme, it is argued that one can strengthen the soundness and robustness of the inferences since each mode of analysis has particular strengths that can make up for the other's blind spots (Cartwright, 2011; Clarke et al., 2014; Steel, 2008).

But how does this division of labor work in research practice? The literature on MMR has produced numerous taxonomies and typologies of different designs (see Bryman, 2006; Creswell & Plano Clark, 2018; Schoonenboom & Burke Johnson, 2017; Tashakkori & Teddlie, 2021, among others). One common defining element is whether the methods are applied in parallel or sequentially. In *parallel designs*, two or more methods are applied simultaneously; in *sequential designs*, one is used after the other. A different feature is whether the parts of an MMR study depend on each other or are performed independently. In the former scenario, insights from one study inform the data collection and/or analysis of the other; in the latter scenario, data collection and/or analysis are performed separately within each method.

The sequential research strategy is probably the most common in political science research. Two variants are typically distinguished (e.g., Beach & Rohlfing, 2018, 11–18; Lieberman, 2005; Rohlfing, 2008; Rohlfing & Schneider, 2018, 44–45; Seawright, 2016). In 'cross-case first/within-case second' designs, the researcher starts with some form of cross-case analysis to identify robust connections between a (set of) explanatory factor(s) X and an outcome of interest Y. This is followed by one or several case studies based on the findings of the first analytic step. On the other hand, 'within-case first/cross-case second' designs follow the opposite logic. Here, the analysis starts at the within-case level to uncover some causal connection and/or mechanisms and then continues with the cross-case analysis to explore whether the identified relationship also holds across a population of cases.

While one of the original motivations behind the methodological work on MMR was to (at least partially) overcome the divide between qualitative and quantitative methods, recent debates have again emphasized the ontological and epistemological differences between research approaches and the challenges they create for integrating methods from the different cultures into an (at least somewhat coherent) MMR design. At least two types of approaches can be differentiated: variance-based and case-based (for the following, see Beach & Kaas, 2020).

Variance-based approaches to MMR build on a counterfactual understanding of causation as developed in the Potential Outcome framework. Counterfactual causation is defined as the claim that a cause produced an outcome because its absence would result in the absence of the outcome, all other things being held equal. Without evaluating the difference that a cause can make between the actual and the counterfactual, no causal inference is possible. Therefore, the main causal inference is established at the cross-case level using controlled comparisons. Put it more bluntly, the cross-case method is in the inferential driver's seat, while the within-case serves as an adjunct method.[2] This does not mean that the within-case study is not important. It fulfills crucial functions such as validating measurement, establishing a case's counterfactual, reconstructing the causal pathways, or searching for confounders (Seawright, 2016; Weller & Barnes, 2016). Causal evidence, however, lies across cases.

In case-based approaches to MMR, multiple understandings of causation exist side-by-side (Baumgartner & Falk, 2019; Beach & Pedersen, 2019; Rohlfing & Schneider, 2018). They usually have in common that the inferential workhorse in MMR designs is located at the within-case level instead. To establish a causal relationship, it must be checked whether the identified explanatory factors indeed exert some causal power over the outcome in a case, and if so, how exactly the causal mechanism plays out (e.g., Beach et al., 2019; Schneider & Rohlfing, 2016, 2019). Here, the analysis at the cross-case level plays an adjunct role, e.g., by establishing an X/Y relation in the first place, guiding the case selection for the within-case study, or mapping the population of cases for further generalization (Box 10.1).

Box 10.1: The Variance-Based and the Case-Based Approach to MMR

The question of variance-based and case-based approaches to MMR needs to be located in the broader discussions within the philosophy of sciences (e.g., Cartwright, 2011; Russo & Williamson, 2011) and political science. In this sense, it connects to the seminal readings like King et al. (1994), which argued in favor of a shared understanding of causal inferences across quantitative and qualitative (i.e., empirically oriented case-based methods). This has been challenged in recent debates, which (again) points out the ontological and epistemological differences between the qualitative and quantitative methods (Brady & Collier, 2010). Consequently, there has been a rise of methodological guidelines for different MMR designs depending on the research tradition in which it is grounded (see Beach & Kaas, 2020 for an overview).

Variance-based approaches to MMR (e.g., Lieberman, 2005; Seawright, 2016; Weller & Barnes, 2014, 2016), as pointed out in the main text, usually

(continued)

[2] It is important to note that there are alternative proposals. For instance, Runhardt (2015, 2021) envisages a design where controlled comparisons are used at the within-case level where two or more cases are examined to see whether the proposed mechanism made a difference.

Box 10.1 (continued)
are grounded in the potential-outcomes framework (*aka* counterfactual causation). It applies a top-down perspective where the main goal is to identify robust causal effects in a population of cases, or a sample thereof (with randomized controlled trials as a gold standard). This is followed by an assessment at the within-case level of whether the causal relationship holds or not. The cross-case analysis using controlled comparisons is the main workhorse for causal inference, focusing on difference-making. To align cross-case and within-case analysis, variance-based approaches often understand causal mechanisms as intervening variables whose difference-making can be assessed using controlled comparisons between cases.

For case-based approaches to MMR, the ontological underpinnings are varied, relying on regulatory theory (e.g., QCA) or mechanisms (e.g., process tracing) (Beach & Rohlfing, 2018; Goertz, 2017; Rohlfing & Schneider, 2018; Schneider & Rohlfing, 2016; see also Chaps. 1, 2, 6, and 7). However, what is shared by all existing frameworks is that the main causal inference happens at the within-case level through case study methods like process tracing. In this regard, case-based approaches are bottom-up in their focus on causation as it plays out within single cases, after which generalizations might be made to other cases. As regards the understanding of causal mechanisms, there is an emerging consensus on a productive account of mechanisms – which we also subscribe to in this chapter – that understands mechanisms in the form of actors engaging in activities that link a cause and outcome together in a productive causal relationship. Nevertheless, epistemological discussions are still ongoing about how to identify the working of mechanisms (see also Chaps. 2, 6 and 8).

10.3 The Problem of Mechanistic Heterogeneity for External Validity in MMR

Making generalizations about the working of mechanisms from one studied case to other cases which are not studied is a crucial problem in the social sciences and beyond (e.g., Cartwright, 2011; Khosrowi, 2019; Steel, 2008; Wilde & Parkkinen, 2019). Knowing how a policy intervention works in one case does not necessarily tell us how it would work in other, non-studied cases.

The relevance of this issue is evident in case-based approaches, where the examination of mechanisms is the main inferential workhorse. But the ability to make generalizable claims about mechanisms is also essential for the variance-based approach. For instance, Weller and Barnes (2014, 21) argue that one goal of within-case analysis is "to understand substantive relationships at the level of individual cases and to use those insights to learn something about the population of cases that feature that substantive

relationship." Therefore, large-N mediation analysis (see Chap. 6) is often used to study mechanisms. However, by studying many cases using variance-based methods, one learns about the average causal effects of X (or the intervening variable) on the values of Y. An average does not tell us how the linkage works in any given case. In Cartwright's words, average causal effects tell us that "it works somewhere" while leaving us in the dark about how it actually works in any given case (Cartwright, 2011).

Once we find a causal mechanism in a studied case using within-case analysis, the key question asks whether we can infer that a similar – *nota bene*: not exactly the same (!) – mechanism also connects X and Y in other cases. In other words, how do we ensure the external validity of findings about causal mechanisms? The answer heavily depends on the degree of causal heterogeneity at the within-case level.

We speak of mechanistic homogeneity if two or more sufficiently similar mechanisms are operative in all the cases that exhibit the same relationship between X and Y. Mechanistic heterogeneity, on the other hand, refers to two situations: (1) the same X and Y are linked together through different mechanisms (*mechanistic equifinality*), or (2) the same X triggers different mechanisms leading to a different Y (*mechanistic multifinality*) (Beach, 2020; Beach et al., 2019; Beach & Rohlfing, 2018; Falleti & Lynch, 2009; George & Bennett, 2005; Gerring, 2010; Goertz, 2017; Sayer, 2000; Weller & Barnes, 2016).

It is important to note that we do not understand causal mechanisms as chains of events, but instead as process-level causal explanations that provide an account of what actors are doing. This account explains why the actors' activities are linked together and how they contribute to producing the outcome in the case. Of course, these process-level explanations can have varying levels of detail (*aka* abstraction). At the most abstract level are schematic theories that focus on the most critical interactions, describing actors and what they are doing in very abstract terms (e.g., "a political entrepreneur engages in speeches that attempt to frame a debate"). At the other extreme are very detailed, case-specific accounts that use formal nouns to describe actors, include many different parts, and where activities are specified in great detail (Box 10.2).

Box 10.2: Causal Heterogeneity

The term *causal heterogeneity* includes a range of phenomena linked to complex causal patterns that can characterize any X/Y relationship. In the statistical literature, the problem of causal heterogeneity plays a significant role, for example, when considering whether different subgroups in a given population react differently to a specific treatment, e.g., an administered policy instrument (e.g., Seawright, 2016; Pearl, 2017; Xie, Xie et al., 2012). Issues of causal heterogeneity are also prominent in the context of QCA, where they are discussed concerning conjunctural causation, equifinality, and asymmetry (Ragin, 2008; see also Chap. 7). Yet, researchers must be aware that causal heterogeneity not only pertains to X/Y relations but also to the level of mechanisms (e.g., Beach et al., 2019; Beach & Rohlfing, 2018; Goertz, 2017; Weller & Barnes, 2016).

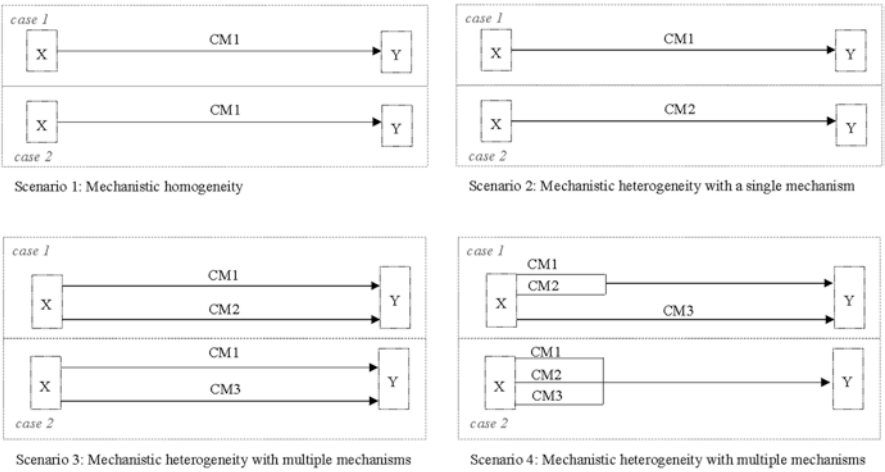

Fig. 10.1 Abstract examples of mechanistic homogeneity and heterogeneity. Own depiction

Figure 10.1 illustrates the issue of mechanistic homogeneity and heterogeneity using causal diagrams in a stylized form for a simple X/Y relationship.
The first scenario displays one variant of mechanistic homogeneity where X and Y are connected via the same mechanism (CM1) in both cases. In contrast, the next situations all refer to different forms of mechanistic heterogeneity.

In the second scenario, two single but different mechanisms connect the same X to the same Y, CM1 in one case and CM2 in another case.[3]

The situation turns more complex in the third scenario. Here, the same X triggers multiple mechanisms in two cases, i.e., mechanistic multifinality, yet there is only one mechanism that is shared by both cases (CM1), whereas the two cases differ on the second mechanism triggered by X, namely, CM2 versus CM3.

Finally, the fourth scenario shows how different mechanisms might interact with each other in different ways across cases – CM1 and CM2 in one case, and CM1, CM2, and CM3 in the second case.

These illustrations are, of course, very simple scenarios. More frequently, explanatory models do not involve one individual factor, but instead several factors X1, X2, X3…, Xi. Here, patterns can become much more complex. Causal mechanisms can work additively or interact with each other, appear in a different sequential order, show complementary instead of conflicting effects (among others, see Beach & Rohlfing, 2018, 18–25; Goertz, 2017, 53–57; Mikkelsen, 2017, 429–34; Weller & Barnes, 2016, 433–37 for further illustrations). For instance, X1 and X2 might trigger two mechanisms, CM1 and CM2, but in one case, this happens simultaneously, whereas in other contexts X1 happens before X2, or even that X1 triggers CM1, which then leads to X2 triggering CM2 – highlighting temporal or causal

[3] Of course, heterogeneity applies when the whole process is different from case to case, but also when parts of it display meaningful diversity.

ordering as reflections of mechanistic heterogeneity. Another example is discussed under the label of 'masking' (Clarke et al., 2014; Steel, 2008, 68; see also George & Bennett, 2005, 145–47). Masking means that a given X might be linked to the same Y through multiple mechanisms with opposite effects on the Y. For instance, a crisis might trigger a process where some actors engage in a frantic search for solutions and advocate for them. At the same time, the same crisis can push other actors to become risk-averse, thereby starting a process of resistance to any change. In the case, both processes might be operative, and the outcome is a compromise on some modest change that either group did not desire.

10.4 Sources of Mechanistic Heterogeneity in MMR

When combining cross-case analysis and within-case analysis in MMR to identify causal mechanisms and make generalizable claims about them, a crucial problem is that the information utilized at the cross-case level is usually uninformative about what is going on at the within-case level of mechanisms. Let us revisit the abstract example displayed in Fig. 10.1: there is simply no way to establish how exactly the mechanisms connecting X and Y play out just by looking at the X/Y relations. Against this backdrop, examining how a mechanism works by studying how it works within one case and generalizing to other unstudied cases is extremely risky. Very different mechanistic scenarios might lurk underneath the same X/Y relationship.

Before we sketch out a generalization strategy sensitive to mechanistic heterogeneity in the next section, we discuss three primary potential sources of mechanistic heterogeneity so that researchers are informed about where to look for heterogeneity pitfalls when generalizing mechanistic claims (Box 10.3).

Box 10.3: Potential Sources of Mechanistic Heterogeneity

As in cross-case analysis, the assumption of causal homogeneity at the level of mechanisms is usually too heroic to be met in the social sciences. We, therefore, argue that mechanistic heterogeneity should be the default assumption when conducting within-case analysis in general and MMR in particular (Beach et al., 2019). Instead of simply assuming that things work in the same way in different cases, researchers should engage in empirical testing of whether mechanistic heterogeneity is present in a population if they want to avoid making flawed generalizations about the working of causal mechanisms.

A *non-exhaustive* list of *non-exclusive* sources of mechanistic heterogeneity includes, inter alia, complex concepts and measures based on multiple attributes with particular causal properties, qualitative hedges within concepts

(continued)

Box 10.3 (continued)
and measures triggering multiple different mechanisms, omitted causal factors and confounders, varying contexts and differences in scope conditions, factors which are identified as redundant or insignificant at the cross-case level, but still have a causal impact at the level of mechanisms, or different forms of temporal and/or causal dynamics which underlie an X/Y relationship.

10.4.1 Complex Concepts or Measures

The first source of mechanistic heterogeneity is that concepts and measures used at the cross-case analysis capture more than one causal property and can trigger multiple mechanisms. Concepts in the social sciences are usually thought of as multidimensional constructs that have several analytical levels, i.e., attributes and indicators (Adcock & Collier, 2001; Goertz, 2020). The literature on concepts and concept formation has developed various strategies for systematizing the constitutive properties of a concept so that they can be fruitfully applied in empirical research.

In the so-called *classical approach* to concept formation, the constitutive attributes of a concept are individually necessary and jointly sufficient (Goertz, 2020; Sartori, 1970). The Venn diagram in Fig. 10.2a illustrates the underlying logic, whereby we start from three constitutive attributes (A, B, C). For a case to be captured by a concept using the classical approach, all three properties must be present – i.e., A and B and C. If only one of the three attributes is missing, the given social phenomenon does not qualify as a manifestation of the concept.

On the other hand, the *family resemblance approach* offers an alternative strategy to concept formation. In contrast to the classic approach, concepts only have sufficient attributes without a specific feature being individually necessary. Under family resemblance, a case is described by a concept when it has at least one of the constituent attributes, regardless of which one. The Venn diagram in Fig. 10.2d illustrates this approach: the presence of either A or B or C – or any combination of the three – is sufficient for the concept to be present (Barrenechea & Castillo, 2019; Goertz, 2020).[4]

Beyond these two standard approaches to concept formation, *mixed types* can also be possible.

In a variant, for instance, there is no single sufficient attribute for having a concept; instead, several conceptual properties must be present, none of which is necessary. To witness, if we require that two out of three attributes need to be present for

[4] In formal terms, the classical approach to concept formation relies on a logical AND combination, marked by the Boolean '*'; i.e., A*B*C. The family resemblance approach is based on the logical OR combination, marked by the Boolean '+', i.e., A+ B PLUS_SPI C. See also Chap. 7 on Qualitative Comparative Analysis.

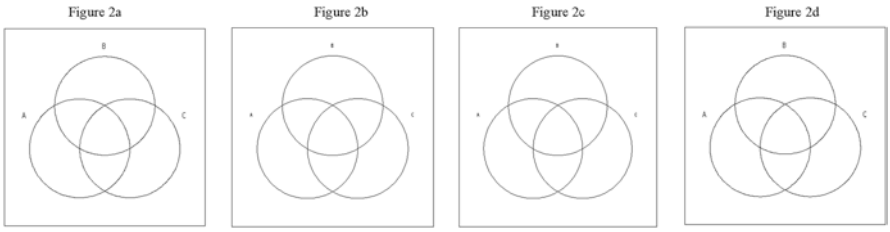

Fig. 10.2 Concept formation strategies and conceptual heterogeneity. Own depiction based on Barrenechea and Castillo (2019)

a concept, this may mean that the concept describes any case showing A and B, or A and C, or B and C, or A and B and C. Figure 10.2c exemplifies this logic based on three ('*n*') conceptual attributes out of which at least two ('*m*') must be given for the concept to apply.

Another mixed type of the two standards approaches is based on the idea that one or more constitutive properties of a concept are necessary, but additional attributes are required but not necessary. For example, thinking again of a concept made up of three attributes A, B, C, we can envisage that A is necessary, but either B or C must be added for a case to be described by the respective concept. As demonstrated in Fig. 10.2b, the concept only applies if another attribute is fulfilled in addition to A.[5]

What does this have to do with mechanistic heterogeneity? The point is that these structures can introduce different levels of (causal) heterogeneity into concepts (Barrenechea & Castillo, 2019; Beach et al., 2019; Collier & Mahon Jr, 1993; Goertz, 2020). As Figure 10.2a highlights, concepts based on necessary and jointly sufficient conditions are very homogeneous since cases are described by this concept only if they show all three attributes. On the other end of the spectrum, concepts that follow a family resemblance logic show a high degree of potential heterogeneity because a total of seven characteristic combinations lead to the presence of the concept – i.e., all combinations except ~A* ~ B* ~ C (Fig. 10.2d). The two mixed types can be located in between. Since different attributes have different causal properties and can trigger different causal mechanisms, it does not need much imagination to envisage that this also leads to mechanistic heterogeneity.

A study by Binder (2015) on the conditions for robust UN interventions in international conflicts illustrates this. Here, the factor 'spillover effects' is conceptualized via three attributes that capture different spillover aspects. The three aspects are, first, refugee flows; second, transnationally operating rebel groups; and third, other negative effects such as drug traffic, terrorism, and economic downturns. To count as a conflict with spillover effect, any of the three factors is sufficient following a family resemblance approach. In such a situation, the cases included in the cross-cases analysis which are coded as experiencing spillover effects contain mechanistic heterogeneity by design: some suffer from only one of these factors,

[5] Formally, this can be expressed by A*(B PLUS_SPI C).

i.e., refugee flows or transnationally operating rebels or economic downturns, others from a combination of two or even all three factors. But the causal mechanisms triggered by each attribute are most probably very different even though they all are coded as cases of 'spillover effect'.

In situations like these, we do not know which mechanism is actually present in a given case just by looking at the relationship between X (here, spillover effects) and Y (here, UN intervention). Hence, we cannot generalize from one case to any other since it is unclear whether cases that only show high refugee flows trigger the same mechanism(s) as cases with only transnationally operating rebels or all three attributes present. At best, we might generalize to cases that share the same configuration of conceptual attributes. But even this is difficult, as we highlight below, since there might still be different dynamics at play among cases that share the same attributes.

The problem of (causal) heterogeneity pertains to various concept formation strategies and complex measures. It also occurs if subtypes are constructed and then used in the form of a ranked scale (Collier & Levitsky, 1997; Møller & Skaaning, 2010). It is inherent to index building which rests on the assumption of homogeneity at different levels of the index (Barrenechea & Castillo, 2019). It may also apply to lexical scales where the defining attributes are hierarchically arranged so that the attribute at the lower level is necessary to the next higher level (Skaaning et al., 2015).

All in all, we should expect that causal heterogeneity, and consequently mechanistic heterogeneity, is pervasive when studying public policy phenomena, especially against the backdrop of the widespread use of complex concepts in cross-case analysis. While this might not be a problem if one is only interested in establishing X/Y relations, it becomes a crucial pitfall in MMR if the aim is to generalize the insights gained at the within-case level to a larger sample of unstudied cases. Simply assuming that causal mechanisms play out in similar ways across all cases would not be warranted in this situation.

10.4.2 Known and Unknown Omitted Conditions

The second source of mechanistic heterogeneity comes from known and/or unknown omitted conditions in cross-case analysis. The problem of *unknown* omitted conditions, i.e., contextual or explanatory factors that are not part of the original model, is frequently discussed in the methodological literature as a problem for MMR (Kuehn & Rohlfing, 2009; Radaelli & Wagemann, 2018; Seawright, 2016; Weller & Barnes, 2016). *Known* omitted conditions, i.e., factors that are not considered in the within-case analysis because they do not make a difference in the cross-case analysis, are less frequently problematized in the literature (but seeÁlamos-Concha et al., 2021 ; Beach et al., 2019 ; Schneider & Rohlfing, 2019).

Conditions omitted in cross-case analysis can substantially impact the within-case level as they can introduce additional mechanisms or interact with existing mechanisms. The problem is straightforward with factors omitted from explanatory

models and is widely discussed, for instance, in the literature as potential confounders (e.g., Goertz, 2017; Radaelli & Wagemann, 2018; Seawright, 2016; Weller & Barnes, 2014). Yet, contextual (*aka*, scope) conditions that are omitted can also play an important role because they can impact how mechanisms operate (i.e., Bunge, 1997; George & Bennett, 2005; Gerring, 2010; Goertz & Mahoney, 2009; Sayer, 2000). This line of thinking also fits nicely into the context-mechanism-outcome (CMO) framework developed by Pawson and Tilley (1997) concerning realistic evaluations. In a nutshell, the framework posits that mechanisms underlying any cause–effect relationship need to be properly contextualized, and whether they work in similar or different ways across varying contexts remains an empirical issue. Returning to the above example of spillover effects and the strength of UN intervention (Binder, 2015), one question concerning the generalizability from one case to another would ask whether the mechanisms differ according to the temporal duration of the conflict. For instance, during a protracted conflict, the intensity of violence might ebb and flow, and there might be several waves of refugees where each wave builds up more and more pressure for international action. A different dynamic might be observed during a short but extremely violent conflict. Of course, whether this is meaningful for treating mechanisms as different depends on the theoretical perspective.

While conditions that are not considered in the analysis can play a crucial role in mechanistic heterogeneity and the generalizability of mechanisms across cases, they are not the only source. One problem we might think of when integrating within-case and cross-case analysis to make generalizations about mechanisms is that explanatory factors might turn out as redundant, irrelevant, or insignificant at the cross-case level, but still have an important causal role to play at the within-case level. This is because, strictly speaking, the level at which causes are operative is always within a single case. Therefore, establishing patterns of difference-makers using statistical techniques or QCA tells us nothing about what is going on within cases. Instead, they only allow us to observe patterns of (in)variation across cases.

For instance, a QCA model might show that condition C is irrelevant since the outcome Y appears together with the presence of C (e.g., ABC) and its absence (e.g., AB~C). In short, C is not a difference-maker from a cross-case perspective (Baumgartner & Falk, 2019; see also Chap. 7). However, once we move down to the case level, the presence or absence of C might be causally relevant for the operation of the mechanism as it still constitutes an analytically important context in which the causal mechanism is embedded (Álamos-Concha et al., 2021; Beach et al., 2019; Schneider & Rohlfing, 2019). The same holds for variables that turn out as (in)significant in regression analyses. All that regressions say is that X has, on average, a particular effect Y, or that it does not; but whether a given factor impacts how the mechanism operates within a given case is an entirely different question that can only be addressed through means of within-case analysis, as this information cannot be derived from the statistical effects (Goertz, 2017; Seawright, 2016; Weller & Barnes, 2014).

In sum, issues like context-sensitivity, proper scoping, or omitted factors as a source of causal heterogeneity are widely acknowledged in the literature discussing

various forms of cross-case and within-case methods. From the perspective of MMR and the task of generalizing causal mechanisms, the problem is aggravated since researchers need to be aware of the limited homogeneity beneath the effect of X on Y and the possibility of multiple mechanisms connecting X and Y across subsets of cases.

10.4.3 Causal and Temporal Dynamics

A third problem when generalizing insights about the working of mechanisms in MMR is that an X/Y relation identified at the cross-case level usually tells us (next to) nothing about the underlying causal and/or temporal dynamics. A look at the literature on within-case studies and MMR discusses a variety of different dynamics that can lurk underneath the same X/Y relationship (Beach & Rohlfing, 2018, 18–25; Beach et al., 2019, 125–28; Blatter & Haverland, 2012, 94; Falleti & Mahoney, 2015, 217; Goertz, 2017, 123–69; Grzymala-Busse, 2011, 1275; Mikkelsen, 2017, 429–34; Weller & Barnes, 2016, 434–35). If unnoticed, they can have a tremendous impact on the generalizability of mechanistic claims since the researcher would assume that the same patterns are linking X in Y in all cases while, in reality, they differ across cases.

One example of mechanistic heterogeneity that can hide behind the same X/Y relation is the temporal sequence of conditions and mechanisms. For instance, a cross-case analysis based on QCA or standard regression techniques might indicate that three factors A, B, C are associated with Y. For illustrational purposes, we use the example of large refugee flows, transnationally operating rebel groups, and other negative effects such as an increase in drug traffic, terrorism, and economic downturns that provoke a robust UN intervention. We can envisage a case where the three factors follow a temporal sequence, according to which the rise of transnational rebel groups (B) first causes an increase in refugee flows (A), which then leads to economic downturns and other negative consequences (C), which finally causes a robust UN humanitarian intervention. Can we now assume that the same sequence is present in all cases? This would probably be a pretty heroic assumption, since many other sequences can still be plausibly theorized. For instance, it might be the case that all three factors appear simultaneously, or the ordering of conditions might be different.

Interaction patterns might be another way that mechanistic heterogeneity manifests itself. For instance, mechanisms might work independently versus conjointly in different cases. Revisiting the example again, the increase in refugee flows, the rise of transnational rebel groups, and negative effects such as an increase in drug traffic, terrorism, and economic downturns might each trigger separate causal mechanisms through different actors and venues that ultimately lead to UN interventions. In other words, A leads to Y, B leads to Y, and C leads to Y through three independent causal mechanisms CM1, CM2, and CM3. However, in other cases, we might find a different situation. One reasonable alternative might be that the three

factors do not show an independent effect, but instead work conjointly, so that each causal mechanism adds or reinforces each other until the UN decides on a robust humanitarian intervention.

It is important to note that these challenges cannot merely be fixed by including interaction terms in regressions or using configurational methods like QCA.[6] Regarding the latter, conjunctions in QCA only tell us that two or more conditions are jointly associated with an outcome; however, they do not tell us anything about the interactions present among the individual conditions within the configuration. Yet the same applies to interaction terms in a regression analysis where we learn that a factor's average causal effect depends on the level of another factor; however, this contains no information on what dynamics and interplays we should expect at the level of mechanisms.

10.5 Taking Mechanistic Heterogeneity in MMR More Seriously

In all the situations described in the previous section, generalizing from one studied case to other cases that have not been studied risks making flawed inferences about which causal mechanisms are operative in different cases. Strictly speaking, we can only know which mechanisms are operative in a given case by investigating that case. This means that researchers are confronted with an inherent trade-off when establishing the external validity of mechanistic claims: examine all cases within a given population at tremendous analytical costs, or make a mechanistic generalization based on hope, with no empirical evidence to substantiate it (Khosrowi, 2019). The trade-off is of special relevance to public policy, where the complexity of processes in different contexts (both across space and time) makes mechanistic heterogeneity likely pervasive.

To engage with this inherent trade-off, we propose a generalization strategy that pays close attention to mechanistic heterogeneity using a sequential, 'cross-case analysis first/within-case analysis second' design. Building on the work by Weller and Barnes (2014, 2016), we advise engaging in multiple follow-up case studies that assess which causal mechanisms are present in strategically selected cases within a population, thereby gradually establishing the boundaries of the external validity of our mechanistic claims. In situations where we find mechanistic heterogeneity, we should map the different causal mechanisms operating in various subsets of the population to clarify why different mechanisms are operative in different

[6] Techniques like mediation analysis, structural equation modeling, or coincidence analysis offer a partial remedy by mapping (causal) chains and sequencing factors. However, other aspects, like whether the speed of events influences the unfolding of complex dynamics between multiple mechanisms, remain open. Additionally, the other sources of mechanistic heterogeneity still play a role.

sub-sets of cases (see Beach et al., 2019, 133–54 for a more detailed discussion) (Box 10.4).

Box 10.4: Strategy for Testing the Generalizability of Mechanisms Under the Assumption of Mechanistic Heterogeneity

The rationale behind the suggested snowballing-outwards procedure is to use findings from within-case analysis to revise the knowledge of the boundaries in which particular mechanisms are operative and progressively update the confidence in the external validity of the mechanistic claims which can and which cannot be made.The proposed strategy consists of the six steps, starting after the cross-case analysis has produced a robust X/Y relationship:

(i) Theoretical unpacking of all potential plausible mechanisms that could link X and Y.
(ii) Mapping of the potential population of cases.
(iii) Initial process tracing of most-similar with population positive case.
(iv) Second process tracing of the positive case that is as similar to initial case as possible.
(v) Gradually probing more dissimilar positive cases, paying close attention to potential sources of mechanistic heterogeneity.
(vi) Concluding with a mechanism-focused comparison of the deviant case(s) to explore potential necessary factors by tracing the breakdown of the mechanism(s) previously identified.

After a robust X/Y relationship is identified at the cross-case level via statistical or configurational methods, the first step of the proposed generalization strategy starts with theoretically unpacking various potential mechanistic explanations. Unpacking mechanisms involves disaggregating causal processes into parts composed of actors doing things.[7] What is necessary at this stage is that researchers make the causal logic underlying the linkages in a mechanism explicit. Doing so also sheds light on all kinds of factors (causal and contextual) that we might expect to be relevant for whether and/or how a given mechanism works. For instance, one pathway might include a part where, to table a proposal that frames a debate, the expert needs to be a trusted epistemic authority by the policymakers. In fact, by theorizing and empirically tracing how a mechanism works, we also shed light on the conditions required for it to work in a particular way.

[7] How to define causal mechanisms is debated within the methodological literature (instead of many, see Beach & Pedersen, 2019; Bennett & Checkel, 2015). Although we cannot get into detail, the problem of generalization and mechanistic heterogeneity is independent of whether one follows a productive account or envisages causal mechanisms rather in terms of intervening factors or very abstract one-liners.

Of course, throughout the next steps, one should still cast the net widely and be open for further evidence about causal mechanisms which have not been hypothesized at this early stage; however, the first step should include a theoretical mapping of the most plausible different mechanistic scenarios and the respective settings in which they might occur.

In the next step, a cross-case mapping of the potential population of cases is undertaken. This involves scoring cases based on values of the explanatory factors X and the outcome Y and potential contextual and causal conditions that might affect how mechanisms work. Here it is crucial to go beyond the identified X/Y relations and to include all analytically relevant (causal or contextual) conditions. In principle, it should be the goal of this mapping to identify clusters of cases as causally homogeneous as possible to minimize the a priori risk of mechanistic heterogeneity.[8]

Based on this mapping, we can select a case for tracing the underlying mechanisms between X and Y. At the initial stage, all positive cases that are members of the X(s), Y, and the given context are potential candidates for process tracing since mechanisms can only be observed in cases where X and Y are present. Ideally, this process tracing identifies one or several mechanisms linking X and Y in a given context C.

However, it might also be the case that no mechanism is identified in the chosen case. Here, we would advise proceeding to another similar case study and checking whether there is also no mechanism linking X and Y. If this is the case, the evidence points towards a mere correlation. Additionally, it could also be that the process tracing reveals one or more contextual factors that impact the working of the mechanism(s), but have not been considered so far. These new contextual features should then be added to revise the mapping of the cases and define more homogeneous subsets.

Based on this initial process tracing of one case, if resources allow it, we should conduct a second study of a case that is as similar as possible on as many relevant causal and contextual factors with the initially studied case. Finding a similar mechanism(s) operative in the second case increases our confidence that the process works similarly across cases. This way, we reduce the risk of missing important factors that might impact how the mechanism works. If, on the other side, we find a different (or no) mechanism(s) operative in a similar case, we would need to look for omitted conditions that differ between the two cases and which explain the difference in the underlying mechanism.

The exploration of mechanistic heterogeneity then continues by strategically selecting more and more different cases to identify the boundaries within which the mechanism operates. When we find different mechanisms operative, we would then want to assess what conditions differ between the cases to understand under which conditions different mechanisms are operative.

[8] To make the mapping compatible with mechanistic explanations when working in variance-based designs, qualitative thresholds for all explanatory, contextual factors, or analytical dimensions need to be established at which a specific mechanism is expected to trigger.

This exercise of empirically testing for mechanistic heterogeneity should be done with an eye to those sources which seem particularly problematic for the research design. For instance, if one of the main explanatory factors is operationalized via a complex concept, one should check whether different causal attributes impact the unfolding of a mechanism. Similarly, researchers should pay close attention to potential interactions, sequencing, and other dynamics among mechanisms that are hidden behind simple X/Y relationships if there is some theoretical or empirical argument that would lead researchers to expect this. In other words, instead of assuming that the same causal mechanism is present in all cases showing X and Y, we encourage researchers to look beyond the results of the cross-case analysis and leverage additional theoretical and empirical insights and probe whether the mechanistic homogeneity or heterogeneity is present in their MMR design.

10.6 Concluding Remarks

One reason for the popularity of MMR is that its main objective coalesces with the evolving consensus in the social sciences that strong causal explanations require evidence of an association between X and Y and evidence for the underlying causal mechanisms between X and Y. The main objective of this chapter was to familiarize researchers with the notion of mechanistic heterogeneity and the challenges this causes when conducting MMR based on some type of cross-case analysis in combination with some form of within-case method. After discussing some basic logics of MMR, we introduced the idea of mechanistic heterogeneity. We highlighted several sources that can bring about causal heterogeneity at the mechanism level in MMR designs. We contend that mechanistic homogeneity is typical when conducting social science research. Starting from the assumption that the social world is characterized by causal complexity, which might be present both at the cross-case level and the within-case level, we must pay more attention to mechanistic heterogeneity when making generalizations about mechanisms. Otherwise, we risk ending up with flawed inferences about the working of causal mechanisms across a sample of cases.[9]

Assuming causal homogeneity at the level of mechanisms makes MMR designs considerably easier. But, as tempting as it might sound, we simply do not know a priori whether this assumption is correct in any given MMR design which strives to integrate insights derived through within-case studies and results from a cross-case analysis. To put it more bluntly, "[...] merely assuming that populations are similar at lower levels would amount to an extrapolation based on hope" (Khosrowi, 2019, 48). Against this backdrop, we call upon researchers to do better than assuming mechanistic homogeneity. Instead, we engage in empirically testing the limits to

[9] Looking beyond the social sciences, causal heterogeneity at the level of mechanisms also plays a crucial role in the life sciences, as the discussions in Steel (2008) and Wilde and Parkkinen (2019) highlight.

which we can generalize mechanistic claims, transparently map out the presence of mechanistic heterogeneity, and establish the proper boundaries for the generalization.

The debate about how to achieve this goal is just beginning. We hope that the guidelines and insights presented in this chapter help to improve research practices and encourage more explicit guidelines on how to address mechanistic heterogeneity while deploying different combinations of methods.

Suggested Readings

1. Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process-Tracing Methods: Foundations and Guidelines*. Second Edition. Ann Arbor: University of Michigan Press.
2. Beach, Derek, and Ingo Rohlfing. 2018. Integrating Cross-Case Analyses and Process Tracing in Set-Theoretic Research: Strategies and Parameters of Debate. *Sociological Methods & Research* 47(1): 3–36.
3. Lieberman, Evan S. 2005. Nested Analysis as a Mixed-Method Strategy for Comparative Research. *American Political Science Review* 99 (03): 435–52.
4. Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools. Strategies for Social Inquiry*. Cambridge: Cambridge University Press.
5. Schneider, Carsten Q., and Ingo Rohlfing. 2016. Case Studies Nested in Fuzzy-Set QCA on Sufficiency: Formalizing Case Selection and Causal Inference. *Sociological Methods & Research 45* (3): 526–68.
6. Weller, Nicholas, and Jeb Barnes. 2016. Pathway Analysis and the Search for Causal Mechanisms. *Sociological Methods & Research* 45 (3): 424–57.

Review Questions

- What are the primary purposes of multimethod research? Can you illustrate the main strengths?
- What pitfalls and trade-offs come with multimethod research?
- How do variance-based and case-based approaches of multimethod research differ?
- Define mechanistic heterogeneity and homogeneity in your own words. Can you give one or two examples of mechanistic heterogeneity from your field of research?
- Discuss how serious you think the problem of mechanistic heterogeneity is in political science? For instance, is it common or only seldom? Does it depend on the understanding of the mechanism, or is mechanistic heterogeneity a problem irrespective of the existing variants?
- Can you illustrate how mechanistic heterogeneity complicates the task of making generalizations in multimethod research?
- Complex concepts, omitted conditions, and causal/temporal dynamics can be seen as major sources for mechanistic heterogeneity when linking cross-case and within-case analysis. Can you think of examples from your field of research which illustrate the described problems?

- Make a list of advantages and disadvantages that come with the strategy that maps and tests the boundaries for generalization in multimethod research. Discuss whether the additional efforts justify the proposed gains. Is generalizing mechanisms based on hope a better strategy from your perspective?

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review, 95*(3), 529–546. https://doi.org/10.1017/S0003055401003100

Ahmed, A., & Sil, R. (2009). Is multi-method research really 'better'? *Qualitative & Multi-Method Research, 7*(2), 2–6.

Ahram, A. I. (2013). Concepts and measurement in multimethod research. *Political Research Quarterly, 66*(2), 280–291.

Álamos-Concha, P., Pattyn, V., Rihoux, B., Schalembier, B., Beach, D., & Cambré, B. (2021, August). Conservative solutions for progress: On solution types when combining QCA with in-depth process-tracing. *Quality & Quantity*. https://doi.org/10.1007/s11135-021-01191-x

Barrenechea, R., & Castillo, I. (2019). The many roads to Rome: Family resemblance concepts in the social sciences. *Quality & Quantity, 53*(1), 107–130. https://doi.org/10.1007/s11135-018-0732-7

Baumgartner, M., & Falk, C. (2019, October). Boolean difference-making: A modern regularity theory of causation. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz047

Beach, D. (2020). Multi-method research in the social sciences: A review of recent frameworks and a way forward. *Government and Opposition, 55*(1), 163–182. https://doi.org/10.1017/gov.2018.53

Beach, Derek and Jonas Gejl Kaas. 2020. The Great Divides: Incommensurability, the Impossibility of Mixed-Methodology and What to Do about It., International Studies Review, 22(2): 214–235

Beach, D., & Kaas, J. G. (2020). The great divides: Incommensurability, the impossibility of mixed-methodology, and what to do about it. *International Studies Review, 22*(2), 214–235. https://doi.org/10.1093/isr/viaa016

Beach, D., & Pedersen, R. B.. (2016). *Causal case study methods: Foundations and guidelines for comparing, matching, and tracing*. University of Michigan Press. http://www.press.umich.edu/6576809

Beach, D., & Pedersen, R. B. (2019). *Process-tracing methods: Foundations and guidelines* (2nd ed.). University of Michigan Press.

Beach, D., & Rohlfing, I. (2018). Integrating cross-case analyses and process tracing in set-theoretic research: Strategies and parameters of debate. *Sociological Methods & Research, 47*(1), 3–36. https://doi.org/10.1177/0049124115613780

Beach, D., Pedersen, R. B., & Siewert, M. B. (2019). Case selection and nesting of process-tracing case studies. In *Process-tracing methods: Foundations and guidelines* (2nd ed.). University of Michigan Press.

Bennett, A., & Checkel, J. T. (Eds.). (2015). *Process tracing. From metaphor to analytic tool. Cambridge*. Cambridge University Press.

Binder, M. (2015). Paths to intervention: What explains the Un's selective response to humanitarian crises? *Journal of Peace Research, 52*(6), 712–726. https://doi.org/10.1177/0022343315585847

Blatter, J., & Haverland, M. (2012). *Designing case studies: Explanatory approaches in small-N research*. Palgrave Macmillan.

Brady, H. E., & Collier, D. (Eds.). (2010). *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed.). Rowman & Littlefield Publishers.

Bryman, A. (2006). *Mixed methods. Sage benchmarks in social research methods*. Sage.

Bunge, M. (1997). Mechanism and explanation. *Philosophy of the Social Sciences, 27*(4), 410–465.

Capano, G., & Howlett, M. (2021). Causal logic and mechanisms in policy design: How and why adopting a mechanistic perspective can improve policy design. *Public Policy and Administration, 36*(2), 141–162. https://doi.org/10.1177/0952076719827068

Capano, G., Howlett, M., & Ramesh, M. (2019). Disentangling the mechanistic chain for better policy design. In G. Capano, H. Michael, M. Ramesh, & A. Virani (Eds.), *Making policies work. First and second-order mechanisms in policy design* (pp. 2–13). Edward Elgar Publishing. https://doi.org/10.4337/9781788118194.00008

Cartwright, N. (2011). Predicting 'it will work for us': (Way) beyond statistics. In P. M. K. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 750–768). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199574131.003.0035

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339–360. https://doi.org/10.1007/s11245-013-9220-9

Collier, D., & Levitsky, S. (1997). Democracy with adjectives: Conceptual innovation in comparative research. *World Politics, 49*(3), 430–451.

Collier, D., & Mahon, J. E., Jr. (1993). Conceptual 'stretching' revisited: Adapting categories in comparative analysis. *American Political Science Review, 87*(4), 845–855.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Falleti, T. G., & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies, 42*(9), 1143–1166.

Falleti, T. G., & Mahoney, J. (2015). The comparative sequential method. In J. Mahoney & K. Thelen (Eds.), *Advances in comparative-historical analysis* (pp. 211–239). Cambridge University Press.

Fielding, N. (2010). Mixed methods research in the real world. *International Journal of Social Research Methodology, 13*(2), 127–138. https://doi.org/10.1080/13645570902996186

Fontaine, G. (2020). Process tracing for comparative policy analysis: A realist approach. In B. Guy Peters, M. Falk, & G. Fontaine (Eds.), *Handbook of research methods and applications in comparative policy analysis* (pp. 273–291). Edward Elgar Publishing.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. The MIT Press.

Gerring, J. (2010). Causal mechanisms: Yes, But…. *Comparative Political Studies, 43*(11), 1499–1526. https://doi.org/10.1177/0010414010376911

Goertz, G. (2017). *Multimethod research, causal mechanisms, and case studies: An integrated approach*. Princeton University Press.

Goertz, G. (2020). *Social science concepts and measurement*. New and completely Rev. Edn. Princeton University Press.

Goertz, G., & Mahoney, J. (2009). Scope in case-study research. In D. Byrne & C. C. Ragin (Eds.), *The Sage handbook of case-based methods* (pp. 307–317). SAGE. https://www.scholars.northwestern.edu/en/publications/scope-in-case-study-research

Grzymala-Busse, A. (2011). Time will tell? Temporality and the analysis of causal mechanisms and processes. *Comparative Political Studies, 44*(9), 1267–1297.

Hendren, K., Luo, Q. E., & Pandey, S. K. (2018). The state of mixed methods research in public administration and public policy. *Public Administration Review, 78*(6), 904–916. https://doi.org/10.1111/puar.12981

Humphreys, M., & Jacobs, A. M. (2015). Mixing methods: A Bayesian approach. *American Political Science Review, 109*(04), 653–673. https://doi.org/10.1017/S0003055415000453

Kay, A., & Baker, P. (2015). What can causal process tracing offer to policy studies? A review of the literature: A review of causal process tracing literature. *Policy Studies Journal, 43*(1), 1–21. https://doi.org/10.1111/psj.12092

Khosrowi, D. (2019). Extrapolation of causal effects – Hopes, assumptions, and the extrapolator's circle. *Journal of Economic Methodology, 26*(1), 45–58. https://doi.org/10.1080/1350178X.2018.1561078

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

Kuehn, D., & Rohlfing, I. (2009). Does it, really? Measurement error and omitted variables in multi-method research. *Qualitative & Multi-Method Research, 7*(2), 18–22.

Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review, 99*(03), 435–452. https://doi.org/10.1017/S0003055405051762

Lindquist, E., & Wellstead, A. (2019). Policy process research and the causal mechanism movement: Reinvigorating the field? In G. Capano, H. Michael, M. Ramesh, & A. Virani (Eds.), *Making policies work. First- and second-order mechanisms in policy design* (pp. 14–38). Edward Elgar Publishing. https://doi.org/10.4337/9781788118194.00009

Löblová, O. (2018). When epistemic communities fail: Exploring the mechanism of policy influence: When epistemic communities fail. *Policy Studies Journal, 46*(1), 160–189. https://doi.org/10.1111/psj.12213

Mikkelsen, K. S. (2017). Fuzzy-set case studies. *Sociological Methods & Research, 46*(3), 422–455. https://doi.org/10.1177/0049124115578032

Møller, J., & Skaaning, S.-E. (2010). Beyond the radial delusion: Conceptualizing and measuring democracy and non-democracy. *International Political Science Review, 31*(3), 261–283. https://doi.org/10.1177/0192512110369522

Pearl, J. (2017). Detecting Latent Heterogeneity. *Sociological Methods & Research, 46*(3), 370–389. https://doi.org/10.1177/0049124115600597

Pawson, R. and Tilley, N. 1997. Realistic Evaluation. London: SAGE Publications.

Radaelli, C. M., & Wagemann, C. (2018). What did I leave out? Omitted variables in regression and qualitative comparative analysis. *European Political Science* online first (January). https://doi.org/10.1057/s41304-017-0142-7.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. University of Chicago Press.

Rohlfing, I. (2008). What you see and what you get: Pitfalls and principles of nested analysis in comparative research. *Comparative Political Studies, 41*(11), 1492–1514. https://doi.org/10.1177/0010414007308019

Rohlfing, I. (2012). *Case studies and causal inference: An integrative framework*. Palgrave Macmillan.

Rohlfing, I., & Schneider, C. Q. (2018). A unifying framework for causal analysis in set-theoretic multimethod research. *Sociological Methods & Research, 47*(1), 37–63. https://doi.org/10.1177/0049124115626170

Runhardt, R. W. 2015. 'Evidence for Causal Mechanisms in Social Science: Recommendations from Woodward's Manipulability Theory of Causation.' Philosophy of Science, 82 (5): 1296-1307.

Runhardt, R.W. 2021. 'Evidential Pluralism and Epistemic Reliability in Political Science: Deciphering Contradictions between Process Tracing Methodologies.' Philosophy of the Social Sciences, 51(4):425-442.

Russo, F., & Williamson, J. (2011). *Generic versus single-case causality: The case of autopsy*, 25.

Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review, 64*(4), 1033–1053. https://doi.org/10.2307/1958356

Sayer, A. (2000). *Realism and social science*. Sage. https://doi.org/10.4135/9781446218730

Schneider, C. Q., & Rohlfing, I. (2016). Case studies nested in fuzzy-set QCA on sufficiency: Formalizing case selection and causal inference. *Sociological Methods & Research, 45*(3), 526–568. https://doi.org/10.1177/0049124114532446

Schneider, C. Q., & Rohlfing, I. (2019). Set-theoretic multimethod research: The role of test corridors and conjunctions for case selection. *Swiss Political Science Review, 25*(3), 253–275. https://doi.org/10.1111/spsr.12382

Schoonenboom, J., & Burke Johnson, R. (2017). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie, 69*(S2), 107–131. https://doi.org/10.1007/s11577-017-0454-1

Schwartz-Shea, P., & Yanow, D. (2012). *Interpretive research design. Concepts and processes*. Routledge.

Seawright, J. (2016). *Multi-method social science: Combining qualitative and quantitative tools. Strategies for social inquiry*. Cambridge University Press.

Skaaning, S.-E., Gerring, J., & Bartusevičius, H. (2015). A lexical index of electoral democracy. *Comparative Political Studies, 48*(12), 1491–1525. https://doi.org/10.1177/0010414015581050

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science* (Environmental ethics and science policy series). Oxford University Press.

Tashakkori, A., & Teddlie, C. (2021). *SAGE handbook of mixed methods in social & behavioral research*. SAGE.

van der Heijden, J., Kuhlmann, J., Lindquist, E., & Wellstead, A. (2019, April). Have policy process scholars embraced causal mechanisms? A review of five popular frameworks. *Public Policy and Administration*, 095207671881489. https://doi.org/10.1177/0952076718814894

Weller, N., & Barnes, J. (2014). *Finding pathways: Mixed-method research for studying causal mechanisms*. Cambridge University Press.

Weller, N., & Barnes, J. (2016). Pathway analysis and the search for causal mechanisms. *Sociological Methods & Research, 45*(3), 424–457. https://doi.org/10.1177/0049124114544420

Wilde, M., & Parkkinen, V.-P. (2019). Extrapolation and the Russo–Williamson thesis. *Synthese, 196*(8), 3251–3262. https://doi.org/10.1007/s11229-017-1573-y

Wolf, F. (2010). Enlightened eclecticism or hazardous hotchpotch? Mixed methods and triangulation strategies in comparative public policy research. *Journal of Mixed Methods Research, 4*(2), 144–167. https://doi.org/10.1177/1558689810364987

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology, 42*(1), 314–347. https://doi.org/10.1177/0081175012452652

Chapter 11
Conclusions. Causality Between Plurality and Unity

Alessia Damonte and Fedra Negri

Abstract The previous chapters convey the image of causal analysis in public policy and beyond as a fragmented field where research communities seldom learn from each other's findings. This chapter resumes the ontological, epistemological, and methodological evidence that causal analysis is characterized by a plurality of objects and "incommensurable" interpretations. It also argues that the same evidence pinpoints how this plurality is complementary at every level, and causal structures raise as the elements that link ontology and methodology and can organize heterogeneous findings to improve learning across accounts.

Learning Objectives
After reading this chapter, you will:

- Understand the different expectations that history and philosophy cast about plurality and unity in approaching causation.
- Appreciate the variety in the ontology, epistemology, and methodology of causal analysis.
- Recognize causal structures as a possible common ground.

11.1 Introduction

As Daniel Little pinpointed in Chap. 2 and Leonce Röth and Andrew Bennett elaborated in Chaps. 6 and 8, the social sciences are home to a variety of understandings of "causation"—regularity, counterfactual, manipulability/interventionist, mechanistic—that have molded research with their particular definitions, methodological commitments, techniques of choice and often a claim of priority over alternatives. In Chap. 10, Markus B. Siewert and Derek Beach warned that, notwithstanding the

A. Damonte (✉)
University of Milan, Milan, Italy
e-mail: alessia.damonte@unimi.it

F. Negri
University of Milan-Bicocca, Milan, Italy
e-mail: fedra.negri@unimib.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_11

optimistic expectations from the mixed-method quarters, these understandings seldom make research strategies suitable to refine each other's findings, for each sheds its light on the phenomena of interest from a particular height and angle. Therefore, causal analysis looks fragmented into discrete approaches, each yielding its piece of knowledge that seemingly cannot speak to the others.

This chapter asks whether such fragmentation is unavoidable, undesirable, or both. To find its answer, it proceeds in two steps. Section 11.2 introduces two opposite accounts of how science is made. One maintains that fragmentation is an undesirable state of "confusion of tongues" and science can only advance under a dominant paradigm pursuing the unification of disciplines by reducing research fields "all the way down" to a few fundamental objects. The other considers that the independence of the research fields makes reduction unnecessary and the variety of research interests makes it highly undesirable; nevertheless, some learning can pragmatically happen as for a wanderer that updates her map along the way. Section 11.3 considers whether the state of the art in causal analysis fits the confusion of tongues or the wanderer metaphor along three dimensions—the ontological, the epistemic, and the methodological. Section 11.4 concludes that the field is intrinsically plural in every dimension; however, accounts are complementary, and causal structures can offer common points of reference for organizing findings into dovetailing portrayals of the "causal elephant."

11.2 Two Tales About the Making of Science

A captivating narrative maintains that science is made in the tension between the two poles of unity and plurality of research mindsets. However, the story turns in different directions depending on one's viewing angle.

11.2.1 The Viewpoint of the History of Science

The first version builds on the idea that science is a social creation and takes historical forms (Kunh, 1996; see Wray, 2011; Sankey, 2019). The modern form comprises "disciplines"—such as chemistry, biology, or economics. The term denotes the distinct body of knowledge that anyone must master before claiming expertise on a subject matter. Disciplines are usually maintained by departments and faculties within colleges and universities. Their members research the subject matter, contribute to its definition by publishing in specialized outlets, and teach courses to train students in the profession. Hence, a discipline arises from the activities of a community committed to some "matrix" of tenets, theories, and practices.

As Thomas Kuhn argues, disciplinary matrixes emerge from the scholarly competition to respond to foundational questions—about the ultimate entities of a research field, their interactions and organization, and the techniques suitable to

know them. A matrix becomes "normal science," the "paradigm" of reference, or the "received view" when it provides a fruitful definition of some fundamental knowledge problem. Often, such definition lies in books and articles that become "classics" in force of a few crucial features: They offer a successful synthesis of previous efforts, restate the legitimate problems of a field, and leave several questions open for research while establishing the method to tackle them (Kuhn, 1996: 10). As more people are trained to address its questions with the methods of reference, old or alternative approaches are "read out of the profession" (*ivi*:19). As a result, the winning matrix dwarfs its competitors and dictates the agenda. In the short run, normal science simply neglects those research issues that do "not fit the box" (*ivi*: 24). In the long run, however, the cumulation of intractable "anomalies" puts normal science into crisis and opens a stage of "extraordinary research" (*ivi*: 90). Possibly, the stage results in a "revolution" and the emergence of a new normal.

In short, this theory assumes that ideas in science follow evolutionary dynamics and tend toward a single equilibrium point at a time. This assumption rests less on evidence about disciplinary trajectories than on prescriptive considerations. Indeed, Kuhn (1996:18) shares with Francis Bacon the tenet that "truth emerges more readily from error than from confusion": Science under a single dominant paradigm, albeit limited in its grasp of the world, is preferable to science under competition. As Kuhn argues, competing disciplinary matrices grow "incommensurable" to one another. In turn, incommensurability makes disciplines "immature" and incapable of relevant advancements.

The obstacle, to Kuhn, is mainly semantic. A competing matrix develops scientific terms that are only meaningful within its original vocabulary, as each term is minted to connect some phenomena to particular theories. Thus, theoretical terms become idiosyncratic lexical constructs and create a specific classification of the subject matter that proves irreducible to any other. Out of the shadow of a dominant paradigm, the scientific discourse proceeds in a confusion of tongues, and the debate across communities unfolds as zero-sum confrontations.

11.2.2 The Perspective of the Philosophy of Science

From the viewpoint of the philosophy of science, the divide runs between "monism" and "pluralism" instead, and the two are understood as research agendas with alternative motivations but of ultimate equal standing.

The monist agenda revolves around the core tenet that "the ultimate aim of a science is to establish a single, complete, and comprehensive account of the natural world (or the part of the world investigated by the science) based on a single set of fundamental principles" (Kellert et al., 2006: *x*). Corollaries of monism are that, at least in principle, such a comprehensive account can describe or explain the world faithfully and strategies of inquiry exist that can produce such a comprehensive account. Scientific monism then turns reducibility into a yardstick to assess the

worth of methods and theories: "methods of inquiry are to be accepted based on whether they can yield such an account"; moreover, "individual theories and models in science are to be evaluated in large part based on whether they provide (or come close to providing) a comprehensive and complete account" (*ibidem*).

Just the opposite, scientific pluralism advocates for an open mind on the nature of causes. It maintains that "there are no definitive arguments for monism and that the multiplicity of approaches that presently characterizes many areas of scientific investigation does not necessarily constitute a deficiency" (Kellert et al., 2006: *x*). In principle, pluralism does not deny the possibility that an encompassing account of the world can be found that effectively allows reducing complexity to the same objects "all the way down." However, it addresses this possibility as an empirical matter decided by evidence that may never prove conclusive.

Besides, the coexistence of various accounts across and within disciplines does not undermine the standing of the knowledge so yielded. Crucially, pluralism commits to maintaining that theories and methods cannot be rejected as "unscientific" on the grounds that they fail to reduce complexity to the same fundamental principle (e.g., Fodor, 1974; Longino, 2013). Pluralism finds the reason for incommensurable approaches in the diversity of the research questions that can be asked. Considerations about the relative autonomy of research fields (e.g., Dupré, 1993), the irrelevance of reducibility to the validity of findings (e.g., Suppes, 1978), and the dappled nature of the world (e.g., Cartwright, 1999) further reinforced the stance. In short, phenomena might be "too complicated or too indeterminate and our cognitive interests too diverse for the monist ideals" (Kellert et al., 2006: *xi*).

Nevertheless, these considerations do not license the conclusion that literally "anything goes." Paul Feyerabend (1993) minted that dictum as the single pluralist principle in a Dadaist mockery of monism—given that, as such, scientific pluralism remains skeptical about the possibility of single fundamental principles in doing science. Instead, the dictum calls for recognizing that any approach has its limits, even when it seems unquestionable. Therefore, science advances when its rules make room for a pragmatic conversation between theories and evidence of any stripes, as a wanderer that updates her map along the way (*ivi*: 223 ff).

11.3 Can We Learn from One Another?

Both the confusion of tongues and the wanderer metaphors fit the causal landscape of policy studies and social sciences, leaving the question open of whether pragmatic learning can happen across the research communities that inhabit them or strict incommensurability reigns instead. The issue can be addressed along three conventional lines (e.g., Della Porta & Keating, 2008): the ontological, the epistemological, and the methodological.

11.3.1 Ontological Incommensurability?

Causal ontologies are assumptions about the kinds of ultimate "objects" in a causal account. They are crucial as they indicate where causal analysis legitimately "bottoms out" while avoiding the chasm of infinite regress or circularity. However, the concept has long proven contentious, as it can mean a commitment to dogmas that outweigh evidence instead of some ground for meaningful methodological choices (e.g., Woodward, 2015; see also Damonte & Negri, Chap. 1).

As discussed by Daniel Little and Andrew Bennett in Chaps. 2 and 8, of the four approaches to causality (i.e., regularity, counterfactual, experimental, and mechanistic), the mechanistic stands out as it offers a convenient ultimate ground. Beyond evading infinite regress and circularity, mechanisms can prevent causality from being reduced to non-causal objects such as constant conjunctions or methodological criteria such as counterfactual reasoning. Without some mechanist account of the nature of the process that generates the observed outcome, non-causal objects are analytically unsatisfying and offer a rough guide to policy choices. As Eric Battistin and Marco Bertoni discussed in Chap. 2, the experimental approach aims at getting as close as possible to causal identification by manipulating the candidate causal factor under controlled conditions. However, the credibility of the findings obtained through manipulation stems from the credibility of the assumptions about the background whence, as Leonce Röth adds in Chap. 6, unknown confounders can operate that bias causal identification. Mechanisms provide testable hypotheses about the relevant covariates in the background, hence make sense of regularity and circumscribe counterfactual reasoning about the outcome to limited regions of the world (e.g., Cartwright et al., 2020; Glennan, 2017; Illari & Williamson, 2012; Machamer et al., 2000; Salmon, 1994).

Scholars from theory-driven areas find mechanistic assumptions easy to embrace (e.g., Peters, 2022; Dowding & Miller 2019; Busetti & Dente, 2018). The approach is also increasingly accepted within research communities concerned that substantive assumptions may impress biases in conclusions (e.g., Imbens, 2020; Imai et al., 2013). However, the literature contends that the concept can be elusive and its definitions at cross purposes (e.g., Mahoney, 2021; Mayntz, 2020; Seawright 2018; Goertz, 2017; Gerring, 2011; Pearl, 2000; Holland, 1988; see also Little, Chap. 2, Röth, Chap. 6, Bennett, Chap. 8, and Beach & Siewert, Chap. 10 in this volume).

Against this backdrop, Wesley C. Salmon (1987, 1994; Dowe, 2000; see also George & Bennett, 2005) provides an encompassing definition that also proves sensitive to the many desiderata in causal ontologies. His starting point is Bertrand Russell's grasp of causality as the seamless "persistence of something" across space and time (1948:459). To preserve the emphasis on the factual side of causation while improving the ability to distinguish it from non-causal phenomena, Salmon borrows from the physical understanding of energy and defines causality as the seamless transmission of some non-null "conserved quantity" across space and time.

As such, causality is singular and inheres to entities as different as still paperweights, thrown baseballs, sent data packets, enacted policy instruments, or engaged

strategic actors. Moreover, it exists in the time window between two distinct alterations, regardless of how narrow that window seems to an observer. In turn, alterations occur at *intersections*—the concept that allows discriminating between causal and non-causal transmission processes.

Following Hans Reichenbach (1956), Salmon identifies three possible alterations that a causal quantity can undergo when intersected:

- First, it can fork into two or more quantities and transmission processes. An observer understands these λ-intersections as a "common cause" giving rise to different outcomes.
- Second, it can merge with one or more preserved quantities into a new one. An observer appreciates these γ-intersections as the "joint production" of a single outcome from independent causal factors.
- Third and more conventional, it can exchange its quantity with another causal process. The observer recognizes these χ-intersections as chained transmissions of the "conserved quantity" to the outcome.

The movement of the conserved quantity across time and places is the "causal rope" connecting two intersections; the other way round, intersections are the starting and the ending point of any specific causal rope. Albeit the "causal elephant" only arises in force of both, it can be addressed as either the causal line of a conserved quantity or as its λ, γ, and χ generation structures.

These complementary viewpoints make the mechanistic ontology intrinsically plural. Indeed, the transferral of "conserved quantities" and linked intersections require different vocabularies to be spoken of. However, each account implies the other—which, in principle, makes room for pragmatic matching and learning. Whether this happens, however, depends on epistemic conditions.

11.3.2 Epistemic Incommensurability?

The epistemic level comprises the responses to the question of how we know causation. The question implies a further broad distinction between "foundationalists" (e.g., Christensen, 2004; Kaplan, 1994) and "naturalists" (e.g., Kornblith, 1980; Quine, 1969; cfr. Bevir & Kedar, 2008). In the former camp, the main question is how we *should* know causation. The response builds on a vision of scientific epistemology as rules and standards deployed to establish cogent evidentiary arguments. Scholars in the latter camp instead focus on *how it happens* that human beings know causation. They share an interest in knowledge as individual and social belief systems shaped by psychological and interactive sense-making processes.

The plurality of the positions within and across camps is mirrored by the many interpretations of probability deployed over time. Probability turns our conjectures about "something" being such and such instead of anything else into explicit and inspectable conditional relationships (e.g., Hàjek, 2007). Such conditionality supports our efforts to predict or retrodict events and make decisions even when our

understanding of their determination is limited, our information is partial, or the world appears indeterminate. However, the same conditionality can afford a large number of readings. Gillies (2000:1; cfr. Weatherford, 1982; Fine, 1973; Kyburg, 1970; Salmon, 1966) identifies four major interpretations:

– *Frequentism* (e.g., von Mises, 1964; de Laplace, 1820) understands probability as the limit of the relative frequency of a kind of event in a long series of trials—or, in its classic version, as the ratio of the outcome of interest to the possible outcomes of a single trial.
– *Propensity* (e.g., Suppes, 1987; Popper, 1959) reads probability as the inclination to realize an event of interest that inheres in selected repeatable conditions.
– *Logical* probability (e.g., Carnap, 1952; Keynes, 1921) gauges the degree of belief that any rational mind would entertain about the holding of the relationship between any two or more propositions given specific evidence.
– *Subjective* understandings (e.g., De Finetti, 1989; Ramsey, 1964) define probability as a degree of credence or expectation of some event that single individuals can express as consistent betting quotients but that may defy substantive rationality.

The logical and the subjective interpretations are often grouped together for their shared focus on human heuristics. In contrast, the frequentist and the propensity readings both assume that probability is independent of the single individual mind—which, customarily, qualifies it as "objective." However, the propensity interpretation differs from the pure frequentist: The latter limits itself to "collectives," while propensity makes room for the conditional probability of individual events. As a consequence, frequentists tend to commit to parametric analysis to preserve accuracy in estimates, whereas propensity interpretations usually support non-parametric procedures and, as such, trade accuracy for the flexibility afforded by weaker or no assumptions about the true distribution of the phenomenon of interest.

The expectation camp, too, is easily associated with non-parametric procedures; however, the logical diverges from the subjective interpretation. The former considers information from rational inference structures as a reason for dismissing a relationship between sentences, whereas the latter maintains that the only misleading probability is the inconsistent one. Thus, logical interpretations are concerned with the soundness of the conclusion they license, whereas subjective interpretations allow absurd beliefs about the world as long as the relationship between odds against and in favor meets the formal axioms of probability calculus.

All in all, these interpretations patently fit the confusion of tongues. Radical subjectivist assumptions annoy those who see them as a license to retain fallacies in reasoning (e.g., Hájek, 2007). Propensity is in the odor of metaphysical speculation, and its causal assumptions imply asymmetries that do not fit the standard axioms of probability (e.g., Humphreys, 1985). Deceptive is equally deemed the claim that mathematical a priori tenets – such as the Law of Large Numbers and the Central Limit Theorem, or the classical Principle of Indifference—confer priority to frequentist probability because they render the ultimate nature of the world (e.g., Freedman, 2010). Logical interpretations appear as deductive as the frequentist and,

in addition, are charged with entertaining highly implausible assumptions about human heuristics (e.g., van Fraassen, 1989).

However, once again, each interpretation suits a particular research interest and, pragmatically, they all can be deployed to illuminate the whole of the "causal elephant" from different angles and heights. However, this does not imply that the methods through which different interpretations are deployed can yield dovetailing knowledge.

11.3.3 Methodological Incommensurability?

Ascertaining causation has long been a pluralistic matter and has often provided a substitute for ontological assumptions (e.g., Rohlfing & Zuber, 2021, Brady, 2008; see Little, Chap. 2). As recalled by Alessia Damonte and Fedra Negri in Chap. 1 and elaborated by Daniel Little in Chap. 2, the influential Humean ideal establishes that a local causal relationship meets two criteria: First, conditions similar to the observed local ones provide the regular antecedents of the outcomes similar to the observed one (i.e., regularity); second, had our local conditions been absent, then the local outcome should have taken a different magnitude or state than observed (i.e., counterfactual). Otherwise said, the methods to ascertain causation can be reduced to the alternative between "enumeration" and "elimination" (e.g., Hintikka, 1968). Notably, each criterion operates at a distinct level:

- Enumeration turns establishing causation into a *quantitative* issue—in its basic version, it means counting the cases where conditions of the same kind precede outcomes of the same kind in the instances of the condition across time and contexts.
- Elimination relies on a *qualitative* change in the setting of the original situation instead—that is, the switch in the state of the condition to switch the state of the outcome.

In moving from an observation to the claim that the observation is causal, the two criteria have long been recognized with different weights. Enumeration can yield lawlike generalizations that capture the robustness of the relationship between kinds across contexts but that, as such, cannot support the claim that the relationship has a causal standing. Barometer readings and storms, hoaxes and salt dissolving in water, birth control pills and biological male pregnancy—all these relationships can pass enumeration, but not elimination. The storm would have occurred had the barometer been broken, the salt would still have dissolved in water if unhoaxed, and Mr. Smith would not have gotten pregnant had he ingested aspirins instead. Thus, elimination better supports the intuition that the relationship is effective and that Salmon's "conserved quantity" yielded the outcome. However, Humean local elimination confronts the long-acknowledged "fundamental problem of causal inference": We cannot rerun history to observe the local outcome in the absence or under

different local conditions while holding all the other potential confounders constant (e.g., Holland & Rubin, 1987; see also Battistin & Bertoni Chap. 3, Negri Chap. 4, Ornstein Chap. 5).

11.3.3.1 Design-Based Solutions

The purposeful selection or construction of observation units as "instances" or "cases" enter as suitable methodological solutions to circumvent the fundamental problem of causal inference by making counterfactuals somehow observable. John Stuart Mill (1843) famously systematized the practices and knowledge of the time into two primary designs plus three elaborations. The two basic designs build on the Humean standards as they proceed:

1. By *agreement*: The condition and the outcome stand in a causal relationship if two or more instances of the outcome are dissimilar in every relevant feature except the condition—or two or more instances of the condition are dissimilar in every relevant feature except the outcome.
2. By *difference*: The condition and the outcome stand in a causal relationship if two cases that are similar in every relevant feature except the condition also differ by the outcome.

The three further elaborations state that:

3. *Joint agreement and difference, or indirect difference*: A condition and one outcome stand in a causal relationship when either the presence of both or the absence of both is the only common feature of matching groups composed of dissimilar instances.
4. *Residues*: If we know that a set of conditions yields a certain quantity of the outcome in a group of instances, and in a matching group we know that there is the same set of conditions plus one and one only, then the additional part of the outcome can be ascribed to that further condition.
5. *Concomitant variations*: If two phenomena vary in tandem, they are connected by some "fact of causation."

Of the five canons, the latter only suits continuous-valued phenomena—in all the remaining designs, phenomena are units' binary qualities. Noticeably, the method of concomitant variations also stands out as it cannot establish that the relationship is causal in itself—only that it suggests some causal "fact" (see Negri, Chap. 4).

The other designs are deemed more conclusive as they rely on selected combinations of qualitative diversity in backgrounds, outcomes, and conditions to dismiss the hypothesis that the conditions in the background are relevant to the relationship of interest (agreement) or that the relationship includes causally irrelevant elements (direct difference, indirect difference, and residues). Of the two threats, Mill maintained the latter is more harmful to the standing of the claim that the relationship is causal, which makes difference-based designs more conclusive. Agreement

remained the design of reference for studies where the assumptions of the most similar background could prove harder to attain; its double deployment as the indirect method of difference was offered as a strategy to license more credible conclusions.

With a grain of salt, the reasoning behind these canons has been standing the test of time. While comparative strategies seldom made a secret of their debt toward indirect difference as their design of reference (e.g., Mahoney, 2021, also see Damonte Chap. 7), it is also hard not to notice how the estimation of the effect in Randomized Controlled Trials shares the rationale of Mill's residues. The same holds for the weaknesses that Mill himself recognized. Design-based inferences can license claims that a relationship is causal but cannot ascertain its direction, absent further assumptions and information. Moreover, "causes" can prove:

– *Plural*, as the same outcome can be "overdetermined"—which raises causal heterogeneity issues often hard to disentangle (see Beach & Siewert, Chap. 10). The same outcome can follow from alternative conditions and processes: For instance, emission trading and environmental regulation can both compel a reduction of carbon emissions. But it may also be that different processes yield the same outcome under the same conditions: For instance, individuals may comply with the same rule due to sheer calculations of advantages and disadvantages of noncompliance, loyalty toward the government or deference toward authority, or the persuasion that it is the right thing to do—in different mixes, but all at once (e.g., Schneider & Ingram, 1990).
– *Composite*, as a causal factor can comprise different components. Moreover, composition comes in two flavors, as it can follow:

  • A *physical* rationale and result from the algebraic sum of its components pointing in different directions, as in the composition of forces. For instance, someone's calculation about compliance may depend on their preferences for noncompliance and information on how likely the penalty is applied (e.g., Klepper & Nagin, 1989). Or it may be that some catch-22 regulations made the original decision to comply impossible to pursue.
  • A *chemical* rationale and result from interactions raising a qualitatively different outcome. For instance, the individual decision to not comply may prove perfectly rational from the individual perspective in the short term, yet turn into a tragedy when the decision spoils a common good and is made under an institutional design that allows opportunism to spill over (Ostrom, 2009).

To prove that the antecedent has some causal import, difference-based designs have to dismiss plurality and composition as background "noise" or part of some "ceteris paribus" clause. However, without knowing how and under which conditions the causal connection holds, the conclusions are possibly inaccurate as their assumptions about the comparability of instances may not hold (e.g., Dunning et al., 2019; Trampush & Palier, 2016; Morgan & Winship, 2015; Cartwright & Hardie, 2012; Imai et al., 2011; Salmon, 1990; Campbell & Stanley, 1963).

11.3.3.2 Model-Based Solutions

The increasing attention to causal models responds to the need for testable structural assumptions. It revives the factual side of causal analysis and revolves around a few options, all resonating with Mill's intuition of plural and composite factors but seldom corresponding perfectly.

For instance, Patricia L. Kendall and Paul F. Lazarsfeld (1950; see also Morgan & Winship, 2015) introduce structures to "elaborate" a correlation of interest and so improve its credibility. These structures emerge by stratifying the relationship between X and Y by a multi-value test factor T. Thus, T "interprets" the relationship if it occurs after X but before Y, as in physical composition. Instead, T "explains away" the relationship if it occurs before X and Y—a relationship that Mill would classify as a "fact of causation" without an autonomous shape. The further elaboration "specifies" the relationship by considering the circumstances that affect the partial relationship between X and Y within each stratum of T. Morgan and Winship (2015) note that specification implies an intransitive relationship of T with either X or Y, which may resonate with Mill's chemical composition (with X) or plurality (with Y).

Causal structures also are the crux of Pearl (2000; see also Röth, Chap. 6). His approach, too, considers these structures as the solution to the problem of identification. The causal standing of a relationship always builds on three terms—the alleged causal factor X, the outcome factor Y, and the additional term Z—arranged in three fundamental shapes and visualized as directed acyclic graphs—the "chain," the "fork," and the "collider." In the chain, Z is the mediator between X and Y; in the fork, it is the common cause of X and Y; in the collider, it is the effect of Y and, independently, of X. Then, the chain corresponds with Mill's physical composition and the collider with Mill's plurality. In Mill's terms, Pearl's fork again is a "fact of causation." Mill's chemical composition, instead, is discussed as the problem of identifying causal intransitivity in chained structural models (e.g., Halpern, 2016; von Sydow et al., 2016; Hitchcock, 2001).

Albeit the confusion of tongues seems to reign again among model-based strategies, here the translation problem does not seem to imply real incommensurability—just blind spots and labeling issues.

11.4 Wrapping Up and Looking Ahead

This chapter asked whether the different techniques in causal analysis can learn from each other or incommensurability rules instead. The portrayals sketched above suggest that incommensurability hides many complementarities between interests in processes or intersections and between "objective" and "subjective" interpretations of probability. However, interests and interpretations cannot dovetail unless they build on some common ground. Such possible common ground consists of causal structures.

On the one hand, causal structures arise threats to the identification of the effect of a single factor that designs aim to keep at bay; on the other, they offer the scaffolding for testable models of how and why the effect occurs. Moreover, causal structures connect methodologies with ontological assumptions – albeit far from perfectly so, as summarized in Table 11.1.

Table 11.1 highlights how ontological and methodological viewpoints shed their unique blind spots on structural alternatives. Mill does not consider the common cause as a proper causal structure, for it raises the spurious correlation that enumerative strategies mistake for causal, while Reichenbach and Salmon seemingly disregard structures that could be labeled "disjoint" as they depend on alternative processes, thus suggesting an analytical focus on one "conserved quantity" at a time. In turn, Pearl's graphs do not identify Mill's chemical composition as a distinct shape—possibly treating it as a path in a fork or a version of the chain structure and as a matter of the debate on how to identify actual instances of intransitive causation from sheer dependence. Last, Kendall and Lazarsfeld develop their typology as explorations of facts of causation.

Beyond the differences in standing and usage, these structures promise to offer the terrain where otherwise diverse research strategies can trade their findings, provided that they acknowledge the peculiarities of each other's language. Indeed, ideally, structural assumptions can accommodate results generated with different grammar and syntax rules while addressing the same policy concern. Frequentist probability can yield robust estimates of some effect of interest of Salmon's "conserved quantity" and, hence, support decisions on whether the treatment is worth the policy effort. Propensity probability can assess Salmon's intersection or Reichenbach's reference class to yield more fine-graded estimates of the effect in selected subpopulations. The logical probability can establish whether a reference class makes a sound singular account and afford the *ex-post* evaluation of interventions while improving forecasting. Subjective probability narrows on individual expectations and exposes the heuristics beneath our decisions as policytakers and policymakers—which can only be evaluated in light of knowledge and assumptions about logical reasoning and "objective" evidence.

Strategies and techniques create families that can be accommodated into a single low-dimensional space only at the cost of inviting outraged objections. Nevertheless, we are positive that the efforts of the next generation of eclectic causal analyses to elucidate causal structures can contribute to building more integrated multidimensional maps of crucial policy, political, and social phenomena.

Table 11.1 Causal structures

Graph	Reichenbach & Salmon	Mill	Kendall & Lazarsfeld	Pearl
$X \rightarrow Z \rightarrow Y$	χ-Transmission	Physical composition	Interpretation	Chain
$Z * X \rightarrow Y$	γ-Joint production	Chemical composition	(X-)Specification	*Fork path*
$X \leftarrow Z \rightarrow Y$	λ-Common cause	*Fact of causation*	Explanation (away)	Fork
$X \rightarrow Z \leftarrow Y$	*Disjoint production*	Plurality	*(Y-)Specification*	Collider

Source: own elaboration. References in the main text

References

Bevir, Mark and Asaf Kedar. (2008). "Concept Formation in Political Science: An Anti-Naturalist Critique of Qualitative Methodology." Perspectives on Politics 6(3), 503–17. https://doi.org/10.1017/S1537592708081255

Brady, H. E. (2008). Causation and explanation in social science. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 217–270). Oxford University Press.

Busetti, S., & Dente, B. (2018). Designing multi-actor implementation: A mechanism-based approach. *Public Policy and Administration, 33*(1), 46–65.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Rand McNally.

Carnap, R. (1952). *The continuum of inductive methods*. University of Chicago Press.

Cartwright, N. (1995). 'Ceteris paribus' laws and socio-economic machines. *The Monist*, *78*(3), 276–294. jstor.org/stable/27903437

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cartwright, N., Pemberton, J., & Wieten, S. (2020). Mechanisms, laws and explanation. *European Journal for Philosophy of Science, 10*(3), 1–19. https://doi.org/10.1007/s13194-020-00284-y

Christensen, D. (2004). *Putting logic in its place: Formal constraints on rational belief*. Oxford University Press. https://doi.org/10.1093/0199263256.001.0001

De Finetti, B. (1989). Probabilism: A critical essay on the theory of probability and on the value of science. *Erkenntnis, 31*(2–3), 169–223. jstor.org/stable/20012237

de Laplace, P. S. (1820). *Théorie Analytique Des Probabilités*. Courcier.

Della Porta, D., & Keating, M. (2008). Introduction. In Id (Eds.). *Approaches and methodologies in the social sciences: A pluralist perspective*. Cambridge University Press.

Dowding, K., & Miller, C. (2019). On prediction in political science. *European Journal of Political Research, 58*(3), 1001–1018. https://doi.org/10.1111/1475-6765.12319

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., & Nellis, G. (2019). Informational interventions: theory and measurement. In Id. (Eds.). *Information, accountability, and cumulative learning: lessons from Metaketa I* (pp. 50–77). Cambridge University Press. https://doi.org/10.1017/9781108381390.

Dupré, J. (1993). *The disorder of things. Metaphysical foundations of the disunity of science*. Harvard University Press.

Feyerabend, P. (1993). *Against method* (3rd ed.). Verso.

Fine, T. L. (1973). *Theories of probability: An examination of foundations*. Academic Press.

Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthèse, 28*, 97–115. jstor.org/stable/20114958

Freedman, D. A. (2010). *Statistical models and causal inference. A dialogue with the social sciences*. Cambridge University Press.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. The MIT Press.

Gerring, J. (2011). *Social science methodology: A unified framework*. Cambridge University Press.

Gillies, D. (2000). *Philosophical theories of probabilities*. Routledge.

Glennan, Stuart (2017). The New Mechanical Philosophy. Oxford University Press.

Goertz, G. (2017). *Multimethod research, causal mechanisms, and case studies: An integrated approach*. Princeton University Press.

Goertz, G. (2020). *Social science concepts and measurement: New and completely revised edition*. Princeton University Press.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese, 156*(3), 563–585. https://doi.org/10.1007/s11229-006-9138-5

Halpern, J. Y. (2015). *A modification of the Halpern-Pearl definition of causality*. In Twenty-fourth international joint conference on artificial intelligence (pp. 3022–3033). https://doi.org/10.5555/2832581.2832671.

Halpern, J. Y. (2016). Sufficient conditions for causality to be transitive. *Philosophy of Science, 83*(2), 213–226. https://doi.org/10.1086/684915

Hintikka, J. (1968). Induction by enumeration and induction by elimination. *Studies in Logic and the Foundations of Mathematics, 51*, 191–231. https://doi.org/10.1016/S0049-237X(08)71045-0

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy, 98*(6), 273–299. jstor.org/stable/2678432

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series, 1988*(1), i–50. https://doi.org/10.1002/j.2330-8516.1988.tb00270.x

Holland, P. W., & Rubin, D. B. (1987). Causal inference in retrospective studies. *ETS Research Report Series, 1987*(1), 203–231. https://doi.org/10.1002/j.2330-8516.1987.tb00211.x

Humphreys, P. (1985). Why propensities cannot be probabilities. *The Philosophical Review, 94*(4), 557–570. https://doi.org/10.2307/2185246

Illari, P.M., Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. European Journal for Philosophy of Science 2, 119–135.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789. https://doi.org/10.1017/S0003055411000414

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 5–51. https://doi.org/10.1111/j.1467-985X.2012.01032.x

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature, 58*(4), 1129–1179. https://doi.org/10.1257/jel.20191597

Kaplan, M. (1994). Epistemology denatured. *Midwest Studies in Philosophy, 19*, 350–365. https://doi.org/10.1111/j.1475-4975.1994.tb00294.x

Kellert, S. H., Longino, H. E., & Waters, C. K. (2006). Introduction: The pluralist stance. In Idem (Ed.). *Scientific pluralism* (pp. 1–25). University of Minnesota Press.

Kendall, P. L., & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. F. Lazarsfeld (Eds.), *Continuities in social research: Studies in the scope and method of "the American soldier"* (pp. 133–196). The Free Press.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan.

Khalidi, M. A. (2001). Incommensurability. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 172–180). Blackwell. https://doi.org/10.1002/9781405164481.ch27

Klepper, S., & Nagin, D. (1989). The deterrent effect of perceived certainty and severity of punishment revisited. *Criminology, 27*(4), 721–746. https://doi.org/10.1111/j.1745-9125.1989.tb01052.x

Kornblith, H. (1980). Beyond foundationalism and the coherence theory. *The Journal of Philosophy*, *77*(10), 597–612. jstor.org/stable/2025943.

Kunh, T. S. (1996). *[1962] the structure of scientific revolutions* (3rd ed.). University of Chicago Press.

Kyburg, H. E. (1970). *Probability and inductive logic*. Macmillan.

Longino, H. E. (2013). *Studying human behavior*. University of Chicago Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25. https://doi.org/10.1086/392759

Mahoney, J. (2021). *The logic of social science*. Princeton University Press.

Mayntz, R. (2020). *Causal mechanism and explanation in social science* (MPIfG discussion paper no. 20/7). handle.net/10419/218729

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive*. Harper & Brothers.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
Ostrom, E. (2009). *Understanding institutional diversity*. Princeton University Press.
Pearl, J. (2000). *Causality*. Cambridge University Press.
Peters, B. G. (2022). Can we be casual about being causal? *Journal of Comparative Policy Analysis: Research and Practice, 24*(1), 73–86. https://doi.org/10.1080/13876988.2020.1793327
Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25–42.
Quine, W. O. V. (1969). Epistemology naturalized. In Id, *Ontological relativity and other essays* (pp. 69–90). Columbia University Press.
Ramsey, F. P. (1964). Truth and probability. In H. E. Kyburg Jr. & H. E. Smokler (Eds.), *Studies in subjective probability* (pp. 23–52). Krieger Publishing.
Reichenbach. (1956). *The direction of time*. University of Los Angeles Press.
Rohlfing, I., & Zuber, C. I. (2021). Check your truth conditions! Clarifying the relationship between theories of causation and social science methods for causal inference. *Sociological Methods & Research, 50*(4), 1623–1659. https://doi.org/10.1177/0049124119826156
Russell, B. (1948). *Human knowledge*.
Salmon, W. C. (1966). *The foundations of scientific inference*. University of Pittsburgh Press.
Salmon, W. C. (1987). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Salmon, W. C. (1990). Scientific explanation: causation and unification. *Crìtica, 22*(66), 3–23. jstor.org/stable/40104633
Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science, 61*, 297–312. https://doi.org/10.1086/289801
Sankey, H. (2019). *The incommensurability thesis* (2nd ed.). Routledge.
Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics, 52*(2), 510–529. https://doi.org/10.2307/2131904
Seawright, J. (2018). *Multi-method social science: Combining quantitative and qualitative tools*. Cambridge University Press.
Shaffer, P. (2018). *Causal pluralism and mixed methods in the analysis of poverty dynamics* (WIDER working paper no. 2018/115). handle.net/10419/190162
Suppes, P. (1978). The plurality of science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 3–16. jstor.org/stable/192459
Suppes, P. (1987). Propensity representations of probability. *Erkenntnis 26*(3), 335–358. jstor.org/stable/20012084
Trampusch, C., & Palier, B. (2016). Between X and Y: How process tracing contributes to opening the black box of causality. *New Political Economy, 21*(5), 437–454. https://doi.org/10.1080/13563467.2015.1134465
van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford University Press.
von Mises, R. (1964). *Mathematical theory of probability and statistics*. Edited and complemented by Hilda Geiringer. Academic Press.
von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory and Cognition, 44*, 469–487. https://doi.org/10.3758/s13421-015-0568-5
Weatherford, R. (1982). *Philosophical foundations of probability theory*. Routledge and Kegan Paul.
Woodward, J. (2015). Methodology, ontology, and interventionism. *Synthese, 192*(11), 3577–3599. https://doi.org/10.1007/s11229-014-0479-1
Woodward, J. (2016). The problem of variable choice. *Synthese, 193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5
Wray, K. B. (2011). *Kuhn's evolutionary social epistemology*. Cambridge University Press.