Synthese (2022) 200:171
https://doi.org/10.1007/s11229-022-03650-w

ORIGINAL RESEARCH

Limits to evidential pluralism: multi-method large-N qualitative analysis and the primacy of mechanistic studies

Rosa W. Runhardt[1]

Received: 29 April 2021 / Accepted: 8 March 2022 / Published online: 15 April 2022
© The Author(s) 2022

Abstract

Evidential pluralists, like Federica Russo and Jon Williamson, argue that causal claims should be corroborated by establishing both the existence of a suitable correlation and a suitable mechanism complex. At first glance, this fits well with mixed method research in the social sciences, which often involves a pluralist combination of statistical and mechanistic evidence. However, statistical evidence concerns a population of cases, while mechanistic evidence is found in individual case studies. How should researchers combine such general statistical evidence and specific mechanistic evidence? This article discusses a very recent answer to this question, 'multi-method large-N qualitative analysis' or multi-method LNQA, popular in political science and international relations studies of rare events like democratic transitions and cease-fire agreements. Multi-method LNQA combines a comprehensive study of *all* (or most) relevant event cases with statistical analysis, in an attempt to solve the issues of generalization faced by other types of qualitative research, such as selection bias and lack of representativeness. I will argue that the kind of general causal claim that multi-method LNQA is after, however, is crucially different from the average treatment effect found in statistical analysis and can in fact only be supported with mechanistic evidence. I conclude from this that mixed method research, and thereby evidential pluralism, may be inappropriate in this context.

Keywords Evidential pluralism · Causal mechanisms · Case study research · Large-N qualitative analysis · LNQA · Mixed-methods research · Multi-method research · Causal inference · Social science · Political science · International relations · Generalization, · Qualitative research

T.C.: Evidential Diversity in the Social Sciences

✉ Rosa W. Runhardt
rosa.runhardt@ru.nl

[1] Faculty of Philosophy, Theology and Religious Studies, Radboud University, Erasmusplein 1, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

1 Introduction

Evidential pluralists argue that "in order to establish a causal claim one normally needs to establish the existence of an appropriate conditional correlation and the existence of an appropriate mechanism complex, so when assessing a causal claim one ought to consider relevant association studies and mechanistic studies, where available" (Shan & Williamson, 2021, p. 4). For instance, the evidentially pluralist Russo-Williamson Thesis (Russo & Williamson, 2007) states that causal claims in (bio)medical research ought to be corroborated by both evidence of mechanisms and evidence of difference-making. Evidence of difference-making here could consist of statistically significant relations between a (proxy) variable for the putative cause and putative effect. Evidence of mechanisms, on the other hand, needs to show *how* the putative cause produces the putative effect, i.e. it needs to establish the existence of a suitable mechanism complex (cf. Shan & Williamson, 2021). In biomedical and biological research, mechanisms are commonly conceptualized using Peter Machamer, Lindley Darden and Carl Craver's definition, as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, p. 3), meaning that evidence of mechanisms should consist of working out amongst others what the relevant entities and activities are and how they are interrelated. In recent years, evidential pluralism in the form of the Russo-Williamson Thesis has led to a variety of fruitful discussions in amongst others epidemiology (cf. Canali, 2019), pharmacology (cf. Parkkinen & Williamson, 2020), and evidence-based medicine (cf. Clarke et al., 2014; Russo & Williamson, 2011).

Evidential pluralism and the Russo-Williamson thesis have also been analyzed outside of (bio)medical research, in amongst others labor economics (cf. Claveau, 2012), econometrics (cf. Moneta & Russo, 2014), and political science. Philosopher of political science Sharon Crasnow has pointed out some clear parallels between mixed method research and evidential pluralism, as have I (cf. Crasnow, 2010, 2012; Runhardt, 2021). Yafeng Shan and Jon Williamson have recently applied the original Russo-Williamson thesis to the social sciences in general, arguing that it can be fruitfully applied there, amongst others as a motivation for mixed methods research (Shan & Williamson, 2021). Indeed, in mixed method research the combination of qualitative and quantitative methods often comes down to a Russo-Williamson-like combination of statistical and mechanistic evidence. In these research contexts, mechanisms are conceptualized more broadly than in (bio)medical research, as "the causal pathway, process or intermediate variable by which a causal factor of theoretical interest is thought to affect an outcome" (Gerring, 2008, p. 163).[1]

One important problem for evidential pluralists and mixed method researchers alike is what I will call the *problem of generalization*. Evidence of difference-making is typically general, concerning a population of cases. However, evidence of mechanisms is most often found in individual case studies. Thus, the question is how one may combine general difference-making claims and specific mechanistic claims. Under what circumstances can one fruitfully generalize from single case studies to a whole

[1] For evidential pluralists Shan and Williamson, this difference in conceptualization seems unimportant, as they argue that their version of evidential pluralism "makes no direct claims about the nature of mechanisms" (Shan & Williamson, 2021, p. 5).

population? And when can case study results support conclusions about a general hypothesis implied by statistical reasoning? These questions seem especially important in social scientific research, since individual entities (e.g., civil wars, peace treaties, instances of the acquisition of nuclear powers by states) often have idiosyncrasies that complicate across-case comparisons. In the social sciences, the common intuition is that case study results are highly context-dependent and difficult to generalize (cf. Steel, 2008). The use of case studies to support a general causal claim faces such issues as selection bias and lack of representativeness. If evidential pluralism wants to gain a foothold in social science research, for instance as a fundamental defense of mixed method research, the problem of generalization must therefore be faced head-on.

A very recent answer to the problem of generalization for mixed method research in the social sciences is 'multi-method large-N qualitative analysis' or multi-method LNQA, defended by Goertz and Haggard (Goertz, 2017; Goertz & Haggard, 2022). The multi-method LNQA approach is firmly based on a pluralist view of causal evidence. Multi-method LNQA combines statistical research with LNQA as a stand-alone method, which consists of detailed case study analyses to find the causal mechanisms behind statistical patterns.[2] Multi-method LNQA solves the problem of generalization in a unique way. Unlike other mixed-method approaches, multi-method LNQA distinguishes itself by requiring the researcher to do a case study analysis of *all* (or most) relevant cases in a population, which is only possible when the events under study are rare enough to indeed cover all of them.

In this article, I will argue that multi-method LNQA should be seen as a unique way to solve the problem of generalization outlined above, since it is based on a comprehensive study of all or most cases in the potentially heterogeneous population. However, I will argue that the kind of general causal claim that multi-method LNQA is after, the 'mechanistic generality claim', is crucially different from both philosophy of causation's type-level causal claim and the average treatment effect. For corroborating such a mechanistic generality claim, all the evidential weight is on the individual case studies. The statistical step in multi-method LNQA is redundant, and (in heterogenous contexts) possibly misleading.

An important consequence of this analysis is that mixed method research and evidential pluralism are both misguided in the LNQA context. Specifically, this article calls into question the evidential pluralists' claim that in all but a few instances[3] one needs to establish both the existence of a correlation and the existence of a mechanism complex. By arguing that appropriately performed LNQA makes evidence from association studies redundant, I show that one may assess a causal claim with mechanism

[2] The authors use the term 'multi-method LNQA' to refer specifically to a combination of LNQA as a stand-alone method and statistical approaches. Such a combination of qualitative and quantitative methods is sometimes labelled 'mixed-method' research in the methodology literature. I will keep to the terminological convention in Goertz and Haggard's article here, but I will use the term 'mixed-method research' when I refer to a combination of qualitative and quantitative methods in general. In doing so, I follow Creswell and Plano Clark's definition of mixed method research, according to which both qualitative and quantitative data are gathered and merged with one another to answer research questions and test hypotheses (Creswell & Plano Clark, 2018).

[3] Such cases include some rather uncommon instances of overdetermination, in which a causal relation does not co-occur with a correlation. See Williamson (2019).

studies alone.[4] Moreover, this part of the article goes against the claim that "evidence of correlation is (…) required" in process-tracing studies (Shan & Williamson, 2021, p. 21).

This article is set up as follows. Firstly, I will give a short overview of multi-method LNQA's key methodological assumptions and give a few paradigmatic examples of the method. I then discuss how multi-method LNQA researchers approach the problem of generalization outlined above, by describing the very narrow, 'mechanistic generality' approach to general causal claims which researchers using multi-method LNQA assume. I compare mechanistic generality with essential concepts from philosophy of causation, using Dan Hausman's discussions of the limitations of average treatment effects. I show that the general causal claims multi-method LNQA researchers aim at are considerably different from claims about average treatment effects, and that *only* multi-method LNQA's case study step is reliable for testing 'mechanistic generality'. I conclude with some additional questions regarding the feasibility of mechanistic generality in political science and international relations, meant for further research.

2 What is large-N qualitative analysis (LNQA)?

In this section, I first introduce large-N qualitative analysis as a stand-alone method. Next, I show how LNQA is combined with statistical analysis in multi-method LNQA and provide some examples of the method. I end Sect. 2 with a discussion of the assumptions in multi-method LNQA.

2.1 Large-N qualitative analysis as a stand-alone method

Large-N qualitative analysis is a new research approach in political science and international relations that is especially focused on developing and testing general hypotheses about the causal mechanisms[5] by which a putative cause is thought to affect a certain effect of interest. Methodologist Gary Goertz has shown that LNQA is now widely

[4] François Claveau has recently argued against the use of the Russo-Williamson thesis in the social sciences, arguing that establishing the existence of a mechanism complex only can in some cases be sufficient to establish a causal claim (Claveau, 2012). As such, there is an important parallel between Claveau's work and my own here. One of the reasons why evidential pluralists like Shan and Williamson have rejected Claveau's argument is that the general causal claim under study by Claveau is corroborated by limited mechanistic evidence. Claveau only discusses mechanistic evidence from a small number of all the relevant countries in which the claim is said to hold. "More would be needed to be done to establish a general mechanistic claim that holds more widely across countries: it would need to be shown that the mechanisms are extrapolable to other countries" (Shan & Williamson, 2021, p. 21). Coming back to my own position, even if this is an issue for Claveau's of mechanistic research, it may not be an issue for the much more comprehensive LNQA study of all cases of some rare event. As such, the analysis in this article may be seen as an extension of Claveau's criticism of the Russo-Williamson thesis. I come back to this in the conclusion of the article.

[5] Here, I will use the term 'causal mechanism' in the broad sense mentioned in the introduction, i.e. following Gerring's definition as "the causal pathway, process or intermediate variable by which a causal factor of theoretical interest is thought to affect an outcome" (Gerring, 2008, p. 163). This is admittedly philosophically limited, since I will not discuss the rich and diverse ontological theories of mechanisms to be found in the social science literature (cf. Gerring, 2008; Hedström & Ylikoski, 2010). However, I believe my use in this article best suits the intuitive use of the term by LNQA methodologists, who never

used, in research published in some of the fields' top-ranked journals such as the *American Political Science Review, American Sociological Review,* and *International Organization* (Goertz, 2017). However, LNQA is a very recent development, and in 2017 Goertz remarked that it had not yet been analyzed by methodologists and philosophers: "no author has explicitly defended [LNQA's] causal inference strategy" (Goertz, 2017, p. 208). Goertz has since unpicked the method in more detail with a prominent political scientist who uses LNQA himself, Stephan Haggard (Goertz & Haggard, 2022).

The LNQA researcher starts their work from a hypothesis about the relation between some causal variable X and an effect Y, along with the stipulation of the causal mechanism the researcher believes may link X and Y. A relevant population, in which the researcher believes X and Y are so linked, is defined. Next, the researcher tries to establish 'regularities' in this population of cases. The regularity is expressed as either a "percentage of X followed by Y (X-regularity) or percentage of Y preceded by X (Y-regularity)" (Goertz & Haggard, 2022, p. 6).[6] This, then, is the jumping-off point for the researcher's within-case analysis of the causal mechanism:

> The crucial step in the analysis is within-case causal inference to establish that the postulated causal mechanism is present and operates as expected. (…) In short, LNQA is a combination of regularities AND within-case causal inference of cases contained in the regularity. It is the combination that generates valid causal generalizations (Goertz & Haggard, 2022, p. 5)

Unlike other qualitative studies, in which only a few cases in a population of interest are analyzed, LNQA is revolutionary since it seeks evidence for its causal hypotheses in case studies of all or most of the cases in the population of interest. Such case study research often takes the form of process-tracing (cf. Beach & Pedersen, 2013; Bennett, 2010; Bennett & Checkel, 2015; Collier, 2011; Crasnow, 2017; Hall, 2013), in which a researcher looks for the observable implications of their own postulated mechanism as well as the implications of other mechanisms in the literature, in order to judge which mechanism was actually present in the case. LNQA researchers use process-tracing in each of the cases in the population separately.

Since ideally all cases in the population of interest will be studied and process-tracing is a labor-intensive method, LNQA is mainly used when there are few cases in the population. Thus, LNQA is used amongst others in the study of rare events, where such a comprehensive look at the relevant population is practically feasible. Rare events are events which are only said to occur when a complex set of criteria is met, and which as a result have only occurred a few dozen times in (modern) history. For political science, Goertz and Haggard name democratic transitions, coups, civil wars, and social revolutions. For international relations, they mention wars, acquisition of nuclear weapons, and shifts in hegemonic order. The large-N nature of LNQA contrasts clearly with other qualitative case study approaches, which may focus on only one

Footnote 5 continued
define the term "causal mechanism" but who use it to denote the means through which the putative cause has produced the effect of interest.

[6] Goertz and Haggard's notion of 'regularity' is therefore not Humean, despite the authors' own linking of the term to constant conjunction. For them, a regularity "is simply the share of cases that appear to conform with the generalization" (Goertz & Haggard, 2022, p. 5).

or a handful of individual cases, e.g. only on the ceasefire between Turkey, Greece, and Britain in July of 1974 rather than all ceasefire agreements since the 1940s. If only one or a handful of cases were to be used to support general conclusions, this would bring up the problems of generalization outlined in the introduction, including selection bias and lack of representativeness. So, in short, LNQA is an answer to the problems of generalization by virtue of its comprehensiveness.

Here, it is important to note a limitation to this solution to the problem of generalization. At first glance, we may be tempted to describe the general causal claims aimed at in LNQA as *type causation*, i.e. as causal claims in which the causal relata are generic, not referring to any particular case. Ellery Eells, for example, describes type causation as "a relation between event types, or factors, or properties" (Eells, 1991, p. 6), and distinguishes type causation from token causation, which is "a relation between particular, actually occurring, token events" (Eells, 1991, p. 6). Christopher Hitchcock suggest we can also distinguish the two by describing type causation as "concerned with a full range of possibilities, whereas [token] causation is concerned with how events actually play out in a specific case" (Hitchcock, 2018, Sect. 1.4).

However we construe it, a type-token interpretation of LNQA's general causal claims is problematic. LNQA is aimed at establishing causal claims at the level of an *existing,* predefined population of cases, and as such its causal claims should be seen as unit-population generalizations. Unlike type-token generalizations, LNQA generalizations do not inherently allow for extrapolation *beyond* the population of existing cases to the 'full range of possibilities', i.e. all potential (future) instances and contexts. So, for example, while a LNQA study may corroborate a general causal claim about the set of all cease-fire agreements since 1940, it does not necessarily corroborate causal claims about potential future cease-fire agreements.

While this is indeed a limitation, LNQA methodologists acknowledge it as such. Goertz and Haggard, for instance, claim to be "interested in the more practical question of the empirical scope" (Goertz & Haggard, 2022, p. 9), i.e. the set of actual cases that will be investigated. In preparing the research, some factors or variables are introduced to define this scope. The underlying assumption is that one circumscribes the scope in this way because these factors are what enables the mechanism. In other words, changing any of these variables might invalidate or inhibit the postulated causal mechanism from working.

Goertz and Haggard describe multiple options which may limit the scope of LNQA research, e.g. by "considering tails of distributions, by conceptual engineering, or through (…) choosing samples not on [the putative cause X or effect of interest Y] but by other parameters—such as region—that will of necessity limit cases" (Goertz & Haggard, 2022, p. 14). From this, we may conclude that limiting the scope is at once a practical issue (given that process tracing is so labor intensive) but also a judgement call about the possible enabling conditions of the causal mechanism. I will come back to the matter of scope in Sect. 3. It is worth stressing, however, that the scope as defined by LNQA researchers often does not coincide with what a statistician would normally call the population of cases. The latter refers to *all* the cases in a statistical data set regardless of the values of X and Y, of which the empirical scope of LNQA will be a subset.

2.2 Multi-method LNQA

So far, we have seen that the stand-alone use of LNQA is aimed at establishing general causal claims, based on comprehensive case study research of all the cases in the population. We have also seen that this type of general causal claim is best seen as unit-population generalization, rather than as type-token generalization. In this section, I will show how the stand-alone method of LNQA is combined with statistical research in multi-method LNQA and illustrate the method using several examples from the literature.

Multi-method LNQA combines LNQA case study research with statistical analysis in at least two important ways, thereby making multi-method LNQA part of mixed method research. Firstly, the case study work can be used to falsify a general hypothesis generated by the statistical analysis, as happens when no evidence can be found that there is an underlying pathway connecting cause and effect, despite correlational evidence. This is what Goertz dubs 'large-N qualitative testing' (Goertz, 2017). Secondly, the case study work can be used to *verify* statistical analysis. In the latter, more positive case, multi-method LNQA combines evidence of statistical regularities with evidence of causal mechanisms from the start. This second variant of multi-method LNQA therefore seemingly fits well with evidential pluralist arguments like the Russo-Williamson Thesis.

Like stand-alone LNQA, the positive variant of multi-method LNQA starts from a hypothesized generalization about the causal relation between a causal variable X and effect Y for some population of interest. For the simplest scenario, in which X and Y are either present or absent in any given case (i.e. in which $X = 0$ or $X = 1$ and $Y = 0$ or$Y = 1$), multi-method LNQA proceeds as follows. After a statistical dataset is created for some population of interest, the 'average treatment effect' between X and Y is estimated, i.e. the difference between unit-level outcomes Y in the presence or absence of the cause$X$, averaged over an entire population of interest.[7] Should the average treatment effect between X and Y prove statistically significant, the researcher turns to special subsets of the total statistical population, most often to the cases in which both X and Y are present (i.e. $X = 1$ and$Y = 1$, the (1,1) cell). They then perform process-tracing of the postulated causal mechanism in all these cases.[8]

2.2.1 Examples of multi-method LNQA

Page Fortna's peace time Now that we have seen what multi-method LNQA involves, consider an early example of this research, Page Fortna's study of the relation

[7] For a recent overview of how modern research designs in political science estimate the average treatment effect, as well as these designs' underlying assumptions, see Keele (2015). For more details, see Sect. 3.2.2 below.

[8] While for the (1,1) cell the focus is thus on cases where cause, mechanism, and effect are present, this can be combined with other analyses as well. For example, Goertz argues, one could investigate equifinality and the importance of one's hypothesized mechanism versus alternative causal mechanisms in the literature by also studying cases where the effect is present ($Y = 1$) but the cause and/or mechanism are not ($X = 0$). Furthermore, one could attempt to find cases where the cause and mechanism were present ($X = 1$) but the effect remained absent ($Y = 0$), in order to find out whether the causal mechanism always operates or not. See Tables 2 and 4 in Goertz and Haggard (2022) for further visualization of these options.

between cease-fire agreements and the duration of peace (Fortna, 2004). As Goertz and Haggard point out, this study seems to be the first to use the term LNQA (Goertz & Haggard, 2022, p. 3). Fortna's statistical analysis is based on a duration model, which estimates the effects of the content of cease-fire agreements (independent variables for e.g., whether forces must withdraw, whether demilitarized zones are put in place, whether there is some form of arms control) on the length of peace (the dependent variable). Fortna's statistical analysis and case study research both cover the same set of forty-eight cease-fire agreements and fifteen follow-up agreements (where an earlier agreement was significantly changed) in the period between 1946 and 1998. Fortna argues that this set provides "a comprehensive survey of a population of cases, giving the big picture, indicating general patterns and tendencies, and providing information on whether particular cases are typical or unusual" (Fortna, 2004, p. 41). She welcomes the addition of case study analysis because, she claims, it allows her to include all relevant details about the idiosyncratic cases under study: "[r]educing an issue as complex as why peace lasted or fell apart to a series of numbers for quantitative research entails the loss of much information, information that can be employed in qualitative analysis" (Fortna, 2004, p. 42).

2.2.2 *Stephan Haggard and Robert Kaufman's* Dictators and Democrats

As a second example, consider Stephan Haggard and Robert Kaufman's 2016 book *Dictators and Democrats* (Haggard & Kaufman, 2016), which uses multi-method LNQA to investigate which causal mechanisms played a role in democratic transitions and reversals during the third wave of democracy (which started in the mid-seventies), using a combination of statistical analysis (with a mixed effects logistic regression model) and process-tracing. While Haggard and Kaufman did not use the term LNQA in their book, I include this study here since Goertz and Haggard mention it as an example in their analysis.

Amongst others, in Haggard and Kaufman's study of so-called 'distributive conflict transitions',[9] the authors test whether there is a relation between the level of inequality in a state (measured using the Gini coefficient for income inequality) and the probability of a democratic transition. Haggard and Kaufman perform a logistic regression analysis of the relationship for the entire population, and then use process-tracing in case studies of distributive conflict transitions, tracing their own proposed causal mechanism (density of social organization) and rejecting alternative mechanisms in the literature. In their process-tracing analysis, the authors consider a subset of the entire population: they select on the dependent variable and consider all distributive conflict cases, i.e. all cases where the effect of interest (a distributive conflict transition) is present. They cover 52 cases altogether (Haggard & Kaufman, 2016, p. 103).

I have included both Fortna and Haggard and Kaufman's research projects here since they will prove simple examples of some the overall advantages and disadvantages of multi-method LNQA. However, both research projects, though they combine

[9] A distributive conflict transition is a transition to a democratic system of government in which economically disadvantaged groups have mobilized, demanding redistribution and threatening the continuing rule by the current authoritarian incumbents, and in which incumbents are ousted or forced to concede to democratic elections.

statistical analysis with a purportedly comprehensive case study analysis, may not be the most representative example of current best practice in multi-method LNQA. The main reason these projects are not as representative is that the work covers over fifty cases, arguably too many to conduct thorough enough process-tracing analysis. On the other hand, as Goertz and Haggard point out, this also depends on the level of complexity of the causal mechanism involved; Haggard and Kaufman's hypothesized causal mechanism is "relatively spare" (Goertz & Haggard, 2022, p. 15). I will come back to this issue in the conclusion.

2.2.3 *Dale Copeland's* economic interdependence and war

Dale Copeland's work *Economic Interdependence and War* (2015) is a recent example of multi-method large-N qualitative analysis with a lower number of cases. Copeland studies the relationship between economic interdependence between states (e.g., in the form of trade) and the probability of military conflict (such as war) between those same states. Copeland's own theory, 'trade expectations theory', argues in short that the expectations a state has of potential future trade with other states play a key role in linking economic interdependence and conflict between those states. Research on the relation between trade and conflict has traditionally been purely quantitative. Copeland, however, focuses on in-depth case study analysis to supplement earlier statistical results because, he argues, "causal mechanisms that lead to peace or war will be inadequately understood if [quantitative methods] [are] our sole or primary methodology, given that quantitative methods are inherently about correlations and associations between variables rather than causality per se." (Copeland, 2015, p. 51).

Copeland provides case studies of all serious great power conflict since 1790, including war and crises that made war more likely but eventually dissolved. This comes to a total of forty cases. Using these forty case studies, Copeland aims to investigate the salience of the trade expectations theory's mechanisms versus the salience of mechanisms posed by competing theories. Copeland concludes that in thirty of the forty total cases, economic interdependence played a role ($X = 1$) and that out of those cases, a further twenty-six show evidence of the mechanisms postulated in his trade expectations theory. This evidence consists of causal process observations, including observations of the extent to which trade expectations were part of political leaders' deliberations during periods of conflict or near-conflict.

2.3 Assumptions in LNQA

Now that we have seen both a theoretical description of the method and several examples of the multi-method LNQA research, I will finish this section by turning to two important assumptions which all examples described above (Fortna, Haggard and Kaufman, and Copeland) arguably share: the conceptual monism assumption and the epistemic reliability assumption.

2.3.1 Conceptual monism

Firstly, all researchers discussed above implicitly assume that the evidence of statistical regularities and the evidence of causal mechanisms in case studies both support conclusions about the same causal concept.[10] Sharon Crasnow has made clear that this *conceptual monism assumption* underlies mixed method research in political science more generally. Mixed method researchers believe that combining different methods is valuable since this may serve as a kind of triangulation[11]:

> We might generate statistical evidence of causal connection through multiple regression analysis, a core statistical technique in political science research, but a case study that traced a causal process or identified a causal mechanism would make us more confident that we had indeed established a cause. (…) The idea is that the statistical work and the case study research are methods that support the conclusion about *the same causal connections* (Crasnow, 2010, p. 37, emphasis added)

For example, Haggard and Kaufman in their study of distributive conflict transitions assume that their logistical regression and subsequent process tracing case studies both test the same general causal connection, viz. that between inequality and democratization in (all) distributive conflict cases.

Copeland's case is admittedly less clear-cut. Copeland argues that correlations and associations are not about "causality per se" (Copeland, 2015, p. 51). However, he does intend to *test* previously hypothesized correlations and associations between economic interdependence and conflict, including those postulated by liberal and realist theory. As evidence of this, in later work Copeland has argued that "quantitative research (…) provide[s] a useful 'first cut' test of the possible explanatory value of trade expectations theory" (Copeland, 2017b, p. 34). Therefore, I believe that Copeland's approach reasonably falls under conceptual monism as well.[12]

Conceptual monism is opposed to 'causal pluralism' (sometimes 'conceptual pluralism'), that is, the view that "every evidential method define[s] its own concept (…) [and that] when moving from method to method we would in fact change the hypoth-

[10] There is a subtle question worth exploring here, namely to what extent this assumption is metaphysical versus epistemic for LNQA researchers and methodologists. Williamson has described his view of causation as purely epistemic, and in his more recent work describes evidential pluralism as "a thesis about establishing and assessing causality, not an analysis of the concept of cause nor a claim about the metaphysical nature of causality" (Shan & Williamson, 2021, pp. 4–5). Further analysis of this question is beyond the scope of this paper.

[11] For completeness, it is worth noting two things. Firstly, triangulation is only one of several possible motivations for mixed methods research in political science. For an overview of some of the alternative literature, see Brookes (2017). Alternative approaches are beyond the scope of this article, as they are more dissimilar to multi-method LNQA. Secondly, LNQA is far from the only mixed-method approach that aims at triangulation. For example, 'Causal-Oriented Mixed-Methods Research' or CMMR "uses large-N analysis to establish a robust relationship between [putative cause] X and [effect of interest] Y and detailed process-tracing case studies to probe the X/Y relationship in specific settings" (Barnes & Weller, 2017, p. 1019). As already highlighted, however, unlike CMMR multi-method LNQA coves *all* relevant cases within some scope of interest.

[12] For a helpful situating of Copeland's assumptions about causation in the wider international relations literature, see also Büthe (2017).

esis to be tested" (Reiss, 2009, p. 28). Crasnow has argued against this interpretation of mixed method research, showing that in many political science examples, diverse evidence corroborates the same causal claim in mutually supportive ways (Crasnow, 2010). The role of case study analysis in mixed method research, Crasnow argues, is to provide contextual details which the statistical results lack, i.e. "case study research methodology develops attention to and respect for specific circumstances and thus an awareness of relevant differences (the extent to which a general theory may not pertain) as well as relevant similarities (the extent to which it does)" (Crasnow, 2010, p. 47).[13]

In what follows, I will criticize the conceptual monism assumption, finding that the relevant differences and similarities found by process tracing will, in many instances, support a conclusion about a *different* type of causal claim than the statistical evidence, namely a 'mechanistic generality claim'. This consideration is not a part of Crasnow's defense of causal monism for mixed method research. I will come back to the feasibility of this first assumption in Sect. 3.

2.3.2 Epistemic reliability

The second key assumption in multi-method LNQA is that the pieces of evidence gathered of statistical regularities and of mechanisms can both reasonably reliably give the researcher knowledge about causal relations *in the evidential context of the study*.[14] In other words, they must assume there are no factors which complicate the use of either process-tracing evidence or statistical evidence for testing general causal claims.

For example, Haggard and Kaufman assume that a mixed effects logistic regression model is suitable for examining the general causal connection between inequality and democratization across different countries (see Haggard & Kaufman, 2016, pp. 69–73 and 90–95 for some of the relevant considerations). Moreover, they assume that the causal process observations they conduct are also informative of this causal connection. For example, they assume that the process-tracing they perform on the historical events that occurred in Argentina between the military's seize of power in 1976 to the organization of democratic elections in 1983, provides reliable evidence for the causal connection between inequality and democratization as well.

[13] Goertz and Haggard state explicitly that they "do not take a position on causal pluralism" (Goertz & Haggard, 2022, p. 7). However, given that they argue for a combination of "regularity and mechanistic approaches" (Goertz & Haggard, 2022, p. 8) (for LNQA) and "statistical as well as regularity and mechanism approaches" (Goertz & Haggard, 2022, p. 8) (for multi-method LNQA), we may argue that they must be committed to causal monism: the statistical work, regularities, and mechanisms must all support a conclusion about the same causal connection. Note, however, that since regularities in and of themselves are only a report of the percentage of cases which fit a certain pattern, arguably they are not linked to any causal claim. I will therefore limit myself to evaluating the feasibility of causal monism behind combining statistical and mechanistic evidence.

[14] To assume that both methods are reliable in the same evidential context is not obvious. As Shan (2022) discusses, assumptions underlying mixed methods research vary wildly depending on one's philosophical position. For example, pragmatists argue that researchers are free to choose whichever methods are the best fit for their research aims and context. Yet this position is arguably compatible with the view that statistical methods and qualitative (mechanism-based) methods are both valuable but for *different* aims and evidential contexts.

As with the first assumption, I will argue below that checking whether this reliability assumption holds requires a further study of what, exactly, is meant by a general causal claim in political science and international relations. This deserves careful attention, because individual entities in LNQA (viz., the different cases in the particular population under study) may be markedly different from one another. This complicates comparisons and thereby leads back to the problem of generalization that LNQA researchers aimed to solve.

In the next section, I will describe and analyze the narrow 'mechanistic generality' approach to general causal claims that researchers using LNQA assume. I will compare this definition with a parallel discussion in philosophy of causation, Dan Hausman's distinction between average treatment effects and other types of general causal claims. Using this theoretical analysis, I will then argue that in fact *only* standalone LNQA is reliable for testing general causal claims, given the narrow mechanistic definition of generality employed by researchers using multi-method LNQA. I will thereby cast doubt on both the conceptual monism assumption and the epistemic reliability assumption.

3 General causal claims in political science and international relations

So far, I have shown that multi-method LNQA is a unique way to solve the problem of generalization, based on both a comprehensive process tracing study of all cases within a (usually heterogeneous) empirical scope of interest and a statistical analysis of the wider population. I have argued that multi-method LNQA methodologists and researchers assume that the statistical analysis and case study analysis both provide evidence of the same underlying general causal claim. In this section, I will link multi-method LNQA researchers and methodologists' interpretation of what 'general' means with theoretical concepts from philosophy of causation, following Dan Hausman's analysis of type-level causation and average treatment effects. I conclude that LNQA 'mechanistic generality' claims, found at the level of individual case studies, provide much richer information than statistical claims. Therefore, if and when the LNQA research step has been performed thoroughly and convincingly, the statistical research step is superfluous at best.

3.1 Mechanistic generality

We have already seen some evidence in Sect. 2 that type causation does not neatly fit with the type of generality that LNQA methodologists like Goertz and Haggard have in mind. Goertz and Haggard state that the generalizations they aim at contain two important components: a "hypothesized causal regularity between X and Y and a proposed causal mechanism" (Goertz & Haggard, 2022, p. 10). The hypothesized causal regularities stem, as stated in Sect. 2.2, from an observed statistically significant average treatment effect within some population. Regularities, however, are only part of the story; the second component, the causal mechanism, is more crucial to the author's notion of generalization. Goertz and Haggard argue that "the qualitative component

of the method is clearly its labor-intensive core" (Goertz & Haggard, 2022, p. 16) and state that a significant average treatment effect in itself "does not establish a causal relationship (…). The causal generalization is established through the within-case causal inference, which in turn is organized around a number of further casual mechanism claims linking [the mechanism] to [the effect]." (Goertz & Haggard, 2022, p. 19). The motivation for mechanistic generalization, rather than an average treatment effect, seems to be that the multi-method LNQA proponents believe evidence of mechanisms is necessary for causal inference in the social sciences, since as Goertz argues "[i]f one cannot produce convincing case studies showing the causal mechanism in action, it is hard to find the statistical analyses convincing at all" (Goertz, 2017, p. 215). So, in short, Goertz and Haggard's general causal claims seem to be a kind of 'summing up' of all the causal mechanism claims connecting X and Y in the individual case studies. To them, making a general causal claim means that the same causal mechanism is behind the causal relation between X and Y in all the cases in the population where $X = 1$ and $Y = 1$.[15] Let us call this *mechanistic generality*.

Crucially, the multi-method LNQA researchers introduced in Sect. 2.2 (Fortna, Haggard and Kaufman, and Copeland) also consider the causal mechanisms they uncover to be essentially different from statistical constructs like interaction terms. Copeland, for example, admits that statistical research into the relation between economic interdependence and conflict has become increasingly sophisticated. The inclusion of interaction terms in later quantitative models meant that these models were able to specify particular properties of states (e.g., domestic variables) which helped or hindered economic interdependence's impact on the chances for peace. However, Copeland argues that these quantitative models, while more complex, are not informative enough by themselves: "quantitative findings, in and of themselves, are merely suggestive correlations; they cannot tell us anything directly about the causal mechanisms underlying the correlations." (Copeland, 2015, p. 60) Copeland is not interested in such correlations alone, and argues that:

> Our common scholarly goal must be this: to discover a plausible interpretation that covers as many of the findings as possible. In short, which of the causal explanations makes the most sense of all the diverse quantitative evidence? (Copeland, 2015, p. 60)

So, multi-method LNQA researchers require case study research precisely because they consider statistical research alone (even when it includes sophisticated tools like interaction terms) inadequate evidence for mechanistic generality. However, the emphasis on mechanistic generality speaks to more than just the weakness of correlational, average effect results in supporting causal claims. In the next section, I will argue that mechanistic generality and the average treatment effect are fundamentally different concepts.

[15] Note, here, that LNQA does not require that this subset of cases is otherwise homogeneous. There may exist various heterogeneous variables within the subset, as long as these variables do not influence the causal mechanism and so are in a sense irrelevant to the analysis. I will come back to this in Sect. 3.2.1.

3.2 Mechanistic generality versus average treatment effects

So far, I have argued that in LNQA, researchers do not aim for an average treatment effect claim. Rather, LNQA researchers combine information of each specific case in the empirical scope of interest into a 'mechanistic generality' claim, i.e. they claim that the same causal mechanism is behind the causal relation between X and Y in all the cases that fall under certain scope conditions (e.g., all instances of some rare event). This distinction, between average effects and mechanistic generality, has some support in the theoretical literature in the form of Dan Hausman's philosophical analysis of causal generalizations in the special sciences (Hausman, 2010). In this section, I will first make the theoretical, philosophical case for the distinction between average effects and mechanistic generality. Then, I will use the distinction to stress that statistical evidence is superfluous at best in supporting mechanistic generality. By drawing this conclusion, I strongly question the evidential pluralist position that one must demonstrate the existence of both a correlation and a mechanism complex in order to establish a causal claim. I cast doubt on both the conceptual monism assumption and epistemic reliability assumption from Sect. 2.

To further analyze the distinction between average treatment effects and mechanistic generality, first note that many putative causes X in political science and international relations are not sufficient for the effect of interest Y. The putative cause can only lead to the effect of interest if the circumstances are right, viz., if some set of required background conditions are present. This is why, in the set-up of an LNQA study, researchers must describe the factors they are using to define the empirical scope. In more complex scenarios, X may be part of a set of INUS conditions, where the putative cause X is an insufficient, but necessary part of an unnecessary but sufficient condition for the effect of interest Y.[16] In LNQA, the conditions that defined the empirical scope may be other elements in the set of INUS conditions besides X itself. Problematically, which background factors play a role in the $X - Y$ relation may be unknown. In other words, we may not be aware of all sets of background conditions under which X will be a cause of Y.

3.2.1 The contextual unanimity condition

Dan Hausman (Hausman, 2010) evaluates solutions to the above problem of unknown background conditions from the philosophical literature. In the philosophy of causation for the special sciences, he argues, the most common response to such complexity has been to introduce a theory of probabilistic causality. In this probabilistic theory the problem of unknown background conditions is fixed by maintaining that "C is a positive cause of E (in some population P) if and only if C increases the probability of E in *every* causally homogeneous background circumstance in P" (Hausman, 2010,

[16] In some scenarios, there is no equifinality (no other possible causes that may bring about Y). Therefore, in these simple scenarios we are not dealing with INUS conditions. Arguably, this is the case in what Goertz and Haggard call 'Y regularities', which take the form of "if $Y = 1$ then $X = 1$" (making X a necessary condition for Y). However, the background conditions will still need to be just right for X to affect Y and so most of the considerations that will follow below still apply. Thanks to an anonymous reviewer for pointing out this subtlety.

p. 53). Hausman points out two key reasons for imposing this 'contextual unanimity condition' relevant for the argument in this article. Firstly, the condition "is the easiest way to avoid relativizing causation to particular contexts" (Hausman, 2010, p. 53), since under this definition there are no contexts in which C does not increase the probability of E. Secondly, this condition is "an attempt to *evade* the irregularity of causal generalizations" (Hausman, 2010, p. 54), given that the causally homogeneous background contexts are often not known, and it may not be possible to judge which of the different possible background contexts apply to a particular individual or case.

To give an example, for Fortna the contextual unanimity condition would imply that the probability of a long peace should be increased by the instatement of a demilitarized zone in *all* causally homogeneous background conditions: in any situation with the same level of monitoring by a third party, arms control measures, confidence-building measures, etc. This, the proponents of the contextual unanimity condition would maintain, helps us avoid relativizing the causal claim that demilitarized zones increase the probability of long peace to background conditions. Moreover, we would evade having to specify which homogeneous background condition any given cease-fire agreement is a part of, glossing over such individual idiosyncrasies.

However, as we clearly see in the Fortna example, and as Hausman himself argues, the contextual unanimity condition is likely too strong for the special sciences:

> C can be a cause of E even though its bearing on E differs in different causally homogeneous circumstances. To interpret the causal generalization 'C causes E in population P' as maintaining that C increases the probability of E in every homogeneous circumstance in this population implies that causal generalizations are almost all false or else have such narrow or unclear scope as to be useless (Hausman, 2010, p. 55).

For example, if Fortna were to make the causal generalization that 'demilitarized zones cause a longer duration of peace in the ceasefire agreements drawn up since 1946', we should not take this to mean that demilitarized zones have this effect on peace in all the ceasefire agreements drawn up. In some cases, for example, other putative causes may trump any effect of a ceasefire agreement.

Similarly, Copeland argues that trade expectations are causally related to conflict in the population of forty cases he has studied, but argues that "trade expectations may be playing an important causal role within different [INUS] bundles that lead to war or to the ending of cold war, even if [trade expectations are] not implicated in all [INUS] bundles that do so" (Copeland, 2017a, p. 49). In both the Fortna and Copeland example, insisting on the contextual unanimity condition would lose sight of the other factors that are causally relevant. In general, there will most likely be a great deal of variation on causal variables within a given population under study in multi-method LNQA, making the contextual unanimity condition a poor solution.

3.2.2 The average effect theory

Hausman himself supports a different solution to the problem of unknown background conditions, the 'average effect' theory.[17] In this theory, X is a cause of Y iff, when one holds fixed the frequencies of all the other background factors relevant to Y (apart from X and its effects) at their frequency in population P, there is a significant difference in average outcomes Y between cases where $X = 0$ and cases where $X = 1$. Hausman's average effect theory is a relevant alternative to the contextual unanimity condition since the average effect can be calculated without strong homogeneity assumptions. Modern statistical research designs based around the average treatment effect do not require information about each individual in the population(cf. Keele, 2015). Rather, the average effect serves as guidance, Hausman argues, which "is needed when the details concerning the contexts are not known" (Hausman, 2010, p. 57). He continues that the average effect theory can provide advice in particular cases even when researchers do not know which homogeneous context is relevant to that case: "one has to generalize across contexts in which the effects of causal factors are not uniform" (Hausman, 2010, p. 57). Such generalizations are therefore practical tools for Hausman, and not the endpoint of research.

The average effect theory, then, is intended as a practical solution to the problem of unknown background conditions. Hausman shows clearly that the average effect theory is *not* useful for causal generalization when sufficient information on the background conditions of each specific case in the population is available. "Of course, if one knew what the causally homogeneous circumstances were, the role of the causal factor in each of those circumstances, and which circumstances individuals were in, then there wouldn't be any need to do any averaging." (Hausman, 2010, p. 55) In those cases, constructing the average treatment effect using statistical inferences or randomized experiments would mean 'losing information' from case studies.[18]

The limitations which Hausman outlines for his average effect theory are equally applicable to the average effect condition in multi-method LNQA. However, it is worth noting that Hausman's discussion of average effect theory is not a methodological, statistical approach, but rather a philosophical argument about the advantages and disadvantages of the probabilistic theory of causality for the special sciences. Hausman's work is an answer to the question of how one may best support a causal claim in the special sciences and as such, his discussion of the concrete methods for estimating the average treatment effect is limited.[19] However, for purposes of this essay, Hausman's

[17] See also the work of John Dupré (Dupré, 1984).

[18] This is akin to Fortna's claim, cited in Sect. 2.2.1, that "[r]educing an issue as complex as why peace lasted or fell apart to a series of numbers for quantitative research entails the loss of much information, information that can be employed in qualitative analysis" (Fortna, 2004, p. 42).

[19] A reader more concerned with concrete methods can draw on an extensive literature in statistical methodology (cf. Imbens, 2004; Keele, 2015; Morgan & Winship, 2015). Of particular interest is this literature's approach to 'identification assumptions', i.e. the assumptions under which a statistical estimate like the ATE can be given a causal interpretation (Keele, 2015, p. 314). This literature, too, concerns itself with heterogeneity of causal relationships. Morgan and Winship, for example, point out that "quantities such as the ATE should not be assumed to be equal to the individual-level causal effect for any individual [case]. (…) [W]hen individual-level heterogeneity of causal effects is present, individual-level causal effects (…) will not all be equal to the average of these individual level effects." (Morgan & Winship, 2015, p. 47).

global conclusion is both defensible and relevant: the average effect is not a useful basis for making a general causal claim if the circumstances of each specific individual in the population are known.

Hausman's analysis supports my earlier argument that the average treatment effect is not the kind of generalizability which the multi-method LNQA researchers and methodologists under discussion in this article are after. In LNQA, information of each specific case *is* available. After all, the revolutionary contribution of the LNQA method is to check all occurrences of the (rare) events under study. In those cases, Hausman would urge us to forget about the average effect interpretation of causal generalizations, as indeed Goertz and Haggard prescribe. In other words, if the qualitative case study analysis step in multi-method LNQA indeed provides strong evidence that X causes Y via a certain causal mechanism in all case studies where $X = 1$ and $Y = 1$, then a population level average treatment effect would lose causal information. All the evidential weight is on the individual case studies. It is only when we do not have access to detailed knowledge of the individual cases that we need average effects.

3.2.3 Moving beyond average effects

So far, we have seen that while there is a need for average effect calculations in cases of unknown background conditions, such as cases where the population of cases is simply too large to find out in which circumstances each individual case is in, arguably once we have conducted a thorough large-N qualitative analysis, average effect calculations are extraneous. In the multi-method LNQA research of Fortna, Haggard and Kaufman, and Copeland, the researchers actively sought information of each specific case. After all, the point was to thoroughly investigate all the case studies within the empirical scope of interest. In those cases, I take it that Hausman would urge us to forget about the average effect interpretation of causal generalizations. It is only when we do *not* have access to detailed knowledge of the individual cases that we need average effects. In sum, if LNQA results in causal knowledge about all case studies within the scope of interest, this then makes the statistical results combined with LNQA results in the multi-method approach superfluous at best. As a consequence, Shan and Williamson's belief (mentioned in the introduction) that evidence of correlation is somehow required in process-tracing appears false in this particular context. While it is the case that evidence of correlation is *established* by multi-method LNQA researchers and can inform the scope conditions for subsequent within-case analysis, this evidence is redundant once the process-tracing step is completed.

To stress this point further, consider an example by Hausman of the negative effects of using an average treatment effect when more information is available:

> Generalizations about the average effects may (…) be badly misleading. For example, if, contrary to fact, smoking lowered the risk of lung cancer in women but increased it sufficiently among men that, on average, the risks over the whole

Footnote 19 continued

The statistical methodology literature also includes valuable discussions of how one may define a population for statistical analysis, described as the choice for a particular 'population model' (cf. Morgan & Winship, 2015, pp. 74–76). To what extent the particular population model chosen affects the discussion in Hausman is beyond the scope of this article, but a relevant area for further (philosophical) research.

> population were higher among smokers than non-smokers, then, on the average effect view, smoking causes lung cancer. But women seeking to avoid lung cancer would of course be ill-advised to quit smoking. If this difference in the average effect among men and among women were known, then the population-wide generalization would be irrelevant. If the difference in the average effect of smoking among men and among women were not known, then the surgeon general's announcement that smoking causes lung cancer, though true, would lead women to make bad choices (Hausman, 2011, p. 230)

An equivalent scenario can be constructed using the Haggard and Kaufman study of democratization. In this study, the average treatment effect is used to identify whether medium inequality (rather than low or high inequality), as measured by the Gini coefficient, is a cause of democratization. There are other factors related to democratization; at best, medium inequality is part of an INUS condition. If we did not know which other factors in the population of countries were relevant, then the average effect could be informative for policy purposes. However, the effect may be misleading about individual countries, e.g., when we try to encourage democratization in a certain state without knowing in detail which of the homogeneous background conditions apply to this state. Were we in a situation (as LNQA intends) where the details of each member of the population *are* known, then the average effect is not what political scientists and international relations scholars (or policymakers) should be after. Focusing on the average effect when individual mechanisms are known would be, as Fortna puts it, 'throwing away information', or even, as Hausman says, 'misleading'.

3.2.4 The usefulness of statistical analysis

Before concluding this article, I wish to preempt one potential point of criticism. A critic may ask whether my analysis is not too harsh about the usefulness of statistical analysis. I have so far argued that statistical analysis cannot be reliably used to test mechanistic generality claims, since statistical analysis is focused on the identification of average effects instead. However, the critic may continue, is there not a clear role for statistical analysis yet in *generating* useful hypotheses that the case study step of multi-method LNQA can then test? Looking at the development of multi-method LNQA, we can see that statistical analysis has indeed been used for this in the past. As said in the introduction, there are two ways of combining statistical analysis with LNQA. The first, 'large-N qualitative testing', is the use of LNQA to *falsify* a general hypothesis generated by statistical analysis (cf. Goertz, 2017). In large-N qualitative testing, process tracing adds further information about individual cases; it looks into the potential causal mechanisms behind a statistical correlation. In doing so the average effect and mechanistic generality claim come apart.

I will grant that statistical analysis has been useful in pointing out 'where to look', i.e. in informing how LNQA researchers may set the scope conditions for subsequent process tracing analysis. However, this is arguably different from considering statistical analysis as, in Copeland's words in Sect. 2.4.1 above, 'a useful first cut test'. Specifically, using statistical analysis as a hypothesis-generating exercise does not require the assumption that an average treatment effect claim and a mechanistic gen-

erality claim support a conclusion about the same kind of general causal connections. In the words of Sect. 2.3.1, we do not require the conceptual monism assumption.

The second way one may combine statistical analysis with LNQA is by using case study work to *verify* some causal conclusion drawn from statistical analysis.[20] Here, the aim is to combine evidence of statistical regularities with evidence of causal mechanisms. Yet given that statistical research and process tracing establish different concepts (an average treatment effect versus mechanistic generality), this way of approaching multi-method LNQA is hard to defend. The statistical analysis may have been useful in pointing out relevant areas of research, but that does not mean that we may assume statistical analysis has any reliability for testing mechanistic generality or even (as the Russo-Williamson thesis argues) that one must typically consider statistical analysis when assessing a causal claim.

4 Conclusion

I started this article by noting that evidence of mechanisms is found at the level of individual case studies, while difference-making claims are general. We have now seen that, in fact, the type of general causal claim implied by difference-making (the average treatment effect) is not what multi-method LNQA researchers are after. Instead, they seek mechanistic generality, i.e. they require that the same causal mechanism connects putative cause X and effect of interest Y in all or most of the cases in some relevant population. So far, I have given clear reasons to doubt the use of statistical research for testing mechanistic generality, based on the argument that average treatment effects and mechanistic generality do not support a conclusion about the same kind of general causal claim.

As such, we may conclude that evidential pluralism as defended by philosophers like Russo, Shan, and Williamson and as followed by multi-method LNQA methodologists is flawed. Evidential pluralism only works when the evidence we collect of mechanisms and difference-making speaks towards the same causal concept (i.e. conceptual monism holds) and each method is minimally epistemically reliable. However, we have seen examples here in which these assumptions are untenable. I predict that the conclusions in this article have further consequences for political science and international relations, given these disciplines' often deal with an evidential context of a highly heterogeneous population with relevant background factors, in which causes are often not sufficient for their effects. In such a context, the average treatment effect does not guarantee anything about individual case studies, nor does it support a conclusion about mechanistic generality.

In this article, I have focused on the discrepancies between statistical claims and mechanistic generality. However, I have not evaluated the epistemic reliability of *case study research* for testing mechanistic generality. My argument so far has been that *if* the qualitative analysis stage of multi-method LNQA is epistemically reliable, i.e. results in causal knowledge about all case studies within the empirical scope, the

[20] This approach is thereby arguably similar to an explanatory-sequential approach in general mixed methods research; cf. Edmonds and Kelly (2017).

statistical step of multi-method LNQA is superfluous. However, I have glossed over how likely it is that this antecedent is indeed true, i.e. how likely it is that the qualitative analysis stage is reliable. As I discussed in Sect. 2.2, we may worry that some LNQA projects are less likely to be valid, given how many cases they attempt to cover and the level of complexity of the causal mechanism involved. The question of how the number of case studies and complexity of the mechanism negatively impact the validity of stand alone LNQA is beyond the scope of this article, but undoubtedly an important topic for further research. Further important questions will include: do we have good reasons to believe that individual cases of some rare event (e.g. those in the (1,1) cell) have enough in common with one another to say that the same causal mechanism is present in all or most of the cases, as mechanistic generality requires? How do the known background factors, which will likely differ between case studies, interact with a mechanism? Can these factors thwart mechanistic generalization?

Note that these issues require that we look beyond the epistemic reliability of process-tracing for singular causal claims alone. The use of process-tracing evidence for corroborating singular causal claims has already been widely discussed by methodologists and philosophers of social science alike (such as Bennett & Checkel, 2015; Crasnow, 2017; Hall, 2013; Jacobs, 2016), who generally agree that process tracing *is* reliable. However, the case study step of multi-method LNQA does not only require the reliability of single case study analysis for singular causal claims, but also the reliability of combinations of single case studies for general causal claims. In other words, we may still ask if and under what circumstances one can fruitfully support mechanistic generality with case study evidence.

There is one area where we can see more easily how case study research can support a mechanistic generality claim: single case study evidence can provide negative evidence towards a mechanistically general hypothesis. In the simplest scenario, when no evidence of a causal mechanism under study is found in any of the case studies, this will speak strongly against mechanistic generality. It makes clear that the same causal mechanism does not lead to the same outcome in all or even most cases. We may say, in other words, that process tracing evidence is "negatively reliable" for mechanistic generality (cf. Runhardt, 2021).

To illustrate this negative reliability of case study research, consider Goertz's example of the statistical hypothesis by Mansfield and Snyder (2005), who found statistical evidence that states which are transitioning to a mature democracy (e.g., from an autocracy) while having weak institutions are more likely to fight wars with democracies than states which are *not* transitioning (such as stable autocracies). This statistical generalization, an average treatment effect, was qualitatively tested by Narang and Nelson (2009), who investigated all cases of states democratizing with weak institutions and which fought international war. Narang and Nelson found only six such cases, all of which took place before the First World War, and concluded that "the Mansfield and Snyder causal mechanism has limited scope and is not very general" (Goertz, 2017, p. 197). This speaks strongly against the mechanistic generality of Mansfield and Snyder's work but does not call into question the average treatment effect they found.

While the negative reliability of process tracing for mechanistic generality seems more straightforward, its 'positive reliability' is less clear. Under what circumstances

will individual cases within some empirical scope of interest support mechanistic generality? Typically, philosophers have approached this question as an issue of external validity. Philosophers like Francesco Guala and Daniel Steel (cf. Guala, 2010; Steel, 2008) ask under what circumstances causal relations found in a test population can be extrapolated to a target population. This is especially problematic when it is unclear whether test and target population have sufficiently similar causal structures to make comparison feasible. In other words, this literature is mainly concerned with cases where (in Hausman's terms from Sect. 3), information about causally homogeneous backgrounds is missing. Yet this is arguably not an issue in the multi-method LNQA work discussed in this article. As said above, we do not yet have a framework for how *known* background factors, which will likely differ between case studies, interact with a mechanism and potentially thwart mechanistic generalization. Therefore, further research must be done into the comparability of mechanisms when such information is available.[21]

Acknowledgements I would like to thank the participants at the 'Evidential Pluralism and the Social Sciences' conference held at the University of Kent on 16 -17 July 2020 as well as Stephan Haggard for fruitful discussions on the topic of this article. Moreover, I would like to thank the two anonymous reviewers for Synthese. Any remaining errors are my own.

Funding The author received no financial support for the research, authorship, and/or publication of this article.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author has no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Barnes, J., & Weller, N. (2017). Case studies and analytic transparency in causal-oriented mixed-methods research. *Political Science and Politics, 50*(4), 1019–1022.

[21] A study of the comparability of mechanisms likely requires a deeper study of the ontology of mechanisms assumed by individual authors which goes beyond the assumptions in this article (see footnote 1). Derek Beach goes some way to discussing the relation between ontology and comparability in Beach (2022).

Beach, D. (2022). Evidential pluralism and evidence of mechanisms in the social sciences. *Synthese,* 199, 8899–8919.
Beach, D., & Pedersen, R. B. (2013). *Process-tracing methods: Foundations and guidelines*. University of Michigan Press.
Bennett, A. (2010). Process tracing and causal inference. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 207–220.
Bennett, A., & Checkel, J. T. (2015). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.
Brookes, M. (2017). The road less travelled: An agenda for mixed-methods research. *PS: Political Science and Politics, 50*(4), 1015–1018.
Büthe, T. (2017). Introduction to the symposium. *Qualitative and Multi-Method Research: Newsletter of the American Political Science Association's QMMR Section, 15*(2), 29–33.
Canali, S. (2019). Evaluating evidential pluralism in epidemiology: Mechanistic evidence in exposome research. *History and Philosophy of the Life Sciences*, *41*(4).
Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339–360.
Claveau, F. (2012). The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 806–813.
Collier, D. (2011). Understanding process tracing. *PS: Political Science and Politics, 44*(4), 823–830.
Copeland, D. C. (2015). *Economic interdependence and war*. Princeton University Press.
Copeland, D. C. (2017a). Rare events and mixed-methods research: Shaping the agenda for the future. *Qualitative and Multi-Method Research: Newsletter of the American Political Science Association's QMMR Section, 15*(2), 48–57.
Copeland, D. C. (2017b). The central methodological claims and contributions of economic interdependence and war. *Qualitative and Multi-Method Research: Newsletter of the American Political Science Association's QMMR Section, 15*(2), 33–35.
Crasnow, S. (2010). Evidence for use: Causal pluralism and the role of case studies in political science research. *Philosophy of the Social Sciences, 41*(1), 26–49.
Crasnow, S. (2012). The role of case study research in political science: Evidence for causal claims. *Philosophy of Science, 79*(5), 655–666.
Crasnow, S. (2017). Process tracing in political science: What's the story? *Studies in History and Philosophy of Science Part A, 62*, 6–13.
Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research*. Sage.
Dupré, J. (1984). Probabilistic causality emancipated. *Midwest Studies in Philosophy, 9*(1), 169–175.
Edmonds, W. A., & Kennedy, T. D. (2017). *An applied guide to research designs: Quantitative, qualitative, and mixed methods* (2nd ed.). Sage.
Eells, E. (1991). *Probabilistic causality*. Cambridge University Press.
Fortna, V. P. (2004). *Peace time: Cease-fire agreements and the durability of peace*. Princeton University Press.
Gerring, J. (2008). Review article: The mechanismic worldview: Thinking inside the box. *British Journal of Political Science, 38*(1), 167–179.
Goertz, G. (2017). *Multimethod research, causal mechanisms, and case studies: An integrated approach*. Princeton University Press.
Goertz, G., & Haggard, S. (forthcoming). Large-N qualitative analysis (LNQA): External validity and generalization in case study and multi-method research. In H. Kincaid & J. Van Bouwel (Eds.), *The Oxford handbook on the philosophy of political science*. Oxford University Press.
Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science, 77*(5), 1070–1082.
Haggard, S., & Kaufman, R. R. (2016). *Dictators and democrats: Masses, elites, and regime change*. Princeton University Press.
Hall, P. A. (2013). Symposium: Tracing the progress of process tracing. *European Political Science, 12*, 20–30.
Hausman, D. M. (2010). Probabilistic causality and causal generalizations. In E. Eells & J. H. Fetzer (Eds.), *The place of probability in science* (pp. 47–63). Springer.
Hausman, D. M. (2011). How can irregular causal generalizations guide practice? *Preventive Medicine, 53*(4), 229–231.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*, 49–67.
Hitchcock, C. (2018). Probabilistic causation. In *The Stanford encyclopedia of philosophy* (Fall 2018 Edition). Retrieved from https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*(1), 4–29.
Jacobs, A. M. (2016). Introduction: Mechanisms and process tracing. *Qualitative & Multi-Method Research, 1*(2), 13–15.
Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis, 23*(3), 313–335.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Mansfield, E. D., & Snyder, J. (2005). *Electing to fight: Why emerging democracies go to war*. MIT Press.
Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology, 21*(1), 54–76.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference* (2nd ed.). Cambridge University Press.
Narang, V., & Nelson, R. M. (2009). Who are these belligerent democratizers? Reassessing the impact of democratization on war. *International Organization, 63*(2), 357–379.
Parkkinen, V. P., & Williamson, J. (2020). Extrapolating from model organisms in pharmacology. In A. LaCaze & B. Osimani (Eds.), *Uncertainty in pharmacology: Epistemology, methods, and decisions* (Vol. 338, pp. 59–78). Springer.
Reiss, J. (2009). Causation in the social sciences: Evidence, inference, and purpose. *Philosophy of the Social Sciences, 39*(1), 20–40.
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. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*(2), 157–170.
Russo, F., & Williamson, J. (2011). Epistemic causality and evidence-based medicine. *History and Philosophy of the Life Sciences, 33*(4), 563–581.
Shan, Y. (2022). Philosophical foundations of mixed methods research. *Philosophy Compass*, *17*(1).
Shan, Y., & Williamson, J. (2021). Applying evidential pluralism to the social sciences. *European Journal for Philosophy of Science*, *11*(96).
Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.
Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science, 32*(1), 33–61.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.