A More Conscious Experimental Design: How More Explicit Use of Process Tracing Can Improve Experiments

Derek Beach (Aarhus University), Levente Littvay (Central European University)
emails : derek@ps.au.dk / levi@littvay.com

Abstract

Randomized experimental design is seen by many researchers in the social sciences as the gold standard of causal inference, but as a method, it is blind to the mechanistic processes that lead from cause to outcome. While experimental researchers are highly conscious of the mechanisms underlying their theorization and inference, they are mostly implicit in their research design and analysis. We argue that process tracing used as an adjunct method alongside an experiment can be used to make the mechanistic assumptions explicit in experimental design and analysis. Through this combination, experimental researchers can become more conscious about causal mechanisms in their theorization, leading to better experimental designs, more effective ways to discuss problems with causal heterogeneity, and even strengthening the internal and external validity of findings. The approach helps produce more effective treatment reinforcements, condition checks and other process information that is often collected but rarely utilized. Finally, these process tracing steps could help augment inference with mechanistic information at the unit (case) level, which would help disentangle and further theorize why and through what mechanism a treatment has an effect, or in the cases of null findings, why the theorized mechanisms broke down, thereby offering better insight even from null findings - an issue experimentalist struggle with since the replication crisis pointed to the importance of null findings. The propositions here would make implicit steps in theorization and design and analysis done by many experimentalists more explicit, allow for better designs, more effective use of all information collected and offer higher transparency for experimental researchers making their implicit processes more explicit.

1. Introduction

The use of experimental designs has become increasingly popular within the social sciences. The strength of an experimental design is the ability to make strong causal inferences about the average causal treatment effect within a sampled population due to the ability to control for potential confounders through design. While portrayed by many as the 'gold standard' (Gerring, 2011; Angrist and Pischke, 2009), an experimental study cannot not - and should not - stand alone. Instead, to properly design, implement and interpret the findings from an experiment requires the use of other methods acting as adjuncts to answer questions that the experimental design cannot by itself (Deaton and Cartwright, 2018).

Scholars engaging in experimental research in the social sciences – irrespective of whether they involve lab, survey or field experiments – frequently use other methods *implicitly* as tools to assist the development of hypotheses, design treatments, monitor the implementation of treatments, and interpret results. However, when presenting their research, most scholars present their design 'as if' the experiment was deductively developed without the assistance of other methodological tools, and thereafter interpret the findings 'as if' it was only the experimentally manipulated data that enabled inferences to be made. In the interests of research transparency, and to improve the application of experimental methods through more explicit recognition of the role that other methods play in developing good experimental designs and interpreting their results, this paper contends that scholars should *explicitly* acknowledge the role played by other methods used in an adjunct fashion alongside their experimental design and present this adjunct data used in the design, implementation and interpretation phases in a more transparent fashion.

This paper develops one such combination; the use of process-tracing case studies (PT) in parallel with an experiment, where PT is used in an adjunct fashion for particular tasks during different stages of the experimental research. At its core, PT involves theorizing a process-type theory about what mechanism(s) link a given cause and outcome together, unpacking theoretically what is going on in-between X and Y, and then tracing how it worked empirically using the observational traces that activities associated with parts of the process leave within a given case.

Note that our argument is not about *mixed-methodology*, defined as research where methods from different methodologies are combined in a manner in which very different types of causal claims are made that are evidenced using very different types of empirical material in seamless combination with each other (Beach and Kaas, 2020). Instead, this paper explores a *mixed-methods* design, where a method from another methodological tradition is used as an *adjunct* to improve the inferences made using the core method. In other words, one methodology - here counterfactual-based experimental design - is in the driving seat, with PT used as a secondary, adjunct tool to improve the design, implementation and interpretation of findings gained from a social science experiment.

In this paper, we discuss the potential uses of explicit PT as an adjunct method to improve experimental research in the lab, through surveys, or in the field. We discuss in detail how PT can assist in developing stronger theoretical hypotheses by unpacking hypothetical processes and mechanisms, how these insights can be used to develop meaningful 'treatment' interventions for the experiment, and how observational process-related data collected in parallel with experimentally manipulated data can be used to assist in evaluating how the treatment was implemented and interpreting what causal effects found/not found by the experiment actually mean. We present step-by-step guidelines for how researchers can productively use PT as an adjunct tool for each step of the research process from the initial theorization of a process-model for the intervention to interpreting results. We present ideas for how this theorization and process-related data can then be used and presented more explicitly in the transparency appendix as a supplemental data source to answer questions that the experimental design by itself cannot.

2. An short introduction to experimental design and the well-known challenges

Randomized experiments are the gold standard of scientific inquiry when making counterfactual-based causal claims. This is due to experiments' ability to test the average effect of a researcher-introduced treatment free of potential selection bias into the receipt of the treatment. The experimentally manipulated data allows for the elimination of reverse causation or spurious findings driving covariation between the cause and the outcome under study, allowing for the inference to be truly causal (Woodward, 2003). Building on the counterfactual understanding of causation, experimental designs are powerful tools to enable testing of causation that have been formalized within the potential outcomes framework (Woodward, 2003; Morgan and Winship, 2007).

A randomized experiment compares two groups randomly sampled from a population and hence expected to be identical, on average, in every way within the bounds of sampling error, where one of the groups (the treatment group or T) receives a treatment (i.e. the key independent variable of the study, Xt) while the other group (the control group or C, Xc) does not. Then the two groups are compared on an outcome, or the dependent variable Y. The power of the design is that when a large number of units are randomly assigned to different groups, the units in both the treatment and the control group are statistically expected to be equal, on average, on all of their characteristics. This means that short of a "by chance" failure in randomization, the units in the treatment group and control group will be completely comparable, also referred to as 'in balance', on all possible covariates. This means that there will be no confounding effects driving the differences in their outcome measures. This is not to say, nothing (else) can go wrong in the design or the implementation of an experimental study. These issues we discuss after a brief introduction of process tracing.

The Replication Crisis

In the past decade experimentalists of all the social sciences had to struggle with a deep introspective evaluation of what is wrong with the paradigm. The replication crisis swept through the social sciences, and although it is not limited at all to experimental research, experiments were at the center of the critique. The field has met the challenges posed

admirably by extensively exploring what led to the failure to replicate so many studies. Countermeasures, such as pre-registration, strict standards were put in place to minimize the effects of multiple testing and researcher degrees of freedom allowing scholars to analyze data enough ways to squeeze out positive findings and are becoming the norm and expectation for experimental studies. Concurrently there is a push for stronger theorization and also the publication of null findings. Scholars have proposed the institution of "registered reports", papers reviewed and accepted based on theoretical grounds before the collection of the data. In these efforts, the community is having difficulty with the cooperation of the journals (or even the journal reviewers in the few attempts that journals made in this direction). PT methods have the potential to help with stronger theorization potentially allowing to push forth on this path. Additionally, PT also has the potential to elevate insights from null studies to more publishable levels.

The validity of experimental inference also depends on the quality of the research design. Threats to validity can be grouped into two categories, internal and external validity and the history of science is riddled with experimental design failures of both types of validity. The struggle to perfect experimental design is largely a struggle to maximize internal and external validity and we believe PT can aid experimentalists further along this path.

Internal Validity

Randomized experiments rely on the assumption that the assignment of the treatment is (1) random where each sampled case from the population has equal chances of being sampled into either the Treatment and in the Control groups and (2) the treatment assigned is the treatment and only the treatment. Most common examples of failures here include *differential drop out rate* for treated and control individuals. While the treatment group and the control group has equal chance of being assigned to their group status, if there is a differential selection bias in being observed, balance on the covariates suffers if different types of people drop out of the study once the treatment is introduced than who drop out of the control group. The most classic example of *assignment failure*, on the other hand, is what led to the introduction of double-blind experiments for the evaluation of medicine. Originally, individuals assigned to the treatment group received a medicine as the treatment,

the control group received no drug. Over time, scholars found that the receipt of a drug constituted more than the chemical input of the drug into the body. The receipt of care by a physician, the receipt of a pill often also lead to improvements in health outcomes despite the chemical compounds having no effect. This led to the introduction of the placebo, an equivalent looking and feeling intervention that allowed for the isolation of the chemical compound from the additional care components of the intervention. The introduction of other treatment components going beyond the chemical compound added was a failure of internal validity, but the fact that the story did not end here highlights well how difficult it is to get internal validity right. Today, the standard for drug trials is a double-blind study where neither the participant, nor the administrator of the treatment knows if the study participant is receiving the treatment or the placebo. These administrators often still subliminally communicated expectations to the study participants contaminating the internal validity of the experimental study. These expectations, in addition to the actual effects of the drug, also translated into outcomes. The isolation of the specific treatment is extremely difficult especially in social situations. Process tracing could help theorize, systematize and therefore isolate the processes through which we expect a treatment to exhibit its effects.

Internal validity is more difficult in the social sciences than in the natural sciences. Lab experiments, through the presence of the study administrator, have the potential to introduce potential sources of contamination. Even in survey experiments, different stimuli (or the lack thereof) could affect different stimuli, having differential impacts on attention and engagement. Process information, such as eye and mouse tracking, could inform the presence of a possible failure (potentially offering a viable explanation for a null finding).

External Validity

A second threat to the validity of an experimental study is external validity, which is the study's ability to *generalize* to other populations or to broader applications of the treatment beyond the confounds of the experimental study's implementation specifics. Lab and survey experiments, for example, are notorious for external validity problems. Studies in the lab's (or worse, a survey's) artificial environment littered with a healthy dose of Hawthorne effect may not generalize to some real life field implementation of the same treatment as one would

initially expect. For example, in communication lab or survey experiments, we receive visual, textual or auditory stimuli in a vacuum, in an unnatural environment taken out of context. Researchers conducting such studies take a leap of faith of external validity when they suggest the stimuli will work the same way in a more natural environment. Applying even a well implemented field experimental situation to another population also has its dangers (Cartright, 2012). PT offers tools to evaluate stimuli both theoretically and empirically allowing the identification of mechanisms through which the stimuli worked (Steel, 2008; Khosrowi, 2019; Beach and Pedersen, 2019). These mechanisms can inform how much we can expect the findings to generalize beyond the artificial research environment or to other populations with potentially different characteristics.

Heterogeneous Treatment Effects

While experiments are highly effective in estimating average treatment effects causally and void of all confounds. But the average treatment effect is exactly as it's name suggests, an average effect. A small average effect can be acquired both through the small effect homogeneously observed among all treated units but it can also be a large effect for a small sub-group of the treated units leaving the rest of the treated units (and, of course, all the controls) without the tested impact. In extreme circumstances the treatment could have an opposite effect on some individuals even in the majority of the sampled population as long as the positive effect in others, on average, outweighs the negatives. This problem poses an aggregation-disaggregation problem. Just because an aggregate average treatment effect is present, we cannot even probabilistically advise any individual to take the treatment if their desired outcomes are in line with the past experimental findings. This problem is known as the heterogeneous treatment effect. Let's break the problem down further.

The design underlying a randomized experiment is unable to identify individual treatment effects as it relies on the comparison of otherwise similar samples with and without the treatment. In other words it cannot compare each treated unit to its counterfactual, the same unit under all the same circumstances without the receipt of the treatment. For this reason, somewhat uncharacteristically, experimentalists have turned to the exploration observed covariates' (and their higher order interactions') impact on the treatment effects, even

deploying big data approaches to reduce the number of models that have to be deployed. This approach, of course, suffers multiple problems.

The exploration of many covariates and their higher order interactions turn a single statistical test of the treatment effect to a very large number of statistical tests even with a very limited set of covariates deployed exponentially leading to a multiple testing problem. With the replication crisis, multiple testing has turned into one of the main criticisms of experimental research. One of the causes of studies' failure to replicate' is assumed to be the file drawer effect, where many studies that were conducted (but yielded null results) ended up in file drawers, with only positive findings published. This process probably led to the publication of more type I error findings that set levels of significance suggested, without any malicious intent by the researchers making it hard to detect. While corrections for multiple testing exist, all of these require more and more statistical power, larger sample sizes. The irony of using a large number of potential observational covariates to understand treatment heterogeneity should not escape anyone. Experimental research built its foundation in the social sciences on the critique of correlational and observational approaches for not being free of confounds, even with many observational controls introduced. While understanding the variation in the treatment effects with regards to a large number of covariates may be informative in understanding how much variance is in the treatment effect, the problem of omitted variable bias remains. Still, just like with necessary control variables in observational studies, the number of covariates to consider in these studies quickly bubbled out of proportion. The main problem is that experimental researchers do not have good theory generating approaches for understanding causal heterogeneity. So it is natural to resort to exploratory research relying on available covariates as opposed to generating the appropriate theories and corresponding data to better understand causal heterogeneity. Process tracing offers a systematized approach to theorize causal heterogeneity. At minimum, it is easy to argue that both close theorization of heterogeneous treatment effects and the limitations to potential covariates would produce large benefits to scholars wishing to explore treatment heterogeneity. For this reason, any approach that helps scholars theorize heterogeneity offers a lot of promise to the field.

3. The core elements of process tracing

Most political scientists believe that process tracing was developed within the discipline, but it actually has its roots in cognitive psychological experiments. In order to try to understand the mechanisms behind observed behavior in these experiments, process-related data were collected using techniques such as mouse and eye tracking (including measures of eye fixation and pupil dilation) to help understand observable manifestations of reasoning processes, including people's attention, interaction with surveys and other treatments. Less high tech verbal reports of observational traces left by reasoning processes are also common in the psychological variant of process tracing. One form of this, while not explicitly labeled process tracing, is the use of cognitive interviews in survey research, which is one of the foundational techniques of effective survey question development. While analysis of public policy or international relations and cognitive neuroscience might seem worlds apart, what they have in common is that they take process mechanisms seriously, something that experimental researchers explicitly keep in a black box.

At its core, process tracing has a theoretical and empirical dimension. At the theoretical level, the word 'process' refers to the mechanistic type of causal claim being made. In contrast to a counterfactual, process tracing involves a mechanistic claim about the *linkage* between a cause and outcome. In the philosophy of science, mechanistic causal claims are understood as the pathways or a via media in which entities perform activities that link causes and outcomes together in a relationship of production or generation (Machamer et al 2000; Machamer 2004). In medicine, whereas the experiment enables us to assess the average treatment effect, the medical version of process tracing would involve tracing the chemical reactions induced in different parts of the body that eventually produce the effect.

The critical components of process theories are *entities,* which are social actors that do things, and the *activities* that transfer what can be termed 'causal forces' to the next entity in the process. The *productive* element of a process theory comes from the focus on the activities performed by entities that provide the causal linkages between parts of a process (Kaiser, 2017; Piccinini, 2017). In other words, the process theory provides enough information about the component parts and the activities linking them together that we are

able to answer the 'how does it work' question (Craver and Darden, 2013). If the experimental treatment is a vignette that provides a partisan cue to a participant, our process theory would unpack the reasoning process whereby this cue is theorized to produce a particular outcome (e.g. attitude change).

In an experiment in which the research focuses on how an individual responds to a given treatment (e.g. a textual frame), the only entity being theorized would be the individual themself. In this type of experiment, what the treatment is doing in a causal sense disappears into the reasoning processes going on in the subject's head - the challenge that led psychologists to develop versions of process tracing methods initially. Yet even though the reasoning process is internal to the subject, theorization of what is going on helps us understand how the cause might be working, which then also enables us to think about potential traces that might be observable for different parts of the hypothesized process. Other experiments might involve *interaction* between individuals, in which case there would be multiple entities involved in a process, and the activities that they undertake to impact each other are what binds them together in a causally productive process.

At the theoretical level, exposing the micro-foundations of processes requires that they are unpacked through developing a blow-by-blow theory for the causal process linking X and Y together. This means going beyond depicting processes as causal graphs, in which the linkages are depicted as arrows but whose content is not described theoretically (e.g. Pearl, 2000; Waldner, 2015). As an example, Waldner produces a causal graph linking liberal ideas with the democratic peace, in which one path is depicted as: liberal ideas -> ideology (TRUE) -> refusal to fight democracies (TRUE) -> constraints on government (TRUE) -> democratic peace (TRUE) (p. 247). However, by using a causal graph, we are in effect black-boxing who is doing something and what they are doing that could lead to the next part of the process. Who is refusing to fight other democracies? What are these persons doing that can impose constraints on a government?

Using the typical notation of "variable" followed by an arrow in a causal graph, we lose sight of the causal linkages theoretically, which also means we do not study them empirically!

Theoretically unpacking a process forces the analyst to make transparent the causal logics that are theorized to bind the process together in a productive relationship - in other words, *why* a particular activity is expected to have a particular impact on the entity in the next part of the process. Additionally, making key linkages explicit also flags for us where our process theorization is fuzzy, thereby focusing our attention on what requires further theorization and/or empirical probing of a case to figure out how it worked.

Note that 'unpacked' process theories can have varying degree of analytical abstraction, ranging from a quite abstract process theory that only "…describes some of the internal details of the mechanism but has black boxes signifying that one or more relevant component parts, activities, and organizational features are unknown" (Craver and Kaplan 2020:299), to a highly detailed process with many parts, in which "[…] all of the entities, properties, activities, and organizational features that are *relevant* to every aspect of the phenomenon to be explained" (Craver 2006, 360 [italics inserted]; see also, Craver and Kaplan 2020). At a minimum, a process theory has to enable the analyst to answer the 'how does it work' question (Craver and Darden, 2013), a question often ignored in counterfactual empirical explorations using randomized controlled trials, even when the initial theorization involves a causal graph describing what is hypothesized to be going on in-between treatment and outcome.

Theorizing a causal process is not merely a logical thought experiment; instead they should be built upon all of the available relevant theorization and empirical knowledge that one possesses. Table 1 and 2 produce two hypothetical examples of what 'unpacked' process theories could look like; both of which could in principle be investigated using experimental designs. The first deals with an individual-level categorization learning process inspired by research in cognitive psychology (Gentner et al 2009), in which a single individual is the entity, and the activities are what the person is doing cognitively when learning from failure (Y). In this process theory, the cause is that an individual is confronted with a situation in which a chosen course of action has failed to achieve the objectives. The first part of the process involves recognizing failure, followed by assessment and identification of a factor (X) that is understood by the person to have caused failure (Y). The critical part of a

categorization learning process occurs when the X:Y lesson is 'stored', where the mind's tendency to focus on similarities while ignoring differences comes to the fore (Gentner et al 2009: 1344-5; Sagi et al, 2012). The outcome is that a person has learned a 'lesson' from the failure of Y that they can draw on when they encounter a new situation in which Y might be a relevant action, but where the lesson leads them to choose another action instead even though the context might be different in the new situation.

Cause	Part 1	Part 2	Part 3	Outcome
New information (failure of Y)	Person *recognizes* failure	Person *engages* in assessment to *identify* the X that caused Y with little consideration of importance of context	Person *stores* lesson that is stripped of contextual information	Person draws on X:Y lesson when they encounter case of Y in future

Table 1 - A hypothetical categorization learning process

Table 2 depicts a hypothetical shaming process in which a person who makes a prejudiced statement within a group that shares norms related to anti-prejudiced attitudes. In the process theory, the cause (or trigger) is the actual prejudiced statement made by a person, which is followed by one or more persons in the group noticing the clash between the statement and the shared norm. The unease produced by this clash triggers the person (or persons) to signal verbally and/or non-verbally that the prejudiced statement is normatively wrong. The next part is that the rest of the group affirms the 'wrongness' of the statement, producing the outcome of shaming into silence of the individual who made the original statement.

Cause	Part 1	Part 2	Part 3	Outcome (shaming)
Person makes prejudiced	Person(s) in group notices	Person(s) signals verbally and/or	Rest of group affirms verbally	Person who made

ethnic/racial statement that clashes with norms held by rest of group	that statement clashes with group norm, upsetting the person (TERM)	non-verbally that statement is 'wrong' that acts as a cue for the rest of the group	and/or non-verbally 'wrongness' of statement	statement is shamed into silence

Table 2 - A hypothetical shaming process

It is important to reiterate that the understanding of process tracing that we develop in this article builds on the *mechanistic* understanding of process found in the philosophy of science; also termed the *productive* account in the literature (e.g. Machamer et al, 2000; Russo and Williamson, 2007; Clarke et al 2014). If a mechanistic claim is treated as a counterfactual, this defines away the analytical value-added of using process-tracing in conjunction with an experiment. Scholars like Woodward otherwise contend that mechanistic claims can be understood as lower-level counterfactual claims by breaking the bigger counterfactual X->Y into smaller bits (e.g. X->M1->M2->Y) (see Woodward, 2003; Imai et al, 2011). However, to be methodologically in alignment with the underlying causal claim, evidencing a counterfactual requires controlled comparison - ideally using a randomized controlled trial! This is in effect causal mediation analysis (Imai et al, 2011), which involves treating causal mechanisms - either as a whole or broken into distinct parts - as intervening variables that in effect are counterfactuals claims. The analytical implication then is that these counterfactual claims have to be investigated by assessing the difference that variation in the intervening variable(s) has for values of Y *across* a set of cases, controlled for other variables (e.g. Imai et al, 2011; Reinhardt, 2015). However, as the effects are assessed *across* cases, the *actual operation* of the mechanism and the critical causal linkages between parts are *not* explored empirically in individual cases (Illari and Williamson, 2011; Green et al, 2010). Further, in this understanding, mechanisms are typically theorized in such a superficial fashion that the process itself is in effect black-boxed theoretically because the causal linkages are not unpacked (Baetu, 2016).

As an example, in a recent article in *APSR*, Szakonyi (2018) explores which mechanism links Russian businessmen becoming legislators (cause) with economic benefits for their firm while they serve in office (outcome). He suggests two pathways and develops empirical proxies for each: either bank lenders look more favorably on businessmen with political success, or political success opens doors to the bureaucracy, resulting in favorable treatment in regulations and procurement (p. 323). He then tests which of the two mechanisms are operative in a mediation analysis using a large number of cases. He finds that when controlled for all other factors, winning candidates are more likely to win more state contracts (p. 333). However, his analysis does not shed light on *how* political success led to more state contracts in actual cases because we do not know what activities the businessmen, bureaucrats and other politicians were doing in the process that can have led to the award of state contracts. In other words, we have no evidence of how the mechanism actually worked; instead, we merely measure empirically that a proxy is present that suggests something might be going on in-between. Tracing how the process worked, understood as a mechanistic claim, would be focused more on understanding the linkages between the activities of politicians, businessmen and bureaucrats, enabling us also to better understand the impact of context in which the activities had particular effects (Khosrowi, 2019). Additionally, because the mediation analysis assesses *average* causal effects, we gain no information about how the mechanisms actually operate within any *particular* case (Leamer, 2010; Green et al, 2010).

At the empirical level, process tracing as a method involves making causal inferences by collecting what some philosophers of science term 'mechanistic evidence' (Russo and Williamson, 2007; Clarke et al, 2014). Mechanistic evidence is defined as the *traces* left by the activities for each part of a process within a given case (Ibid). Mechanistic evidence is thereby *observational* data, and it can take many forms depending on what the activity is and the evidential context in which the activity takes place. In contrast, the evidence used to make causal inferences using experimental designs can be termed 'evidence of difference-making', which takes only one form: measures of values of X and Y *across* treated and non-treated individuals. A key point here is that process tracing involves assessing whether hypothesized traces left by activities were present *within* a case in which Xt was present.

While the assessment of negative (Xc) cases is necessary in experiments, using observational evidence in process tracing to try to find traces of activities in a negative case is like trying to observe parts of the smoking -> cancer mechanism, e.g. the part where smoke destroys the hair-like cilia in lungs, in a case where no smoke is not present.

The inferential value of evidence of difference-making garnered with an experiment is straightforward, being the magnitude of differences of case scores on Xt/Xc and Y. In contrast, mechanistic evidence can have varying inferential value depending on how directly the activities and their linkage to the next part of the process are observed. Actual observation of the activity in which the researcher in real-time witnesses the impact it has on the next entity in the process would be much stronger evidence of a causal linkage than more indirect evidence - what is termed circumstantial evidence in law. For example, we might only be able to observe a sequence in which something like a policy frame was not present before, then the activity such as a speech act occurred, after which the new policy frame was present in the debate. Here, instead of directly observing the *linkage* provided by an activity, in this example we would be *assuming* that the speech act produced a new policy frame in the debates in a causal sense using indirect, circumstantial evidence. In particular, it would be the sequence that would allow us to assume linkage. In contrast, if we were able to directly observe through our participation that the frame of the debate radically shifted after the speech act, and there was nothing else going on at the same time that could have done it, we would be able to make a stronger inference about the linkage.

As an example, in the hypothetical shaming mechanism, direct evidence of part 2 and the causal linkage that the activity provides would be the observation by the researcher of some verbal/non-verbal cue from the person that noticed the clash which is then picked up by the rest of the group, who then engage in the activities associated with part 3. In contrast, for part 1 we might only be able to observe that some person made a prejudiced statement that was then followed by a person in the group reacting to it with a cue. Here the sequence makes it very plausible that someone became upset, and that this led them to take an action in part 2, but we have not actually observed the linkage. The methodological point here is that not

all mechanistic evidence is created equally in terms of the strength of causal inferences that they enable.

Concluding, if process tracing is to be used in parallel with an experiment, this involves: 1) developing a process theory for how the treatment (X) could be linked in a productive (causal) relationship to an outcome (Y) through a series of parts composed of entities engaging in activities that provide the productive linkages in the causal theory, and 2) operationalization of what potential observable manifestations (traces) might be left by the activities associated with each part of the process.

The theoretical development of processes is a relatively straightforward extension of the theorization that we would do anyway in good experimental research. At the empirical level, experimental researchers often do collect (or could collect) mechanistic evidence without realizing it is process data. What's more, this data is often discarded as lacking a useful contribution to the experiment. Decisions about the collection of process data are most often theorized at a superficial level and this is one area where more conscious usage of process tracing can improve experimental designs and augment inference.

4 - How more explicit use of process tracing can help dealing with the well-known challenges to experimental designs

[NOTE TO THE READER - the following is still relatively preliminary ideas that will be fleshed out in the coming months. We had planned on submitting a more-or-less finished manuscript to APSA, but COVID came in the way. Both authors are academic co-convenors of the ECPR Method School, and we spent most of the spring and summer developing an online version of our method school event. We are very proud of the event we were able to hold! (see https://ecpr.eu/SummerSchool). But the following sections suffered as a result!]

We divide an experiment into four phases, depicted in table 1. First, the development phase is when a cause and effect is theorized. This phase includes theoretical considerations of what the treatment and outcomes are, with special attention focused on what the treatment is not, or, in other words, all possible ways the treatment may be contaminated threatening the internal validity of the experiment. This is the phase where the need for a placebo is considered at the theoretical level by pondering what would constitute equivalent stimulus to both the treatment and the control group, but where only the treatment group receives the theoretical cause. Additionally, at this phase experimentalists should also ask what is it about the cause that can constitute a causal effect.

	Development phase	Design phase	Implementation phase	Interpreting the results
analytical tasks	- theorize treatment/outcome - what is the actual treatment? - what might work as a placebo?	- developing a selection rule for sampling a population - operationalizing the treatment, placebo and outcome	- selection into groups - administering the treatment/placebo - collection of data	- assessing the ATE - what do the findings mean?

Table 3 – Analytical tasks for each stage of an experiment.

Second, the design phase mirrors the development phase, except the theoretical decisions are here transferred into a specific empirical experimental design, an experimental procedure. The carefully considered cause is developed into the actual treatment participants will receive and decisions are made on the operationalizations of the outcome. Is the treatment a proxy for some other underlying cause? How crude is the proxy? Is there an underlying mechanism that is not, though maybe should be tapped by the design? If parts of the underlying mechanisms are black-boxed, that should be a conscious and not an intuitive decision.

Third, the implementation phase is where the experimental procedures are implemented and the data is collected. This involves the selection of participants into the treatment and control groups, the administration of the study and the collection of the data. Special attention needs to be paid to what, if anything, goes wrong in the implementation. Was the randomization successful in producing balance between the treatment and control groups? Is there potential post-treatment dropout which may be different for the treatment and control groups, for example.

Phases 1-3 are where the experiment's internal validity may suffer. The final phase involves the interpretation of the results, the statistical calculation of the average treatment effect, usually through the comparison of the groups on the outcome. And this phase also includes the interpretation of the results and what they mean in relation to our existing knowledge. Both internal and external validity needs to be considered in this phase. Potential challenges in this phase include an understanding about the potential heterogeneity of the treatment effect, questioning if the overall ATE findings can be implemented on any individual with the expected results. Interpretation should go beyond asking if the treatment does (or does not work) and answer why it does (or does not) work? This is where potential generalizability of the finding should also be considered.

We now turn to a more practical discussion of the explicit use of process tracing methods in conjunction with an experiment, and how the combination can help tackle well-known challenges facing any experimental design.

The development phase

While the controlled manipulation of a treatment within an experimental design is enough to enable causal inferences to be made (Woodward, 2003), developing an explicit process theory during the development phase that explains the causal linkage between a treatment and outcome can help improve an experiment. In Popperian language, the experimental design has its strengths in the 'context of justification', but is less strong in the pre-experimental 'context of discovery'. The counterfactual causal claim can be evidenced by merely assessing the empirical difference that variation of X makes for Y in the controlled experiment, meaning that it is not necessary to actually develop a strong theory for why X makes a difference. The implication of this is that experimental research often neglects more robust theorization about *what it is about* the treatment X that could actually produce variation in Y. The analytical result of this theoretical oversight is that many experiments proceed with relatively crude theoretical proxies instead of more focused theorization about what it is about a cause that could be a cause (Deaton and Cartwright, 2018: 4; Clarke et al, 2014). Simply speaking, while knowing that something works is a contribution, without an adequate understanding of how it works, why it works and under what conditions it works, the scientific value of the finding is more limited in its scope and generalizability to other situations, applications or populations. The more these questions are developed (either merely at the theoretical level, or through inferences made using mechanistic evidence), the stronger the overall causal inference.

Theorizing a causal process in terms of parts composed of entities engaging in activities in sufficient detail to be able to answer the question 'how does it work' therefore has multiple uses during the development phase and onwards. First, and most obviously, if we are unable to formulate a plausible process theory for the linkage of a hypothesized treatment with an outcome, there is no reason to engage in an experiment in the first place.

Second, theorizing the process sheds light on what attributes of a treatment might actually be causal, helping the researcher figure out the appropriate level of analytical abstraction at which to conceptualize the treatment, and to design the subsequent manipulation of the

treatment in the RCT. Using an example from the natural sciences, there is much research on the relationship between drinking red wine (X) and heart attacks due to blood clots (Y). However, before a good experiment could be designed it would be important to theorize more about what it is about X that could produce variation in Y because red wine has multiple causally relevant attributes that might produce differences in Y, including alcohol and resveratrol (Andriantsitohaina et al, 2012). Given that there were multiple potential causally relevant attributes that would be present if the presence/absence of red wine was manipulated in an experiment, researchers first developed a process theory linking one attribute (resveratrol) and how it could potentially be causally linked to Y, after which an experiment was undertaken to explore the causal effects of varying dosages (Andriantsitohaina et al, 2012).

It is of course possible to develop a compelling process theory that is just that - a nice causal story (Clarke et al, 2014: 350). This is why the literature on process theories in medicine insists that what is important is what comes next - the *evidence* of the process theory, and not the theory by itself (Ibid). The methodological implication of this point is that it can be a good idea during the development phase of an experiment to conduct a pilot study using process tracing to investigate whether there is any mechanistic evidence that suggests the theorized process might actually be taking place in real-world cases. Given space constraints, we do not elaborate on how process tracing case studies can be undertaken, but instead reference other work (e.g. Beach, 2017; Beach and Pedersen, 2019).

Third, developing a plausible process theory also improves our theoretical understanding of a given phenomenon, which in principle also makes nul results interesting when they contradict a plausible causal process theory. One of the recommendations of the replication crisis is that null results should, in fact, be published. But neither reviewers, with a status quo bias of wanting to see positive findings, or journal editors who understand that negative findings will likely receive fewer citations, seem to be convinced about the utility of publishing null results. Attempts to overcome the issue include a call to consider studies as registered reports: theoretically developed and empirically well-designed study plans put together before the collection of the actual data. But it is questionable if the social sciences

in general have matured to a level of scientific inquiry that studies can be judged solely on their theoretical foundations and study designs, independent of the findings. If the discipline wants to take a step in this direction, experimental researchers will likely be at the forefront of this. However, experimentalists are badly in need of stronger process theories for what pathways can potentially produce differences. The status quo of studies published solely on the merits of the empirical findings (and their causal identification strategies) is not going to suffice to move the field forward [EXPAND A BIT ON THIS ARGUMENTATION].

The design phase

Theorizing a process also helps in designing how the treatment, placebo and outcome will be measured empirically in an experiment. In contrast, the result of using crude theoretical proxies for treatments in experiments is that the actual measurement of the treatment can be invalid in relation to the theoretical cause we actually are interested in measuring. This relates to what Morton and Williams term 'construct validity', which is how valid the inferences based on the data are for the theory being tested (Morton and Williams, 2010: 260).

While this can be a problem in all types of experimental research, it is often particularly problematic in survey research, where the experimental manipulation is to deploy different questions to respondents. As an example, Tomz and Weeks (2013) used a survey experiment to attempt to test whether citizen perceptions of democracy (X) made respondents less likely to support military action (Y). While they do claim they are theorizing the 'causal mechanisms' linking X and Y, they never theorize them beyond headlines like 'threat perception' and 'likelihood for success'. For instance, they write that, '..For example, Lake (1992), Reiter and Stam (2002), and Bueno de Mesquita et al (1999) argue that wars against democracies are especially costly...Following this logic, citizens may be deterred from using force against democracies because they anticipate high costs of war and a low probability of victory…' (p. 851). However, this leaves us in the dark regarding how it might actually work. Should we expect that average citizens have learned (through schooling or their own analysis based on recent (or historical) events?) that democracies are tougher opponents than autocracies? Do citizens have to be aware that they have these attitudes for them to

make a difference in their willingness to support military action, or can it function as a set of subconscious background assumptions?

In the study, they then asked the survey question 'There is much concern these days about the spread of nuclear weapons. We are going to describe a situation the [U.K./U.S.] could face in the future. For scientific validity the situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. After describing the situation, we will ask your opinion about…' '…attacking the country's nuclear development sites now [they could] prevent the country from making any nuclear weapons.' (p. 853). The survey manipulated questions related to political regime type (Xt), along with a set of control variables (military alliances and military power/levels of trade). They find a difference produced by regime type. In relation to the 'causal mechanisms', they do not find a strong relationship for likelihood for success.

However, given the lack of theorization of the processes linking perceptions (X) to support (Y), the study suffers from a range of problems, irrespective of its other merits. Most obviously, is the study even measuring perceptions of democracy by telling respondents in a survey that a political regime is democratic or not? One can question whether an average respondent would be able to define the differences in regime type in abstract terms, or whether the respondents are actually using real-world examples of regimes to understand the question. For example, as the surveys were deployed in April-May 2010 (UK) and October-December 2010 (US), there was extensive coverage in the media of the Iranian nuclear program and the threat that it potentially posed (the UN Security Council imposed harsher sanctions on Iran in June 2010). Therefore, while the theory claims that perceptions of regime type *in itself* can make a difference, the empirical measurement arguably taps more into the complicated issues of how citizens perceive current events and their views of the potential dangers of Iran possessing nuclear weapons. In other words, we are unsure what the survey treatment is actually measuring empirically.

Related to this, the study claims to measure the 'mechanism' of perceptions of costs and likelihood of success by asking citizens to estimate the likelihood of events such as whether

the country would respond to attacking the US. However, if the underlying process theory operates more at the subconscious level, should we expect that citizens will provide differing estimates when asked off the cuff in a survey?

Stronger process theories would have tried to unpack the reasoning process of citizens related to the mechanisms hiding behind headlines like 'threat perceptions' and 'costs of fighting'. Related to citizen anticipation of costs of war, an unpacked process theory would dig into the types of information the citizen might use (input could range from things they learned in school, past experiences, or current events), whether they engage in some form of rational calculation and how it might work, and whether the process might interact with other processes (e.g. morality concerns, or threat perceptions). Doing this would help the researcher design better measures of the treatment and proxies for the process itself. Of course this requires significant theoretical work - and perhaps a range of process tracing pilot studies to explore how the relationship is working - but nobody said good research was easy!

Theorizing a process - and ideally also engaging in an actual process tracing pilot study in the early phases of an experiment - can also help the experimentalist think more carefully about what type of individuals would be relevant to recruit for the sample. Put simply, the sample should be composed of units in which the mechanism(s) linking X with Y can potentially work (Morton and Williams, 2010). Further, the units selected should not be too heterogeneous. For example, if we are investigating how the beliefs of individuals structure how they perceive political information, and our process-level theorization that built on existing theoretical/empirical knowledge suggested that there might be different processes taking place depending on whether an individual had strong beliefs or less-strongly held beliefs. In the former, the process might plausibly be a more-or-less automatic reaction that does not even involve any cognitive thought, whereas in the later, it might involve considerable reflection or cue-seeking etc (Evans, 2008). In this example, we might then choose to focus only on individuals with strong beliefs, designing the selection of individuals in order to screen for this characteristic.

The implementation phase

Engaging in some form of process tracing while running an experiment can help answer questions relating to whether the experimental design was faithfully implemented - termed manipulation checks in the literature (Hauser et al, 2018). While these forms of manipulation checks are often conducted in parallel with experiments, we contend that experimentalists lack a vocabulary for evaluating them both at the theoretical and empirical level, with the result that this information is only treated as 'checks' without real inferential value.

In contrast, using the process theory developed during the development/design phases, effort can be put into collecting different forms of mechanistic evidence for all or parts of the process *in parallel* with the conducting of an experiment. This information can strengthen the internal validity of the experiment by checking whether the treatment is actually being implemented, but also enabling us to gain clues about what potential mechanisms might be operative across different types of participants. For example, while an experimentalist might view the collection of information about how participants perceived an intervention merely as a validity check of whether we are measuring what we think we are measuring, seen in process tracing terms this could also act as mechanistic evidence that sheds more light on the underlying mechanisms binding the treatment with the outcome. In other words, the data graduates from being merely a 'check' on measurement validity to being data enabling some inferences about the underlying mechanisms themselves.

There are a variety of tools commonly used to check that the manipulation is faithfully implemented. We contend that this information also has evidential value if treated as mechanistic evidence. For example, attention checks, as simple as a trick question to see if the respondent is an actual person or a robot randomly clicking, or as complex as mouse and eye tracking, can also be considered mechanistic evidence. Survey methodology has been using eye tracking to see if the survey questions are adequately read and processed. The time with which a respondent answers a question could also function as mechanistic evidence, in particular if we are working with a dual-process model of cognition.

Condition checks and treatment reinforcements are often part of experiments, added intuitively to serve the sole goal of seeing if the treatment was received by the participant or to make sure the participant takes the time to process the intended treatment. Condition checks developed in line with theorized processes offer some, minimal, evidence with which respondents could explore alternative theorized processes. Treatment reinforcements, such as elaboration tasks making the respondent write about what was present in the stimulus, often offer rich contextual information on how a respondent thought about the stimulus and if the stimulus achieved their intended process. But these elaboration tasks are included intuitively, whereas more explicit process tracing would force the researcher to home in on and specifically enquire about the theorized underlying processes. This mechanistic evidence can then be assessed qualitatively to either reinforce the experimental evidence or, in the instances of null results, to enable us to trace where the theorized processes broke down, to better understand the null finding.

In the lab and in the field, extensive behavioral evidence is potentially available that has the potential to shed light on treatment effects at the process level. Information is often automatically volunteered in exchanges between the person administering the experiment and the participant. But such information is rarely recorded systematically or even superficial as part of a debrief of the study administrator. Lab (and increasingly, with the miniaturization of biotracker devices, also field) experiments also allow for the recording of physiological information that could include heart rate, skin response and startle response and is not limited to eye tracking (Oxley et al 2008, Lahey and Oxley 2016). Exploration of the assignment process can help you with evaluating whether the selection of individuals into groups was actually random in your experiment.

Further, theorization with PT enables the researcher to be more conscious about what variables we need to check about to ensure we have balance. Returning to the example on the survey experiment of the democratic peace, stronger process level theorization could help in the proper selection of individuals in order to ensure balance. [MORE HERE]

Of course there is the risk that gathering all of these types of PT data might interfere with the treatment itself. Hauser et al (2018: 9) write that, 'By adding additional measures the researcher may change the internal psychological processes. There is more than one way that a manipulation can be validated, and researchers should give the same careful consideration to their choice of a manipulation check as they do to their choice of dependent variable measures. Authors should justify including a manipulation check with an experiment if they chose to do so, explaining why it is necessary and why it is unlikely to affect their conclusions. Often the best choice may be to forego including a manipulation check in the actual study by establishing the effectiveness of the manipulation through other means such as in pilot work. Editors and reviewers should evaluate whether a particular manipulation check improves or impairs the quality of any given study rather than assuming that using a manipulation check automatically improves it.'

To avoid this critique, we suggest that explicit process tracing of a strategic sample of individuals in parallel with an experiment - what Hauser et al (2018) term 'pilot work' - can be a helpful solution to this dilemma, also enabling the researcher to be able to make sense of what the evidence means theoretically and empirically.

Interpreting the findings

Deploying PT case studies as an adjunct tool can enable better interpretation of the findings of the experiment itself. As regards interpreting findings, the use of either superficial forms of process tracing, in which certain observable traces are collected for all units, or more in-depth process tracing of a select number of units that involves tracing how the treatment actually worked/did not work. It is important to note the critical methodological differences between the experiment and PT. Whereas the experiment compares ATE between treatment and control groups, and the data that enables inferences to be made involves what can be termed 'experimentally manipulated data' that measures the difference that treatment/control make for the outcome across the groups, PT involves the collection of observational data of the traces left by the activities associated with parts of a causal process as it operated within a single unit of analysis. This means that the two methods are making very different types of inferences using very different types of empirical material.

There are a variety of analytical tasks that PT in parallel with an experiment can help deal with. These include: 1) evaluating causal heterogeneity and making sense of the ATE, 2) providing a stronger theoretical grounding for how the treatment works, and

First, If there is causal heterogeneity in the treated sample, with large differences in the causal effect that X had on Y, process tracing of a strategic sample of individuals can help shed light on why there were these differences. Experimental designs enable inferences to be made about the average treatment effect (ATE) of X on Y within the sampled population (Morgan and Winship, 2007: ???). However, an ATE can mask significant differences in the treatment effect within sub-groups, both as regards direction of the effect and magnitude (Haarding and Seefeldt, 2013: 99-100; Deaton and Cartwright, 2018: 4, 10-13; Leamer, 2010). For instance, there can be significant outliers in treatment effects that pull the ATE in a particular direction [MORE AND BETTER HERE]. In effect, the validity of ATE isonly as good as the treated population. The experiment is usually blind to potential differences that can be produced by differences across individuals that might influence whether and how the treatment works. While certain sampling techniques are available to try to ensure a more representative sample in relation to the full population (Mutz, 2011; MORE), the core challenge remains. The methodological point here is that our experiment – by itself – does not enable us to understand why there are differences in net effects across groups.

Ideally, process tracing case studies would be done in parallel with the experiment on a range of cases with different causal effects in an attempt to see whether the mechanism(s) were different across the cases. If we for instance found different mechanisms in cases with large positive effects versus small negative effects, we would then want to assess what factors differentiate these two groups of cases, thereby also enabling us to understand the bounds within which one type of causal effect was present.

One scenario can be that a cause triggers multiple processes that are linked to the same outcome, but where each has a different effect (Steel, 2008: 68). For example, exercise (X) is

linked to weight (Y) through at least three different causal pathways; each of which has a different effect:

1) exercise -> calorie burning -> weight loss
2) exercise -> muscle-building -> weight gain
3) exercise -> psychological guilt relief -> eat more -> weight gain.

Depending on the individual and the type of exercise, there might be strong negative net effects (exercise->weight loss) because the first pathways dominates the others, whereas in other individuals and types of exercise, the second and third pathways might dominate, resulting in significant weight gain.

Second, an experiment only sheds light on the effect of controlled manipulation of the treatment *within* the sampled population (Deaton and Cartwright; 2018: 4; Harding and Seefeldt, 2013: 98-99; ???). In effect, an experiment is a black-box into which a treatment goes in and emerges in the form of ATE. The experiment does not shed light on *why* the treatment worked or not (Clarke et al, 2014; MORE REF HERE). To understand how a treatment works (or does not work), we need to open up the black-box by tracing the causal processes that link causes and outcomes together, or trace them until they break-down to understand why the treatment did not work (Anderson, 2011: 421-2).

Finally, the experiment does not provide us with data upon which to assess the external validity of the findings. To quote Clarke et al (2014: 348), 'In sum, it is far from obvious whether a treatment will be efficacious outside the population in which it has been tested, or whether a successful policy action will be as good in a different context. No matter how well RCTs are designed and implemented, they do not on their own allow one to establish external validity. Evidence of mechanisms supplies information crucial to setting up the study and deciding how to adjust a policy action for a different population.'

By getting closer to the process itself, process tracing case studies are vital tools for the exploration of external validity (Aronson, 2018: 1172; Clarke et al, 2014:346-7; Khosrowi, 2019). The key point is that theorizing a causal mechanism also involves thinking about the

particular capacities required for entities (actors) to engage in activity which are the product of the context within which the process takes place. The ability of actors to engage in activities is sensitive to *context* (Sayer, 2000; Falleti and Lynch, 2009; Illari, 2011; Steel, 2008). Context can therefore relate to anything in a case that impacts on the ability of entities to engage in particular activities. For example, a policy expert might be widely perceived to be an epistemic authority in context A, enabling the recommendations to shape policy discussions, whereas in context B, the same type of policy expert might be perceived more as a partisan actor, reducing the persuasiveness of the same type of speech act. Here this factor would delimit the bounds within which the given process theory could travel.

A real world example of the importance of context can be found in White (2009). He describes a causal process that links a policy intervention (cause = education of mothers in nutrition) with an outcome (improved nutritional outcomes for children) that was found to have worked in a case (the Tamil Nadu Integrated Nutrition Project in India). The unpacked mechanism can be described as: Cause (mother participates in program) → 1) mother receives nutritional counselling → 2) exposure results in knowledge acquisition → 3) knowledge used to change child nutrition → Outcome (improved nutritional outcomes) (based on White, 2009: 4-5). Based on the success of the program in the Tamil Nadu case in India, it was then attempted to use the same policy intervention in Bangladesh. However, the process did not function as expected in the different context; instead it broke down. The reason for this was a key contextual difference. In Bangladesh, mothers were not the key decision-makers in households, with men doing the shopping, and mother-in-laws' in joint households (sizeable minority) acting as decision-makers about what food went onto the table. The process therefore 'worked' until part 3, but because of a contextual difference, it broke down in the Bangladesh case.

4. Conclusions

[To be written]

References

Baetu, Tudor M. 2016. 'From interventions to mechanistic explanations.' *Sythese,* 193: 3311-3327.

Cartwright, N.. 2012. Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps. *Philosophy of Science* 79(5): 973-989.

Clarke, B., D. Gillies, Phyllis Illari, Federica Russo, Jon Williamson. 2014. 'Mechanisms and the Evidence Hierarchy.' *Topoi*, 33(2): 339-360.

Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms: Discoveries Across The Life Sciences*. Chicago; London: University of Chicago Press.

Craver, Carl F, and David M Kaplan. 2020. "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations." *The British Journal for the Philosophy of Science* 71: 287–319. https://doi.org/10.1093/bjps/axy015.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach.* Cambridge: Cambridge University Press.

Evans, 2008). , J.. 2008. 'Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology* 59: 255-278.

Gentner, Dedre, Jeffrey Loewenstein, Leigh Thompson, and Kenneth D. Forbus. 2009. 'Reviving Inert Knowledge: Analogical Abstraction Supports Relational Retrieval of Past Events.' *Cognitive Science*, 33: 1343–82.

Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. 'Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose.' *Annals of the American Academy of Political and Social Sciences*. 628(1): 200-208.

Hauser, David J., Phoebe C. Ellsworth and Richard Gonzalez. 2018. Are Manipulation Checks Necessary? *Frontiers in Psychology.* 9(Article 998): 1-10.

Illari, Phyllis McKay and Jon Williamson. 2011. Mechanisms are real and local. In *Causality in the Sciences*, edited by Phyllis McKay Illari, Federica Russo and Jon Williamson. Oxford: Oxford University Press, 818-844.

Illari, Phyllis and Jon Williamson. 2013. 'In Defense of Activities.' *Journal for General Philosophy of Science*, 44 (1): 69-83.

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.

Leamer, Edward E. 2010. 'Tantalus on the Road to Asymptopia.' *Journal of Economic Perspectives*, 24(2): 31-46.

Machamer, Peter. 2004. 'Activities and Causation: The Metaphysics and Epistemology of Mechanisms.' *International Studies in the Philosophy of Science*, 18(1): 27-39.

Machamer, Peter, Lindley Darden and Carl F. Craver. 2000. 'Thinking about Mechanisms.' *Philosophy of Science*, 67(1): 1-25.

Morten, R. and K. C. Williams. 2010. *Experimental Political Science and the Study of Causality.* Cambridge: Cambridge University Press.

Pearl, J.. 2000. *Causality: Models, Reasoning and Inference.* Cambridge: Cambridge University Press.

Piccinini, Gualtiero. 2017. 'Activities Are Manifestations of Causal Powers.', in M.P. Adams, Z. Biener, U. Feest and J. A. Sullivan (eds) *Eppur si muove: Doing History and Philosophy of Science with Peter Machamer.* Cham, Switzerland: Springer, pp. 171-182.

Russo, Federica, and Jon Williamson. 2011. 'Generic versus single-case causality: the case of autopsy.' *European Journal of the Philosophy of Science,* 1 (1): 47–69.

Sagi, E., D. Gentner, and A. Lovett. 2012. 'What Difference Reveals About Similarity.' *Cognitive Science 36,* (6): 1019–50.

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

Waldner, D.. 2015. Process Tracing and Qualitative Causal Inference. *Security Studies,* 24(2): 239-250.