The Version of record of this manuscript has been published and is available in the *Journal of Economic Methodology* (2021) https://www.tandfonline.com/doi/10.1080/1350178X.2021.1952290

What's (Successful) Extrapolation?

Donal Khosrowi

Institute of Philosophy, Leibniz University Hannover

donal.khosrowi@philos.uni-hannover.de

Abstract

Extrapolating causal effects is becoming an increasingly important kind of inference in Evidence-Based Policy, development economics, and microeconometrics more generally. While several strategies have been proposed to aid with extrapolation, the existing methodological literature has left our understanding of what extrapolation consists of and what constitutes successful extrapolation underdeveloped. This paper addresses this lack in understanding by offering a novel account of successful extrapolation. Building on existing contributions pertaining to the challenges involved in extrapolation, this more nuanced and comprehensive account seeks to provide tools that facilitate the scrutiny of specific extrapolative inferences and general strategies for extrapolation. Offering such resources is important especially in view of the increasing amounts of real-world decision-making in policy, development, and beyond that involve extrapolation.

JEL: B40, C10, C90, C53

Keywords: Extrapolation, Causal inference, Extrapolator's circle, Evidence-Based Policy, development economics

1 Introduction

Extrapolation means various things: one can extrapolate a trend beyond the support afforded by existing data or extrapolate a claim that holds for some domain to other, novel domains. Here, I focus on a particular type of extrapolation: extrapolating causal effects measured in a study population to a distinct target population. This type of extrapolation is increasingly common in Evidence-Based Policy, development economics, and microeconometrics more generally, where researchers and analysts often use experimental and quasi-experimental methods to estimate the causal effects of policy or development interventions to determine their effectiveness[1]. The evidence of causal effects generated by these methods (henceforth *effect evidence*) is supposed

[1] I understand an intervention being 'effective' to mean that it achieves some desired effect (qualitatively or quantitatively). By 'effect evidence', I mean evidence concerning the magnitude of the causal effect of an intervention, such as produced by a randomized controlled trial.

to help decision-makers clarify which interventions ‘work’, including in novel target populations of interest to them. Yet, as is well-known, study and target populations can differ in relevant ways, so it is usually implausible to assume that interventions will be similarly effective in new environments as in those studied so far (Reiss 2019; Vivalt 2020). To bridge the gap between study and target, one needs to engage in extrapolative inference, which involves clarifying whether populations exhibit relevant similarities, and predicting how any differences bear on the effects to be expected in a target.

A substantial literature elaborates the problems encountered in extrapolation and proposes strategies for overcoming them (Shadish et al. 2002; Hotz et al. 2005; Steel 2009; 2010; Cartwright 2013a; 2013b; Bareinboim & Pearl 2012; 2016). Others have criticized these strategies in turn (Muller 2013; 2014; 2015; [reference blinded]). Despite these advances, two basic questions have received surprisingly little attention thus far. First, what does extrapolation of causal effects consist of, at the most general level? Second, what constitutes *successful* extrapolation? Making progress on these questions is important because it can help promote the critical appraisal of general strategies for extrapolation as well as specific inferences. Since extrapolation is routinely used to underwrite decisions about implementing policy and other interventions, which often involves significant epistemic risks, it seems important to develop tools that help us scrutinize both an increasingly common type of inference and the increasing number of real-world decisions that it grounds.

In this paper, I aim to make progress on these methodologically important issues. *Section 2* sketches a working analysis of extrapolation, which helps highlight several important dimensions along which problems of extrapolation and extrapolative inferences can vary. Distinguishing these dimensions allows us to recognize that some problems and some kinds of inferences are substantially more challenging than others, posing distinct obstacles for existing strategies aimed at addressing such problems. *Section 3* elaborates two crucial challenges in the way of successful extrapolation. Building on Steel's (2009) *extrapolator's circle*, I argue that strategies for extrapolation 1) must avoid being empirically overdemanding with regard to the supplementary resources they require to support their assumptions and 2) need to ensure that the effect evidence we wish to extrapolate from remains relevant to an envisioned conclusion rather than being displaced by these supplementary resources. *Section 4* offers several refinements to Steel's *extrapolator's circle*, yielding a more general challenge, which I call the *extrapolator's bind*. With this challenge articulated, I then propose an account of successful extrapolation that accommodates the nuances presented by this challenge. *Section 5* outlines how my account provides important tools for the appraisal of strategies for extrapolation as well as specific inferences. *Section 6* concludes.

2 Extrapolation: Many Problems, Many Inferences

There are surprisingly few explicit attempts in the literature to characterize extrapolation. The literature on *external validity*[2], for instance, (e.g. Campbell and Stanley 1963; Shadish et al. 2002) has made important progress in elaborating *threats* to our ability to generalize experimental results to novel settings and offers detailed recommendations for how study design can keep such threats at bay. But it has also treated external validity largely as a property of study designs themselves or particular estimates (Deaton and Cartwright 2018, 87-88) and has said little on how *inference* to

[2] Understood here as "[…] the extent to which [an] effect holds over variations in persons, settings, treatments, or outcomes" (Shadish et al. 2002, 22)

new populations proceeds, particularly in cases where important differences between populations remain. These issues have been addressed extensively by others (e.g. Hotz et al. 2005; Steel 2009; 2010; Cartwright 2013a; 2013b; Bareinboim & Pearl 2012; 2016; Muller 2013; 2014; 2015), but while plenty of details have been furnished to spell out strategies for extrapolation, there has been little in the way of a general, systematic treatment of the nature of extrapolation and its success conditions. For instance, building on some typical examples, Steel proposes the following broad characterization: 'In each of these cases, one begins with some knowledge of a causal relationship in one population, and endeavours to reliably draw a conclusion concerning the relationship in a distinct population.' (Steel 2009, 3). Yet, while this tells us what extrapolation aims to achieve, at the most general level, it does not provide systematic details on what ingredients are needed and how they must work together in enabling successful inferences[3].

To characterize extrapolation in more detail, it is useful to distinguish between problems of extrapolation and extrapolative inferences. Problems come first and are characterized primarily by how the populations of interest are constituted causally and how their causal features relate to one another in terms of similarity and difference. Inferences come second and are supposed to overcome the problems so defined. Let me expand on each in turn.

A problem of extrapolation consists of two populations A and B where a causal effect learned in A shall be used to infer a causal effect in B.[4] The crucial challenge is that A and B might differ in causally relevant ways, so it is important to learn in which

[3] However, in developing his extrapolation strategy in later chapters, Steel makes important progress in characterizing a general stricture on successful extrapolation – this will be discussed in Section 3.

[4] A and B might be entirely distinct populations, partly overlapping, or might consist of the same individuals at different times.

respects they are similar or different, and how these similarities and differences matter. We can think of these similarities and differences as part of a relation $\boldsymbol{R}$ obtaining between A and B.[5]

R has ontic and epistemic aspects. Focusing on its ontic aspects, we may understand R as comprising of an extensive list of causal features of A and B, e.g. whether causal relationships between particular variables, say X and Y, exist at all, what functional form these relationships take (e.g. linear, nonlinear), and what values and distributions causally relevant parameters and variables take in both populations (see [reference blinded]). Based on these features, R encodes particular sub-relationships of similarity and difference between individual features, such as that two variables X and Y are causally related in the same way in A and B or that the distribution of a certain variable Z differs in a certain way. We can see that R also has irreducibly epistemic aspects when considering that not all causal features matter for an extrapolation. So among the universe of A and B's features, R focuses on those that are relevant for answering a particular query. Nor must relevant features be captured in their full detail to promote our epistemic purposes: choosing a level of descriptive detail for features comprising R also involves epistemic interests. Finally, encoding similarities and differences between particular features involves non-trivial judgments that depend importantly on our epistemic interests, too: what is similar enough for one purpose, e.g. X being positively relevant for Y in A and B, might be too different for another, e.g. when X and Y are linearly related in A but quadratically in B. So, R is

[5] In particular, we may understand R as comprising of an extensive list of causal features of both A and B, including individual causal relationships between variables as well as finer-grained parametric and distributional details. Based on these facts, R can encode particular sub-relationships of similarity and difference between individual features, such as that two variables X and Y are causally related in the same way in A and B or that the distribution of a certain variable Z differs in a certain way. Of course, encoding similarities and differences requires non-trivial judgments that depend importantly on our epistemic interests, which is why R also has an epistemic nature.

best thought of as an epistemically pertinent abstraction of the true, underlying causal makeup of A and B.

To illustrate, consider a toy case from microfinance where the production of household welfare Y through microfinance access X is mediated by households' investments in durable goods Z. Even if the basic structure of the causal mechanism[6] $X \rightarrow Z \rightarrow Y$ is the same in A and B, Z might play different causal roles, e.g. when individuals' tendency to invest is higher in A than in B. If we are interested in predicting the X-Y-effect in B, R should hence focus on similarities and differences concerning certain causal features, e.g. the functional form of the relationships $X \rightarrow Z$ and $Z \rightarrow Y$, rather than, say, on other, orthogonal causes of Y that might differ between A and B. The similarities and differences captured by R are hence partly provided by the causal makeup of A and B and partly selected, represented, and interpreted according to our epistemic interests.

With R in place, we can now characterize extrapolative *inference* in more detail, including what role R plays in it. Extrapolative inference begins with effect evidence pertaining to the causal effect of an intervention in A, such as an average treatment effect estimated in a randomized controlled trial (RCT) or a credible quasi-experimental study, and endeavours to infer a causal conclusion about the effect of the same or a similar intervention in a novel population B. Such an inference will usually involve various assumptions about R. For instance, assume from our previous example that a causal mechanism governing a microfinance effect in A is learnt to be $X \rightarrow Z \rightarrow Y$ and that the $X \rightarrow Z$ relationship is moderated by a individuals' age M;

[6] I understand mechanisms loosely here, in the spirit of minimalist accounts such as Illari and Williamson's (2012), as consisting primarily of causes and relationships between them that can have more specific functional forms. I use upper-case symbols for causal variables [X, Y, Z, etc.] and arrows to denote causal relationships.

higher values of M induce larger marginal effects of X on Z, since older individuals tend to invest more in durable goods. Assume further that A and B differ in their distribution of M, so our interest is in predicting the X-Y effect in B, given B's particular distribution of M. In performing such an extrapolation, we might need to assume that A and B are similar in certain respects, such as that the two causal relationships $X \rightarrow Z$ and $Z \rightarrow Y$ are instantiated in B, that the functional form of these relationships is similar in A and B, that the mediating variable Z is not bypassed by any counteracting causal pathways in B, etc. ([reference blinded]).

It is not enough, of course, to *merely* assume these things about R. For an extrapolation to be justified, we must *support* these assumptions; usually by invoking some combination of background knowledge and, more importantly, supplementary empirical evidence that, together, help clarify R's details. While background knowledge comprises general theoretical or empirical resources relevant to clarifying R's details, supplementary empirical evidence comprises evidential resources specifically sought out to clarify particular causal details of study and target populations. For example, a middle-range background theory of the conditions under which microfinance interventions can alleviate credit constraints faced by households of the rural poor can elucidate what general characteristics of a population favor the effectiveness of microfinance interventions. Supplementary empirical evidence, on the other hand, could help clarify whether behavioral response to capital in a particular target might involve undesirable substitution effects, or whether target households' demand elasticity for durable goods is sufficiently high to promote welfare-improving investments. Since pertinent background knowledge and theory are often scant or insufficiently detailed, I assume that empirical evidence usually plays a more important role in supporting extrapolation (though see Cartwright 2020 for a more optimistic outlook on the role of theory).

With these ideas in place, let me sketch a working analysis of extrapolation detailing the roles played by the different ingredients outlined so far:

> ***EXT***: extrapolation is an inference I that uses effect evidence E obtained from a study population A to infer a conclusion C about a target population B, with the help of assumptions P pertaining to the relation R between A and B, as well as background knowledge K and supplementary evidence S that help support P.

This analysis can help refine our understanding of extrapolation in several important ways. As a first step, it helps us recognize that extrapolation is a highly heterogeneous family of inferences that can differ significantly across several of the variables figuring in ***EXT***. For one, there can be important differences in R. Two populations might differ minimally, e.g. when the distributions of some causally relevant variable, e.g. age, differ slightly. Differences can be more dramatic, however, when the basic structure of the causal mechanisms governing the outcomes of interest is entirely dissimilar, e.g. when X is causally relevant for Y in A but not at all in B. Intuitively, some problems of extrapolation are more difficult to overcome than others depending on how R is constituted.

Further important differences can arise at C, P, K, and S. First, the nature of the conclusion sought can differ importantly, as extrapolations may seek to address a wide range of queries, including:

1) Will an intervention have *some* effect in B if it does so in A?

2) Will an intervention have *a similar/the same* effect in B as in A?

3) What is the magnitude of the causal effect of an intervention in B if it is such-and-such in A, and taking into account any differences between A and B?

Differences in the type of conclusion pursued often entail a whole array of further differences concerning what assumptions P are needed for an inference. For instance, an inference that enables us to answer 3) will generally need to involve stronger assumptions than those seeking to answer 1) or 2).

Stronger and more extensive assumptions, in turn, will often demand more extensive or more fine-grained support S and K. For instance, it might be relatively easy to support that a certain qualitative relationship between two variables X and Z holds in A and B. However, ascertaining the quantitative similarity of this relationship, e.g. that there are similar marginal causal effects of X on Z, might be more challenging. So, the appropriate character of S and K will depend both on the kinds of assumptions required to infer a certain type of conclusion and on the amount of support needed to adequately underwrite these assumptions. In sum, extrapolations can differ importantly in the conclusions they aim for, in the assumptions they require to license these conclusions, and in the support needed to justify these assumptions.

These insights apply not only to specific inferences, but also to whole *strategies* for extrapolation, which can differ significantly in the *kinds* of inferences they can enable, and hence also in the kind and extent of support they routinely demand. For instance, Cartwright's *Argument Theory of Evidence* (2013a) maintains that conscientious extrapolation of causal effects from RCTs is best understood as proceeding by means of valid and sound *effectiveness arguments*, i.e. arguments yielding a conclusion about a target by drawing on an RCT result as well as additional premises asserting crucial similarities between populations. The arguments used to illustrate this strategy often focus on establishing rather modest conclusions, such as that an intervention can be expected to have some non-zero effect for at least some individuals in a target (2013a:14). Similarly, Steel's (2009) *Comparative Process Tracing* strategy aims to extrapolate claims of qualitative causal relevance by using evidence of qualitative downstream similarities in mechanisms between populations.

Other strategies are more ambitious. For instance, Hotz et al.'s (2005) covariate-based strategy aims to enable predictions of causal effects when populations exhibit differences in the distributions of causally relevant variables. For instance, the magnitude of a causal effect of X on Y might depend on individual's age M, and A and B might differ in their distribution of M, so the X-Yeffect in the target would differ from that in the study. In such cases, Hotz et al.'s approach models how the effect of interest depends statistically on M and reweights the effect from A by the observed M-distribution in B. Similarly, Bareinboim and Pearl's causal graph-based approach (2012; 2016) also seeks to enable conclusions about quantitative effects under conditions where populations differ in relevant ways. It does so by means of a sophisticated algorithmic procedure to derive so-called *transport-formulae*, i.e. expressions that help 'shield away' unimportant differences between populations and accommodate any remaining differences by means of reweighting.

In licencing more ambitious inferences, these two approaches require a host of stronger and finer-grained assumptions. For instance, they not only have to assume that the causal mechanisms governing the outcomes of interest have the same structure in both populations (see Hyttinen et al. 2015), but also that particular causal relationships are similar up to the level of functional form and structural parameters (see [reference blinded]) – otherwise precise quantitative conclusions are not permitted. These assumptions, in turn, are significantly more burdensome to support.

Knowledge of functional form and parameter values is rarely handed down by background theory, so empirical resources will usually have to supply the information needed to get extrapolative inference off the ground. This is where extrapolation faces serious empirical challenges, but also where existing strategies say little on how to overcome them. They clarify the abstract conditions under which certain extrapolative inferences are valid, but they do not tell us how to make any inferences so enabled sound. Successful extrapolation, however, is not achieved by supplying valid inference templates alone; it is achieved only when such templates are joined by resources that make their assumptions plausible. Yet, acquiring such support is not only often challenging, but also raises deeper problems that call into question whether a large class of inferences sanctioned by existing strategies can in fact be successful in practice. Let me expand on this concern.

3 Challenges And Strictures

While the strategies for extrapolation outlined above can provide valid and useful inference templates (or so I will grant), two important challenges remain in the way of successful extrapolation. These challenges arise from how much and what kinds of supplementary resources S and K are needed to support an inference.

To elucidate these challenges, it is useful to draw out three distinct aspects of an extrapolation:

1) The relationship R that holds between a study and target population
2) The assumptions P about this relationship that are required for an extrapolation

3) The part of P that is (or can be) in fact supported by supplementary resources S and K.

Ideally, S and K will be jointly sufficient to fully support P. The more S and K manage to support P, other things being equal, the more likely it is that we can draw an accurate conclusion. At the same time, the more extensive the conjunction of S and K *must* be to adequately justify the assumptions P demanded by specific extrapolation strategies, the more empirically demanding such strategies are. In the limit, P would require us to assume all there is to assume about R, and to fully support P, S and K would need to encompass all there is to know about R. This presents us with two general challenges.

3.1 Overdemandingness

First, compelling strategies for extrapolation should not be epistemically overdemanding. For instance, if justified extrapolation would require us to learn the sub-personal neural-level microstructures underpinning individual behaviors and social phenomena of interest, and issues of similarity and difference would need to be settled at these levels, including details on how a neural-level causal basis realizes these phenomena, this would be overdemanding and undesirable.

More generally (and bracketing the role of K), overdemandingness concerns cases where supplementary evidence S is needed to support the assumptions P required for extrapolation, but where acquiring this evidence is extremely costly, difficult, or even impossible, such as measuring individual causal effects (which are typically considered unobservable magnitudes, cf. Rubin 1974; Holland 1986) or other causal features that cannot (principally or realistically) be learned from observational procedures, such as agents' dispositions to respond to interventions never before experienced.

This concern, by itself, could seem to have rather little bite, however, as one may respond that (at least some) strategies for extrapolation were perhaps never intended to overcome concrete problems of extrapolation *on their own*. For instance, Cartwright's (2013a) *Argument Theory* does not aim to provide recipes for extrapolation from start to finish but primarily seeks to characterize a general constraint: it should be possible to cast extrapolative inferences in the form of valid (and sound) arguments.

More generally, we might argue, along the lines of Marcellesi (2015, 1315), that concerns about empirical demandingness are not immediately relevant to abstract extrapolation strategies, as these strategies only aim to specify the general conditions under which certain inferences are permitted. And while there may still be a pressing need for complementary *empirical* strategies to help us acquire what is required for supporting extrapolation, we should perhaps not expect both the general recipes as well as the concrete details for how to do the messy empirical work from a single, overarching strategy.

However, as I turn to argue now, even if there were off-the-shelf empirical strategies for obtaining the supplementary evidence required by existing strategies in a straightforward way, *using* this evidence faces a second important challenge, which suggests that the fault is indeed with the abstract strategies that demand such evidence in the first place.

3.2 The Extrapolator's Circle

This second challenge, called the *extrapolator's circle,* originally highlighted by LaFolette and Shanks (1996) and more recently elaborated by Steel (2009), adds more concrete strictures on how epistemically demanding a strategy for extrapolation may be. Specifically, a strategy for extrapolation should not require supplementary resources that would allow us to answer a causal query of interest based on these resources *alone*.

If extrapolation required such extensive resources, this would clearly be undesirable, either turning the problem of extrapolation into an altogether different inferential problem, e.g. reasoning from background knowledge and piecemeal causal information, including from the target, to the effects of some intervention there. Or we might say that, while we are still in the business of extrapolating, E is rendered redundant to our conclusion. Either way, falling prey to the extrapolator's circle would undermine much of the promise that Evidence-Based Policy and similar approaches in the 'treatment effects-literature' in development and empirical microeconomics hold, i.e. that causal effects learned in some population A can be informative for predicting the effects of the same or similar interventions in other populations B, C, D, etc. For instance, a prevailing hope in Evidence-Based Policy is that we can build so-called 'libraries of evidence' that collate the causal effects of different interventions, and thereby facilitate prediction of the effects of these and other, similar interventions in novel targets. Yet, if the only way to make use of such evidence were to learn so much about target populations that the carefully collated evidence became redundant to answering our questions, then why should we build evidence libraries at all?

This worry becomes more acute when acquiring supplementary evidence about the target requires some form of intervention there, rather than learning from observational data. Beyond involving the risk of harming agents (such as when an intervention that had strictly positive effects in A makes agents worse off in B), it poses an even greater risk of trivializing our inferences. In such cases, effect evidence from a study population would provide little epistemic value beyond telling us about the effects of interventions where they have already been implemented, at most giving us some hope that they *might* be effective elsewhere. But the justificatory burden in making predictions about any novel targets would be carried by evidence that is distinct and unrelated to the effect evidence that we started from.

4 Understanding Successful Extrapolation

Following Steel, I take the extrapolator's circle to be a crucial challenge that any compelling strategy for extrapolation must overcome, and preferably for a large class of cases.[7] Yet, there are also several ways in which this challenge can be detailed to improve our understanding of what successful extrapolation requires. Let me proceed to offer two general proposals that help concretize the underlying problem highlighted by the extrapolator's circle, and generalize it beyond the construal offered by Steel.

4.1 It's a Bind, Not a Circle

First, we should recognize that the extrapolator's circle is often not a circle proper. More specifically, the extrapolator's circle is triggered not only when we must already know C to infer C, suggesting some sort of full-fledged circularity, but also when S and K jointly permit inferring C beyond some threshold of sufficient confidence, thus

[7] Steel's own *Comparative Process Tracing* strategy indeed promises to evade this problem (but see Reiss 2010).

making E redundant to C. To capture this, I propose that we refer to this challenge as the *extrapolator's bind*.[8] The bind generalizes beyond the circle: it captures cases where C must be known to infer C, or C is trivially learned in the process, such as when we need to implement an intervention in the target to learn what its effects there will be. But it is also more general, in that it captures cases where other resources, such as S and K, displace the relevance of E for inferring C. The bind also captures two nuances at once: it characterizes a *problem*, first and foremost. But once the importance of this problem is recognized, the bind becomes *normative*: it is a binding stricture on what we may and may not do when aiming to extrapolate successfully.

My second proposal concerns the assumption that the extrapolator's circle is an all-or-nothing affair (Steel 2008, 78, 85, 86, 99). Both Steel's as well as LaFolette and Shanks' original formulation (1995, 157) suggest that it is triggered whenever we *know* the answer to our query based on evidence from the target alone, making the effect evidence dispensable. However, there can be gradual variations where effect evidence is rendered almost irrelevant to our conclusion, but not entirely.

We can think about the gradual nature of the extrapolator's bind in terms of the *degree of relevance* of E to C, which we can understand in terms of the sensitivity of C to changes in E, i.e. how much C changes over variations in E. For instance, we could ask how C would change when a positive causal effect measured in A were replaced by a negative one. The more sensitive C is to changes in E, other things being equal, the more C is informed by E, and hence the more *relevant* E is to C. Lower levels of sensitivity, on the other hand, suggest that E plays a less important role, and that C hinges relatively more on S and K.

[8] I am indebted to [name blinded] who has suggested this term to me.

There are two problems with this account of relevance, however. One relates to what is called the *weight* of evidence (see Peirce 1878; Keynes 1921). E can be relevant to C in at least two different ways: it can change the *content* of our conclusion, say from C to C', and it can provide more or less *support* for one and the same conclusion. The first is captured by the idea of sensitivity outlined above, i.e. the changes induced in a conclusion C as a response to changes in E. For instance, if E would change to indicate that an intervention had a negative, rather than a positive, effect in A, then C's content might change to assert that our intervention will have a negative rather than a positive effect in B, too. However, the weight for C that E affords also needs to be considered. Here, changing E (or subtracting or adding it from our evidence base) can change our confidence in C, though it does not change the content of C as such. For instance, we might have weak observational evidence S about B that an intervention on X could have a positive effect on Y there. Suppose now that we also obtain high-quality experimental evidence E that an intervention on X has a positive effect on Y in A as well as some further evidence S' to support that A and B are highly similar. Although adding the conjunction of E and S' to our evidence base does not change C's content, it does add significant weight.

More generally, if adding or subtracting E makes a larger difference to the weight in favour of C, then, other things being equal, the more relevant E is to C. Conversely, if E makes no difference to the weight of evidence for C, other things being equal, then it is irrelevant. This could be the case if S and K already warrant C beyond some relevant threshold of confidence α, so that E would not make a difference to whether we are sufficiently confident in C one way or another.[9]

[9] It might seem odd to say that E is irrelevant to C because there is additional evidence S invoked to infer C, but where S was learned after E. Yet, since I assume that E alone is insufficient by itself to

The way in which E changes C's content and the weight in its favour is likely to be interactive. Specifically, adding a token of evidence to our evidence base may not only change the content of C, or only the weight in its favour, but both. For instance, when E contravenes a conclusion that would have been reached by considering only evidence S from the target, then this may plausibly change both the content of our conclusion (say, that an effect is positive rather than zero) from C to C', and our confidence in this conclusion (there was previously no weight in favour of C').[10]

A second potential problem with understanding the extrapolator's bind in terms of relevance is that the very nature of the conclusion pursued can itself play an important role in determining the relevance of E to C. For instance, if C is highly general in nature, such as when asserting only that an intervention on X is causally relevant in *some* way for changing Y in B, then this may itself bear importantly on how relevant E is to C. Specifically, the more general the desired conclusion, the less relevant E will be to it in terms of potentially changing its content, other things being equal. This could seem counterintuitive. How relevant E is to C would not seem to be driven by other sources of support S and K. Indeed, even these other sources of support would be rendered less relevant to C. So does this account of relevance lose track of the problem it is supposed to articulate?

No, these implications merely help us recognize that context matters for whether and how much we fall prey to the extrapolator's bind. If C is more easily reached, say because it is more general and hence less demanding to support, then this will, other

infer C with appropriate confidence – this is what serious problems of extrapolation are all about – whether or not S was available before or after E is immaterial to assessing whether S renders E irrelevant to C.

[10] Further details might be concretized in a Bayesian framework (e.g. Landes et al. 2018), but a formal treatment of evidential relevance is not essential to my goals here.

things being equal, make it more likely that we fall prey to the extrapolator's bind since less or weaker supplementary evidence S is required to reach C unaided by E. Keeping E and S constant, changing the nature of C so that it is easier to support will simply make it more likely that E is rendered redundant. This preserves the way that S and E compete for relevance to C, which is the key problem that the extrapolator's bind highlights.

Taking the above concerns into account, we can now formulate more precisely what it takes to extrapolate successfully: a strategy for extrapolation should steer clear of the extrapolator's bind *as best as it can*. Relevance, in turn, has two facets that need to be balanced. Relevance for C's content is important, but not if E provides little weight in favour of C. Similarly, weight is important, but not crucially so if E does not bear much on C's content and we could have inferred it regardless (although with less confidence). What balance is adequate will, of course, hinge on specifics about the case, so not much more can be said beyond emphasizing that both should be considered.

With these proposals in place, let me refine my working analysis to say more on what constitutes successful extrapolation.

4.2 From Extrapolation to Successful Extrapolation

Recall ***EXT***, which characterized extrapolation as an inference I that uses evidence E obtained from A to infer a conclusion C about a target B, with the help of assumptions P pertaining to R, as well as supplementary resources S and K that help support P.

We can now refine ***EXT*** by adding strictures on *successful* extrapolation: an extrapolation is successful *iff* the following four conditions hold[11]:

1) (***INFORMATIVENESS***) Some conclusion C of the desired kind is inferred.
2) (***JUSTIFICATION***) C is adequately justified, i.e. our confidence in C that is warranted by a combination of E, P, K, and S must exceed some threshold α.
3) (***ACCURACY***) C is accurate, relative to some threshold β.
4) (***RELEVANCE***) C is inferred in such a way that E remains relevant to C beyond some threshold γ.

INFORMATIVENESS ensures that C speaks to what we want to know about the target and not to some other, potentially related question. As outlined earlier, some strategies for extrapolation are more limited in what conclusions they can enable, so success depends on whether their abilities extend to cover those conclusions we are interested in.

JUSTIFICATION demands that C is not obtained by sheer luck, but that it enjoys sufficient support and is arrived at by means of a sound process. As the standards for what counts as sufficient justification are plausibly context-dependent (e.g. on the stakes involved in drawing mistaken conclusions), a threshold α can be used to capture the gradual nature of justification as well as how pragmatic considerations

[11] One might wonder why there is no fifth condition pertaining to the *validity* of the inference schema used. I will simply grant that the schemas supplied by existing strategies are either valid, or invalid (when inductive) but compelling when used properly.

inform how much confidence is needed to justify acceptance of a conclusion and subsequent action (e.g. intervention in the target).

Accuracy requires that what C asserts about the target is accurate with respect to what is (or will be) the case there.[12] It is not enough to have a well-justified conclusion that speaks to queries of interest to us, but that ultimately turns out to be radically mistaken. Accuracy will, of course, be context-dependent, too. It could mean that a causal effect predicted to be positive in the target indeed turns out to be positive, that C correctly predicts the exact magnitude of a causal effect, or that C correctly instructs us that a certain kind of co-intervention is needed to achieve a specific outcome distribution. So accuracy can come in different forms, and we might wish to spell out varying standards of accuracy β for a conclusion to be suitably accurate. Of course, it is also important to recognize that, unlike the other conditions, accuracy can only be determined after the fact. This does not pose any special problems for general assessments of whether strategies for extrapolation meet this desideratum, however. If they repeatedly and consistently fail to provide accurate conclusions, this can tell us something about how successful these strategies might be in future instances.

Finally, ***Relevance*** captures the extrapolator's bind. It is both a constitutive feature of extrapolation, distinguishing it from inductive inference more generally, and it is a (gradual) success condition. No relevance means no extrapolation, but while a little relevance means we are extrapolating, it may not be enough for success. The relevance condition maintains that successful extrapolation requires that effect evidence remains *sufficiently* relevant to our conclusion beyond some threshold γ. Of

[12] I assume that desiderata concerning the *precision* of the conclusion are captured by 1).

course, γ is merely a conceptual placeholder, and it might be difficult to operationalize relevance in such a way that measuring it and determining a meaningful threshold γ is practically feasible. But some threshold that is (perhaps significantly) above full-blown irrelevance of *E* to *C* seems desirable in real-world extrapolations. RCTs, for instance, do not usually come cheap, and if our best available extrapolation strategies would standardly require information that would make RCT evidence largely irrelevant to *C*, then this would seem highly undesirable.

Considered together, then, successful extrapolation requires that an extrapolative conclusion of the envisioned kind is reached, that it is justified to some sufficient degree, accurate to some sufficient degree, and that the effect evidence we are extrapolating from remains relevant to it to some sufficient degree. Let me expand more broadly on how this refined analysis can help us assess the success and failure of specific inferences and whole strategies for extrapolation.

5 Success and Failure

Failure in extrapolation can come in different forms. A failure of 1)-3) might be called a failure to extrapolate successfully (to different degrees, and potentially with different weights). A complete failure of 4), however, may lead to a more undesirable conclusion. We might say that not only does an extrapolation fail, but, particularly when 1)-3) are indeed satisfied, there is a special kind of failure occurring: one fails not only to extrapolate successfully, but one fails to extrapolate *at all*, since *E* does not relevantly figure in inferring *C* anymore.

What does my analysis tell us about the success of *strategies* for extrapolation? This might seem unclear, as what it provides is first and foremost a clarification of what it means to successfully extrapolate in concrete instances. The answer is that the analysis works bottom-up, and in a context-sensitive way. It begins from single instances of extrapolation and, based on how specific strategies handle such instances or indeed whole types of extrapolation, proceeds towards more general conclusions about the success of these strategies. This approach is advantageous as it seems likely that most strategies will be able to achieve some instances of successful extrapolation (and perhaps consistently for certain kinds), but fail in other (kinds of) cases. So success is piecemeal, likely to be heterogeneous, and one failed instance of extrapolation does not make for an entirely failed strategy. But if the circumstances under which specific strategies for extrapolation are prone to failure are important, systematic, and general enough, then this can nevertheless license broad conclusions about how successful they are as strategies.

For instance, Bareinboim and Pearl's graph-based strategy is not only the most ambitious candidate among existing strategies but also the most epistemically demanding. It generally requires that analysts are in possession of sufficient causal/mechanistic knowledge to assert that two populations can be represented by the same causal graph (Bareinboim & Pearl 2016, 7351 fn.) However, it is exceedingly unlikely that such knowledge is routinely available in real-world settings (Hyttinen et al. 2015). Moreover, even if it were, such knowledge might often be sufficient to independently learn the causal effect of interest, thus undermining the success of a potentially large body of inferences enabled by the strategy.

Importantly, the analysis provided here allows us to *predict* the success of extrapolative inferences, as well as of whole strategies more generally. While it is of course hardly possible to predict how accurate an extrapolative conclusion is before learning the effect to be predicted in a target, it is often possible to predict how successful an inference will be with regard to relevance. Specifically, if a strategy, in virtue of the assumptions it makes and the support they require, demands extensive support S and K that would clearly render E irrelevant to C (e.g. detailed causal knowledge from the target that would permit inferring C independently), then, even before engaging in specific inferences and acquiring the resources needed, we can tell that these inferences will fail to be successful. Similarly, if it is foreseeable that a whole strategy is bound to yield unsuccessful extrapolations in virtually every instance (of a kind of extrapolation), we may conclude that it is entirely unsuccessful (for this kind of extrapolation).

The analysis developed here also draws out crucial dynamics bearing on the success of extrapolation more generally, irrespective of the particular strategy employed. As the discussion has made clear, E, S, and K compete for relevance for C. But there are also important tensions between the success conditions governing this competition. In particular, there is a trade-off between, on the one hand, how precise and detailed the conclusion is that we aim for, the envisioned accuracy of that conclusion, and the justification required for reaching it, and, on the other hand, the relevance of E to C. Aiming for more detailed, precise, and accurate conclusions C almost always requires more extensive justification by means of S and K. Yet, adding more and more supplementary resources, in particular resources that say something about the effect of interest in the target independently, risks displacing the relevance of E to C. Recognizing such tensions is important when considering what type of conclusion to aim for, what strategy to adopt, and for deciding whether we should aim to learn an effect by means of extrapolation at all or perhaps by some other way.

Finally, let me emphasise how elaborating the role that relevance plays in extrapolation not only makes conceptual and theoretical progress, but also has bearing on practice. Consider again one of the key promises of Evidence-Based Policy initiatives, which is that building evidence libraries collating information on the effectiveness of different interventions in addressing common policy issues is a key step in facilitating the implementation of better, more effective policy. If a large class of extrapolations that could proceed from the evidence collated would routinely make that evidence redundant, we might need to reconsider some of the basic principles of how evidence-production and -use are organized. These principles might be doubly unhelpful: as long as methodological guidelines say little on the intricacies involved in extrapolation, they continue to invite attempts at naïve extrapolation that simply, but wrongly, assume that an effect established somewhere can be straightforwardly expected in novel populations (Cartwright 2013a). At the same time, by failing to take concerns about what makes extrapolation successful into account, they may also overestimate just how much work the evidence collated in libraries can do for us in speaking relevantly to the effectiveness of interventions in novel targets. Recognizing that conscientious efforts at extrapolation face intrinsic tensions when it comes to maintaining the relevance of evidence for what we want to learn raises important practical concerns about how to allocate scarce resources to our inferential and practical ends. Extrapolation is not the only way to predict causal effects in novel populations. At least in those cases where successful extrapolation in regard to relevance seems unlikely, we are perhaps better off pursuing other strategies from the get-go, such as investing more heavily in building middle-range theories of the phenonema of interest and program theories elucidating how specific interventions achieve their intended effects across important variations between settings (Astbury and Leeuw 2010; Pawson 2013; Cartwright 2020). This option can seem attractive, as stronger theory can not only help us predict how effective a given intervention will be in a novel target, but also help us adapt and tailor interventions to the specific circumstances encountered there.

6 Conclusions

The arguments provided here allow us to make progress in spelling out desiderata for existing and future strategies for extrapolating causal effects. Such strategies should help us learn from effect evidence E and a conjunction of assumptions P, background knowledge K, and supplementary evidence S pertaining to the relation R between populations, to reach an action-guiding, ampliative conclusion C about the causal effects of interest, where *ampliative* means that the conclusion should go beyond what we can already infer about the target on grounds of S and K alone. Existing strategies are liable to fall well short of this goal, however: they are often empirically overdemanding, and likely to fall prey to the extrapolator's bind in virtue of the extensive empirical assumptions they involve.

This should not come as a surprise. In the limit, accurately predicting a causal effect in a target despite potential causally relevant differences might require one to know about all relevant differences and similarities, as any unknown difference may curtail accurate extrapolation. Yet, because some of these differences and similarities are at least extremely difficult to learn without falling prey to the extrapolator's bind, extrapolation with certainty will not only remain an elusive ideal, but is also wholly undesirable, since it is bound to render our effect evidence redundant to predicting the effects we are interested in.

Moving from the ideal to the practically feasible, it is clear that any compelling strategy for extrapolation must stop well short of these extensive requirements. It is also clear that any such strategy must tell us a rich story about the inferential leap that will, necessarily, persist between what we (must) learn for the purpose of successful extrapolation and what we are interested in inferring. Existing strategies provide accounts of the conditions under which accurate prediction of causal effects is possible in principle and what assumptions are needed for this, but they do not tell us how to support these assumptions in a way that helps us to *successfully* extrapolate, i.e. extrapolate without falling prey to the extrapolator's bind. The tools offered here, I hope, can help us make progress in refining our understanding of the challenges that remain in the way of success.

Acknowledgements

I wish to thank Nancy Cartwright, Wendy Parker, Finola Finn, Lukas Beck, the audience at INEM2019 Helsinki, and two anonymous reviewers for their helpful comments, suggestions, and criticisms.

References

Astbury, B., and F. Leeuw. (2010). "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation." *American Journal of Evaluation* 31(3):363-81.

Bareinboim, E. and J. Pearl. (2012). "Transportability of causal effects: Completeness results", In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

— **(2016)**. "Causal inference and the data-fusion problem", *Proceedings of the National Academy of Sciences*, 113:7345-52.

Campbell, D., and J. Stanley. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cartwright, N. D. (2013a). "Knowing What We Are Talking About: Why Evidence Doesn't Always Travel". *Evidence and Policy: a Journal of Research, Debate and Practice*, 9(1):97-112.

— **(2013b).** "Evidence, Argument and Prediction". In: V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham, Switzerland: Springer International Publishing Switzerland.

— **(2020).** "Lullius Lectures 2018: Mid-level theory: Without it what could anyone do?" In C. Martínez Vidal and C. Saborido (eds.) *Nancy Cartwright's Philosophy of Science*, Special Issue of *Theoria*.

Deaton, A., and N. D. Cartwright. (2018). "Reflections on Randomized Control Trials". *Social Science & Medicine*, 210:86-90.

Holland, P. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association,* 81(396):945-60.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer. (2005). "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*. 125:241–70.

Hyttinen, A., F. Eberhardt, and M. Järvisalo. (2015). "Do-calculus when the true graph is unknown", in: UAI'15 Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 395-404.

Illari, P. M., and J. Williamson. (2012). "What is a Mechanism?: Thinking about Mechanisms Across the Sciences", *European Journal for Philosophy of Science*, 2:119–35.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

LaFollette, H., and N. Shanks. (1996). *Brute Science: Dilemmas of Animal Experimentation*. New York: Routledge.

Landes, J., B. Osimani, and R. Poellinger (2018). "Epistemology of Causal Inference in Pharmacology. Towards a Framework for the Assessment of Harms", *European Journal for Philosophy of Science*, 8(1):3-49.

Marcellesi, A. (2015). "External Validity: Is There Still a Problem?". *Philosophy of Science*, 82(5):1308-17.

Muller, S. M. (2013). "External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

— **(2014)**. "Randomised trials for policy: a review of the external validity of treatment effects", Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **(2015)**. "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1):217-225.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. London: SAGE.

Peirce, C. S. (1878). "The Probability of Induction", *The Popular Science Monthly, Illustrations of the Logic of Science*, XII.

Reiss, J. (2010). "Review: Across the boundaries: Extrapolation in biology and social science." *Economics and Philosophy*: 26:382-390

— **(2019).** "Against external validity." *Synthese* 196(8):3103-21.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology,* 66(5):688–701.

Shadish, W. R., T. D. Cook, and D. T. Campbell. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Steel, D. (2009). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

— **(2010).** "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs", *Philosophy of Science*, 77(5):1058-69.

Vivalt, E. (2020). "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association* (online first).