Predicting 'It Will Work for Us': (Way) Beyond Statistics

Nancy Cartwright
LSE and UCSD

1. Introduction

The topic of this paper is 'external validity' and its problems. The discussion will be confined to a special class of conclusions: causal conclusions drawn from statistical studies whose fundamental logic depends on JS Mill's method of difference. These include randomized control trials (RCTs), case control studies and cohort studies.

These kinds of studies aim to establish conclusions of the form 'Treatment T causes outcome O' by finding a difference in the probability (or mean value) of O between two groups, commonly called the 'treatment' and the 'control' groups.[1] Given the method-of-difference idea, in order for the causal conclusion to be justified the two groups must have the same distribution of causal factors for O except T itself and its downstream effects. The underlying supposition is that differences in probabilities require a causal explanation; if the distribution of causes in the two groups is the same but for T yet the probability of O differs between them, the only possible explanation is that T causes O. The studies differ by how they go about trying to ensure as best possible that the two study groups do have the same distribution for causal factors other than T. There are, as we know, heated debates about the importance of randomization in this regard but these debates are tangential to my topic.

I want to separate issues in order to focus on a question of *use*. Suppose, contrary to realistic fact, that we could be completely satisfied that the two groups had identical distributions for the other factors causally relevant to O. I shall call this an *ideal* Mill's method-of-difference study. What is the form of the conclusion that can be drawn from that and of what use is it? In particular of what use is it in predicting whether T will cause O, or produce an improvement in the probability or mean of O, 'for us' – in a population we are concerned with, implemented as it may be implemented there?

The basic problem is that the kinds of conclusions that are properly warranted by the method-of-difference design are conclusions confined to the population in the study. That is seldom, indeed almost never, the population that we want to know about.[2] A difference in the probability of the outcome in this kind of study can at best establish what I call 'it-works-somewhere' claims and the *somewhere* is never where we aim to make further predictions. We want to know, 'Will it work for us in our target population as it would be implemented there?' This questions often goes under the label of an 'effectiveness' claim. I call it more perspicuously an 'it-will-work-for-us' claim. The problem of how to move from an it-works-somewhere claim to an it-will-work-for-us claim usually goes under the label 'external validity' and is loosely expressed as the question 'Under what conditions can the conclusion established in a study be applied to other populations?'

In this paper I shall argue for two claims: a negative claim that external validity is the wrong idea and a positive claim that what I call 'capacities' and Mill called

[1] Naturally only a difference in frequency is observed. There is thus a preliminary question of statistical inference: what probabilities to infer from the observed frequencies. I set this question aside here because I want to focus on the issue of *causal* inference.

[2] Even if the entire target population were enrolled in the study, predictions will be about future effectiveness where there may be no guarantee that this population stays the same over time with respect to the causally relevant factors.

'tendencies' are almost always the only right idea. The currently popular solution to the problem of external validity from philosophers and statisticians alike is to study the 'invariance' characteristics of the probability distribution that describes the population in the study. I shall argue that external validity is the wrong way to express the problem and invariance is a poor strategy for fixing it. Probabilistic results are invariant under only the narrowest conditions, almost never met. What's useful is to establish not the invariance of the probabilistic result but the invariance of the *contribution* the cause produces, where the concept of 'contribution' only applies where a 'tendency claim' is valid. Tendencies, I shall argue, are the primary conduit by which 'it-works-somewhere' claims can support that it will work for us.

This raises a serious problem that I want to stress: Reasoning involving capacities/tendencies requires a lot more evidence and evidence of far different kinds than we are generally instructed to consider and we lack good systematic accounts of what this evidence can or should look like.[3]

In particular I shall argue:

1. We need lots more than statistics to establish tendency claims.

2. The very way tendencies operate means that building a good model to predict effectiveness is a delicate, creative enterprise requiring a large variety of information, at different levels of generality, from different fields and of different types.

3. Correlatively we need a large amount of varied evidence to back up the information that informs the model.

2. What can Mill's method of difference establish, even in the ideal?

I should begin with a couple of caveats. My discussion takes 'ideal' seriously. What can be done in the real world is far from the ideal and I will not discuss how to handle that obvious fact. I want to stress problems that we have even where some reasonable adjustment for departures from the ideal is possible. The second caveat is that I discuss only inferences of a narrow kind, from 'T causes O somewhere' to 'T, as T will be implemented by us, will cause O for us'. For most practical policy purposes, inferences that start from 'T cause O somewhere' need to end up with conclusions of a different form from this, often at best at 'T' will cause O' for us' where T' and O' bear some usually not very well understood relation to T and O. I suppose here that the inferences made assume at least that T and O are fixed from premise to conclusion, though other causal factors may be changed as a result of our methods of implementation.[4] With these caveats in place, turn now to the meat of what I want to discuss.

If the conclusion that we look for in answer to the question in the title of this section is to be a causal claim (as opposed to a merely probabilistic claim) about T and O, then here is at least one valid conclusion that can be drawn using Mill's methods,

[3] Consider as a smattering of examples the evidence use guideliness from the U.S. Dept of Education (2003), the Scottish Intercollegiate Guideline Network (2008), Sackett et al. (2000), Atkins et al. (2004) or the Cabinet Office (2000).

[4] Exactly what counts as changing T versus changing additional factors that were in place in the study but are not in place in the target implementation is a little arbitrary. But drawing a rough distinction helps make clear what additional problems still face us even if T and O are entirely fixed. (Thanks to John Worrall for urging me to make these two caveats explicit. For more on both issues, I suggest looking at Worrall's many papers on these subjects. Cf. Worrall (2007) and references therein.)

supposing them applied ideally (which of course we can only hope to do approximately and even then, we seldom are in a strong position to know whether we have succeeded):

> The treatment, T, administered as it is in the study, causes the outcome, O, in some individuals in the study population, X.

This conclusion depends on the assumption that if there are more cases of O in the subpopulation of X where T obtains (the 'treatment group') than in the subpopulation in which it does not (the 'control group'), then at least some individuals in the treatment group have been caused to be O by T.

Since this conclusion depends on taking causal notions seriously and in particular on taking the notion of singular causation[5] as already given, those who are suspicious about causation tend instead to look for mere probabilistic conclusions. The usual one to cite is *mean effect size*: the mean of O in the treatment group minus the mean of O in the control group ($<O>_T - <O>_C$).

What about the external validity of this conclusion?

> *ESEV* (effect size external validity): When will the mean difference be the same between the study population X and a target population θ?
>
> - *ESEV Answer 1*: If T makes the same difference in O for every member of X and θ.

This however is a situation that we can expect to be very rare. Usually the effect of a cause will be relational, depending in particular on characteristics of the systems affected. Consider an uncontroversial case, well-known and well-understood. The effect of gravity or of electromagnetic attraction and repulsion on the force an object is subject to depends, for gravity, on the mass of that object, and for electromagnetism, on the magnetic or electric charge of the affected object.

A more widely applicable answer than ESEV Answer 1 is available wherever the *probabilistic theory of causation* holds. This theory supposes that the probability (in the sense of objective chance) of an effect O is the same for any population that has all the same causes of O and for which the causes of O all take the same value; i.e. the probability is the same for all members of a causally homogeneous subclass.[6] Loosely, 'The probability of an effect is set once the values of all its causes are fixed'. The set of causes of O that are supposed fixed in this assumption are those characteristics that appear in the antecedent of a complete and correct causal law for O.[7] The probabilistic theory of causation then provides a second sufficient condition for effect size external validity.

> - *ESEV Answer 2*: When X and θ are the same with respect to

[5] That is, that 'T causes O in individual i' is already understood. Alternatively, one could presuppose the probabilistic theory of causality in which T causes O in a population φ that is causally homogeneous but for T and it's downstream effects just in case in φ, Prob(O/T) > Prob(O/-T). Then if Prob(O) in the experimental population with T > Prob(O) in the experimental population with –T, we can be assured that there is a subpopulation of X in which 'T causes O'. (But note that if the two probabilities are equal, we have no reason to judge that T causes O in no subpopulations rather than that its positive effects in some cancel its negative effect in others.)

[6] These probabilities will be zero or one where determinism holds but not in cases where causality can be purely probabilistic.

[7] What counts as 'complete' and correct here requires some care in defining; delving into this issue takes us too far from the main topic of this paper.

a) The causal laws affecting O AND
b) Each 'causally homogeneous' subclass has the same probability in θ as in X.

Sufficiency follows from the probabilistic theory of causation. In addition, these two are also almost necessary. When they do not hold then ESEV is an accident of the numbers. This can be seen by constructing cases with different causal laws (hence different subclasses that are causally homogenous) or with different probabilities for the causally homogeneous subclasses (e.g., shifting weights between those subclasses in which T is causally positive for O and those for which it is causally negative or less strongly positive).[8]

These are strong conditions, and they are recognized as such by many scholars who try to be careful about external validity. One good example appears in a debate about the legitimacy of reanalyzing the results from RCTs on the effects on families from disadvantaged neighbourhoods of moving to socioeconomically better neighbourhoods. In 'What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?'[9] Ludwig et al take the purist position: They oppose taking away lessons that the study was not designed to teach. In a section titled '*Internal versus External Validity'*, these authors further caution --

> …MTO defined its eligible sample as…[see below]. Thus MTO data…are strictly informative only about this population subset – people residing in high-rise public housing in the mid-1990's, who were at least somewhat interested in moving and sufficiently organized to take note of the opportunity and complete an application. The MTO results should only be extrapolated to other populations if the other families, their residential environments, and their motivations for moving are similar to those of the MTO population.

The trouble here is that RCTs are urged in the first place because we do not know what the other causes of the outcome are, let alone knowing that they have the same distribution in the study population as in possible target populations. This is a fact the authors themselves make much of in insisting that only conclusions based on the full RCT design can be drawn. For instance, they explain –

> The key problem facing nonexperimental approaches is classic omitted-variable bias.

and

> A second problem … is our lack of knowledge of which neighborhood characteristics matter…Suppose it is the poverty rate in a person's apartment building, and not in the rest of the census tract…[BUT an experimental] mobility intervention changes an entire bundle of neighborhood characteristics, and the total impact of changing this entire bundle…can be estimated even if the researcher does not know which neighborhood variables matter.

The overall lesson I want to urge from this is that effect size will seldom travel from the study population to target populations and even when it does, we seldom have enough background knowledge to be justified in assuming so.

---

[8] The constructions resemble those illustrating Simpson's paradox. Cf. Cartwright (1979); Salmon (1971).
[9] Ludwig et al. (2008)

Effect size is a very precise result however. Perhaps we would be happy with something weaker, for instance, the direction of the effect. So we should ask:

> *Effect direction external validity (EDEV)*: When will an increase (resp. decrease or no difference) in the probability or mean of O given T in a study population X be sufficient for an increase (resp. decrease or no difference) in a target population θ?

There are a variety of answers that can supply sufficient conditions, including –

- *EDEV Answer 1*: If X and θ
  a) Have the same causal laws AND
  b) *Unanimity:* T acts in the same direction with respect to O in all causally homogeneous subpopulations.

- *EDEV Answer 2*: If θ has 'the right' subpopulations in the 'right' proportions.

Both these answers are still very demanding. Clearly they require a great deal of background knowledge before we are warranted in assuming that they hold. In the end I shall argue that there is no substitute for knowing a lot, though there will be different kinds of things we need to know to follow the alternative route I propose – that of exporting facts about the contributions of stable tendencies. The tendency route is often no more epistemically demanding[10] than what these answers require for exporting effect direction or effect size and tendencies are a far more powerful tool more widely applicable: Tendencies can hold and be of use across a wide range of circumstances where ESEV Answer 1 fails; they also underwrite condition EDEV 1b. when it holds yet can be of use even where it fails; and they do not depend, as EDEV 2. and ESEV 2. do, on getting the weights of various subpopulations right in order to be a reliable tool for predicting direction of changes in the outcome.

Let us turn then to this alternative route, which involves exporting not probabilistic facts but causal facts. Doing so requires that we be careful in how we formulate causal claims. In particular it is important for this purpose to distinguish three different kinds of causal claim.

2. Three kinds of causal claim

The distinctions that matter for our discussion are those among --

1. *It-works-somewhere claims:* T causes O somewhere under some conditions (e.g. in study population X administered by method M).
2. *Tendency claims:* T has a (relatively) stable tendency to promote O.
3. *It-will-work-for-us claims:* T would cause O in 'our' population θ administered as it would be administered.

3.1 T causes O somewhere

[10] Nor, sadly, do I think we can hope for answers that are less demanding epistemically if we want sound and valid arguments. And that's the point: we need to know what the premises are for a valid argument; only then can we get on with the serious job of seeing to what degree they can be warranted.

This is just the kind of claim that method-of-difference studies can provide evidence for; and it is important information to have. In saying this I follow, for instance, Curtis Meinert[11] when he says: 'There is no point in worrying whether a treatment works the same or differently in men and women until it has been shown to work in someone.'[12]

It-works-somewhere claims are the kind of claim that medical and social sciences work hard to establish with a reasonably high degree of certainty. But what makes these claims evidence for effectiveness claims: T will cause O for us? I have reviewed the standard answer: external validity. My alternative is tendency claims: T has a (relatively) stable tendency to promote O.

3.2 T has a stable tendency to promote O

3.2.1 What are tendencies?

I have written a lot about the metaphysics, epistemology and methodology of tendencies already.[13] Here I hope to convey a sense of what they are and what they can do with a couple of canonical examples. For instance,

- Masses have a stable tendency to attract other masses.
- Aspirins have a relatively stable tendency to relieve headaches.

The driving concept in the logic of tendencies is that of a *stable contribution*. A feature, like having a mass, has a stable tendency when there is a fixed contribution that it can be relied on to make whenever[14] it is present (or properly triggered), where contributions do not always (indeed in many areas seldom) result in the naturally associated behaviours. The contribution from one cause can be – and often is – offset by contributions from features as well as unsystematic interferences. The mass of the earth is always pulling the pin towards it even if the pin lifts into the air because the magnet contributes a pull upwards. What actually happens on a given occasion will be some kind of resultant of all the contributions combining together plus any unsystematic interferences that may occur.

Reasoning in terms of contributions is common throughout the natural and social sciences and in daily life. Consider the California class-size reduction failure.[15] Here is a stripped down version of the widely accepted account of what went wrong.

There were well conducted RCTs in Tennessee showing that small class sizes improved reading scores there (that is, providing evidence for an it-works-somewhere claim). But when California cut its class sizes almost in half, little improvement in scores resulted. That is not because there was a kind of holisitc effect in Tennessee

[11] Meinert is a prominent expert on clinical trial methodology and outspoken opponent of the US NIH diversity act demanding studies of subgroups because they generally cannot be based on proper RCT design. I agree with him that about the importance of knowing it works somewhere. But my point in this paper is that that knowledge is a tiny part of the body of evidence necessary to make reasonable predictions about what will work for us.
[12] Quoted from Epstein (2007).
[13] Cf. Cartwright (1989) and (2007a).
[14] Though note that some tenancies can be purely probabilistic and also the range of application can be limited.
[15] Bohrnstedt et al. (2002)

where the result depended on the special interaction among all the local factors there. Rather, so the story goes, the positive contribution of small class size was *offset* by the negative contributions of reduced teacher quality and inadequate classroom and backup support. These latter resulted because the programme was rolled out statewide over the course of a year. This created a demand for twice as many teachers and twice as many classrooms that couldn't be met without a dramatic reduction in quality. The positive contribution of small class size was not impugned by these results but possibly even borne out: The presumption seems to be that scores would have been even worse had the poorer quality teaching and accommodation been introduced without reducing class sizes as well. The reasoning is just like that with a magnet and gravity acting together on a pin.

Tendency claims are thus a natural conduit by which it-works-somewhere claims come to count as evidence for it-will-work-for-us claims. It should be noted however that a stable tendency to contribute a given result is not in any way universally indicated by the fact that a feature like class size participates in causing that result somewhere. Nevertheless, if a result is to be exported from a study to help predict what happens in a new situation, it can seldom be done by any other route.

3.2.2 The big problem for tendency logic

The central problem for reasoning involving tendencies is that we do not have good systematic accounts of what it takes to establish such claims. We have nice histories of establishing particular claims, especially in physics, but little explicit methodology. This contrasts, for instance, with it-work-somewhere claims. We have a variety of well-known well-studied methods for establishing these, methods for which we have strong principled accounts of how they are supposed to work to provide warrant for their conclusions and of where we must be cautious about their application. Recently, for instance, there has been a great deal of attention and debate devoted to Mill's-method-of-difference studies and to the advantages and disadvantages of various methods for ensuring that the requisite conditions are met that allow them to deliver valid conclusions. But if I am right that tendencies are the chief conduit by which it-works-somewhere claims come to support it-will-work-for-us, this attention focuses on only a very small part of the problem. For an it-works-somewhere claim is at best a single rock in the kind of foundation needed to support a tendency claim.

So I want to plead for more systematic work to lay out the kinds of studies and types of evidence that best support tendency claims. As best I can tell ultimately we need a theory to establish tendency claims, though admittedly often we will have to settle for our best stab at the important relevant features of such a theory. That's because contributions come in bundles and are characterized relative to each other. We only have good evidence that gravity is still working when the pin soars into the air because we can 'subtract away' the contribution of the magnet and thus calculate that gravity is still exerting its pull. To do that we need to have an idea both about what other factors make what other contributions and what the appropriate rule of composition for them is.[16]

Of course we most often have to proceed to make it-will-work-for-us predictions without a well-developed theory. In that case we make our bets. My point is that we must be clear what we are betting on and what evidence is available to back up the

[16] Note though the tension here: Most advocates of RCTs like them because, they claim, no substantive theory is required to do what they purport to do – i.e. establish an 'it-works-somewhere' claim.

bet, even what kind of further evidence we should be setting out to learn. Are we betting on, and using the logic of, stable tendencies, and if so, to what extent does our evidence back us up in this? Or are we betting on facts about identical causal laws and correct distributions of other causal factors between study and target populations, and if so, to what extent does our evidence support that?

3.2.3 Tendencies versus external validity

My overall message is that sometimes there are tendencies to be learned about. Where there is a stable tendency, this provides a strong predictive tool for a very great range of different kinds of target populations. It naturally does not tell us what the observed result will be unless we know there are no unsystematic interferences at work, we have good knowledge of the contributions that will be made by the other causal factors present and we can estimate how these contributions combine, which is very seldom the case outside the controlled environment of a physics laboratory. But when we know a tendency claim we can make a prediction about the direction of change. Whatever the result would have been, if the cause is added the new result will differ by just the amount predictable from the contribution. But beware. The comparison we can make is with what the result would have been *post implementation*, just subtracting the effect of T itself. So, even restricting ourselves just to claims about direction of change, we still have not arrived at an 'it will work for us' claim, as I have characterized that.

Let us return to a comparison of tendencies versus external validity – predicting that 'the same' effect, either effect size or effect direction, will hold in the target as in the study population.

- Neither can be taken for granted.
- Both require a great deal of evidence to warrant them, though of different kinds.
- With respect to **effect direction**:
  - **Stable tendencies**: Post-implementation effect direction can be predicted from knowledge that T has a stable tendency to promote O (that it makes, say, a known contribution) without requiring knowledge of the distribution of other causal factors in the target.
  - **External validity**:
    - Recall by contrast that under **EDEV 2**. the distribution of causally homogeneous subpopulations must be 'right' in order for the effect direction to be the same in the target as in the study population; and of course for cases in which some set of right conditions hold, it takes considerable background knowledge of what the other causal factors are and what the target situation is like to be warranted in assuming they do.
    - T has a stable tendency to promote O implies **EDEV 1.b**).
    - What about **EDEV 1.a**)? I have not gone into the issue of the range across which a cause must make the same contribution in order to be labelled as a tendency. Obviously there is no firm answer. What matters is that there should be good reasons to back up whatever range is presupposed in a given application. Many well-known tendencies, however, can survive a change in the other causes that affect the same outcome. Philosophers keen on modularity as a mark of genuine causation often insist that this is a widespread feature and it is often supposed in science as well. For instance most of us are familiar from elementary economics with exercises to calculate what happens if the demand laws change while the contribution to exchange from the supply

side stays fixed, and vice versa. When that's the case tendency reasoning can provide predictions of effect direction that EDEV 1 cannot, though of course the assumptions that a tendency is stable across changes in other causal laws needs good arguments to back it up.

- With respect to **effect size**:
  - **Stable tendencies**. Effect size can be calculated when the contributions of all major tendencies present in a situation are known, or reasonably approximated, along with the appropriate rule of combination. This is typically what we demand from an engineering design but can surely never be supposed for social and economic policies for effects on crime, education or public health. Various narrow medical cases are generally thought of as lying in between these extremes.
  - **External validity**. It is seldom the case that the target and study populations have the same causal laws and same distribution of causal factors, and even more rare that we should be warranted in supposing so. So if the external validity of effect size is our primary method for learning something about target populations from Mill's method-of-difference studies, these studies will be of very little use to us.
- Use of the logic of tendencies is epistemically demanding. But so is external validity, only in different ways. Tendency knowledge, where available, can do more than traditional external validity reasoning and is far more widely applicable. Moreover tendency logic is well established to work well in a variety of domains. So it is wasteful and capricious to refuse to use this logic when evidence is available for it. Of course often some evidence will be available but not enough to clinch our conclusions. That is the human condition and it applies in spades to external validity reasoning as well. When clinching evidence is missing, we had best proceed with caution and, if we can, hedge our bets.

3.3 T will work for us

3.3.1 Counterfactuals: case-specific versus general-purpose causal models

Julian Reiss and I[17] each argue that it-will-work-for-us claims are best supported by case-specific causal models. It is not unusual among causal theorists nowadays to urge that these kinds of claims are best evaluated via causal models. After all, these are singular counterfactual claims: T would cause O if it were implemented in our population as it would be implemented there. The central difference between our claims and many others is the emphasis on 'case-specific' – i.e. on models built specifically for the counterfactual at hand, as it will be implemented.

For contrast consider the models of Judea Pearl, who has developed what must be the most detailed and thorough semantics for causal counterfactuals now available.[18] In Pearl's semantics counterfactuals are, as I advocate, evaluated on the basis of a causal model. I think I can explain the kinds of difficulties that face the use of general-purpose as opposed to case-specific models by reference to Pearl's models, without laying out details of his approach.

Causal models for Pearl are of a very specific form. The form connects neatly with our general probabilistic methods for discovering 'it-works-somewhere' claims; and this is both their strength and their weakness. For the somewhere is never here.

[17] Reiss (2007); Cartwright (2007b)
[18] Pearl (2000)

Even if – contrary to what can ever realistically happen – a study encompasses the entire target population, the population of the study is not literally the same as the one about which future predictions are made. One may suppose that the same causal model will describe the 'same' population in the future as in the past but that is a strong assumption of external validity and it should have evidence, reason and argument to back it up.[19]

Reiss and I both stress that a causal model for evaluating 'it-will-work-for-us' claims needs to be built to the case at hand – for the given cause as, where and when it will implemented. A causal model for the system as it has been functioning or for similar systems is neither necessary nor sufficient. It is not sufficient because implementations of a cause often bring about importantly relevant changes, not only in the arrangement of other causes but also in the basic governing causal principles. It is not necessary because, as with external validity, *the same* as before or as elsewhere is the wrong idea. How the system has behaved so far or how 'similar' systems behave can be a clue to what will happen when the cause is implemented, but only a clue. We often have reason to suppose that it is the central clue; often we have reason to think it is not because we know how easily the system of laws or the arrangement of causes at work in our case might be. 'The same' causal model is just as much a hypothesis about a future case as is any new causal model proffered in its stead.

In the ideal a case-specific causal model to evaluate a specific it-will-work-for-us counterfactual will contain two essential ingredients:

- a list of 'all' the causes (or all the ones that can have a significant effect on the outcome) that will be present once the targeted cause is implemented
- a tool for calculating what happens with respect to the targeted effect when these all act together.

With this information we can predict the effect.

The trouble with causal models of this form is that we are seldom in a position to produce them with anything like a high degree of reliability. It is thus a good thing that for many kinds of predictions they are not necessary. Sometimes there are 'shortcut' models, or what following Gerd Gigerenzer[20] we might call 'cheap heuristics', that predict approximately enough the same result, sometimes even provably so, without mirroring the causal narrative that will unfold in nature as an ideal case-specific causal model does. Alternatively, sometimes there are good partial models that predict aspects of the effect, for example, estimates of effect size difference. Moreover when we are lucky an already constructed model laying out the causal laws that have governed the system till now or that govern similar systems can be taken over wholesale to serve for the specific case. But to repeat, the case-specific model that we get by this strategy is as much in need of justification as any other.

[19] It should be noted that this is not just a reappearance of Hume's problem of induction. For the problem itself presupposes that there are general principles of some kinds at work in nature and even that we can find out about them, understand how they work and predict what kinds of conditions are required for a system to continue to operate as before. This is how we can often be confident that our interventions will not be successful because they will shift the arrangements of causes at work or undermine the operating principles. A better label for the problems for invariance I raise here is 'Mill's problem of induction' since it is the kind of worry that he described in arguing that economics cannot be an inductive science. (Mill (1836); for further discussion see Cartwright (forthcoming.)

[20] Gigenerzer et al. (1999)

3.3.2 Tendencies and causal models for it-will-work-for-us claims

Tendency claims play an important role in constructing causal models for evaluating singular causal counterfactuals. Where causes act with stable tendencies we can be in a powerful situation with respect to either full or partial models because in this case the causes contribute by a systematic rule that we can learn about and encode in our theories. Otherwise prediction is more piecemeal and local and though we often do it well, there is little good philosophical work on how to do so. So where tendencies can be relied on,[21] these will be a huge help in constructing a causal model for the evaluation of a specific causal counterfactual. Even if not all the causes present have a stable tendency so there are unsystematic interferences, if the targeted one has a known contribution then it may at least be possible to calculate an effect direction or even an effect difference. And certainly tendency claims are the central way by which it-works-somewhere claims can come to count as evidence that it will work for us.

Where we know of no stable tendencies then we are more at sea. I take it that we are often good at local detailed causal reasoning but that we need a great deal more concerted research on what strategies are reasonable to pursue in these areas. What matters, I believe, is to recognize the epistemic and ontological situation we are in when we want to judge if a treatment will work for us and do the best we can, hedging our bets and recognizing when we are making heroic assumptions in constructing our causal model and when not.

Given the limitations of tendency logic it is important to recognize that Pearl's semantics, and others like it, presuppose tendencies.[22] Pearl's causal models consist of a set of causal claims in functional form, one for each effect under study, with a dependent variable as effect and the independent variables as causes, plus a probability measure over the exogenous variables (i.e. those variables representing quantities not caused by other quantities represented in the set under study). To evaluate the counterfactual 'T would produce O for us'[23] Pearl substitutes for the law in the model for t ($t = f(x,y,z,\ldots)$), $t = T$, leaving all other laws in the model the same. This represents setting the value of t 'surgically', as should be done in a method-of-difference study. The assumption that this is always possible for any cause in the model is called *modularity.* The value of O that results is ultimately calculated from the law in the model for O, a law of the form $O = g(r,s,t,\ldots)$.

What we should note is that the general assumption that a system of laws is modular presupposes that the causes in that model have stable tendencies, stable at least across all the uses to which the model and its accompanying semantics is put. The contribution of a cause to an effect is given by the term in which that cause appears in the law for the effect; the rule of combination, by the functional form. Consider for instance the law, $acc = GM/r^2 + \varepsilon q_1 q_2/r^2$, for the acceleration of a particle of charge q_2 in the vicinity of the earth (of mass M) and of another particle of charge of q_1. The mass of the earth makes a stable contribution of the size of its mass (M) multiplied by the acceleration of gravity G and the inverse of the square of the distance of its

[21] But be careful. Many tendencies are conditional: They hold relative to an underlying structure that gives rise to them. So in using them we are betting on the stability of the underlying structure – in my language, a 'nomologival machine'; and, as always, it is best to have as much evidence as possible to decide which way and how much to bet. (For a longer discussion see Cartwright (forthcoming) and (1989).)

[22] Although James Woodward (2004) does not offer a detailed semantics for counterfactuals, he is another causal theorist who makes very strong modularity assumptions, hence very strong tendency assumptions.

[23] Here I suppose that T and O are specific values that some random variable, t, o, can take.

centre of mass from the particle. This adds vectorially with the contribution that the charge q_1 makes, which is its size multiplied by $\varepsilon q_2/r^2$. Ask now 'Would setting $q_1 = Q_1$ increase the particle's acceleration?' To answer, following Pearl, calculate acc = $GM/r^2 + \varepsilon Q_1 q_2/r^2$, substituting for the other values in this equation the values they take in the situation at hand. The assumption that the mass of the earth continues to contribute in exactly the same way as the value of the charge is changed is to suppose that the mass has a stable tendency. Similarly, to assume that the functional form for the electrostatic term stays the same, and indeed the overall functional form for acceleration does too, is to assume that charge has a stable tendency.[24] So to assume modularity for changes under every variable in the model is to make very strong tendency assumptions.[25]

I obviously have no quarrel with tendencies, having defended them for well over two decades. But we need to keep clearly in mind the lesson of section 3.2.2. Causes often act holistically; tendencies cannot be taken to be the rule. Mill himself felt that the logic of tendencies applied in physics and in political economy but not in chemistry or more generally in the study of society[26] and Julian Reiss argues that they are not all that common even in political economy.[27] Nor are the conventional methods by which we test causal claims sufficient to establish tendency claims, especially not the wide-ranging claims about tendencies presupposed in taking a causal model to be modular. And I should stress that this is true not only for the method-of-difference methods discussed in this paper but for a wide variety of other valid methods for causal inference as well, including various econometric methods and many that trace causal pathways.

Aside on representation. This brings me to a point about representation that is somewhat more complex than the issues I have discussed so far, but one that matters to the question of how causal models help in the evaluation of it-will-work-for-us claims. Pearl, faced with challenges like mine to strong modularity assumptions, maintains that when the model is not modular that just means it is misspecified; that is, we haven't written down the right model. Whether he is right or not depends on how one conceives of his causal models. One way is to start with an independent notion of 'causal law', one that meshes at least reasonably well with our accepted methods for testing/establishing causal laws. Then one can consider how this model can be used (if at all!) to evaluate singular causal counterfactuals. If we read Pearl this way then it looks as if he offers a semantics that should allow us to evaluate any

[24] This can be a misleading example because these tendencies are, or are often supposed to be basic, hence universal. As mentioned in footnote 21, most tendencies, however, depend instead on some stable underlying structure to give rise to and maintain them. So they are stable across changes that affect only arrangements in the superstructure, not necessarily across those that affect the substructure.
[25] Again, there is a serious caution to be urged. I said that Pearl's equations were of a familiar kind that we have rules for how to estimate and sufficient conditions (as with instrumental variable models or others I describe in Cartwright (2007)) for determining if they can be interpreted causally. But neither these standard methods nor the sufficient conditions I know about warrant the modularity assumptions necessary to use the equations as instructed to draw counterfactual conclusions. This remark is essentially a repeat of my two-fold point that the equations, given their prescribed use in warranting counterfactual predictions, presuppose tendencies and that tendencies need a good deal more evidence to be warranted than that provided by the standard methods that warrant it-works-somewhere conclusions.
[26] Mill (1836). This would have placed Mill in the later methodenstreit (the battle of methods) more on the side of Schmoller and the holists, as opposed to Menger and those who believed in the wide applicability throughout the social sciences of the analytic method.
[27] Reiss (2007)

counterfactual with any variable[28] from the model in the antecedent and any variable from the model in the consequent.

This I believe is what Pearl is generally taken to be doing; and as a strategy it has exactly the problems I have described here. First, we do not have sufficient reason to take tendencies to be the rule, or even the fallback position. The causal laws governing situations are often holistic so that they are not much of a guide about what happens when the whole causal complex is no longer the same. Second is the point I have mentioned but not developed in any detail here,[29] that causal laws generally depend on some underlying structure that gives rise to and maintains them and many of the ways we implement antecedents in counterfactuals can undermine this structure in ways that destroy the very causal laws we hope to use to evaluate the counterfactual.

A usual fix for these problems is to try to extend the variables in the model. In this case the new variables would have to include descriptions of the possible underlying structures that could arise from any method of implementing a change on any variable in the model plus all new variables implicated in causal relations that the various new substructures would give rise to. Of course this is no fix for the problem of holism. As a fix for the problem that causal laws as we usually think of them and test for them depend on vulnerable substructures to support them, it seems impossible. Moreover, it is a cheat. We cannot define a proper variable whose values are the unending open-ended array of possible substructures that could exist once we start to intervene;[30] and if we could, it certainly would not be a random variable of the kind required in Pearl's models: Neither nature nor we supply a probability measure over any such array of possibilities.

The second way to interpret the causal laws in a model is to backread the 'causal laws' for a situation from the proffered semantics and the set of counterfactuals true for a given set of features in that situation. That is, the causal laws are whatever they have to be to allow the semantics to give correct results for the counterfactuals. This interpretation fits more closely with the claim that if the models aren't modular then they are misspecified. Probably it is easy to show that a model of Pearl form can be created that gives the correct results for any targeted counterfactual. But there is no guarantee that such a model can be created for an arbitrary collection of true counterfactuals over features under consideration, let alone a full set of them.

My own version of a causal model falls between these two. It is a model purpose built for evaluating a particular counterfactual as it would be implemented. Write down the causes of the targeted effect that will be in place given the implementation and consider what together they produce. The strength of this proposal is that it is sure to produce correct answers if we can carry it off. This is just the flip side of its chief weakness: We do not have set procedures for doing this and often are at sea.

---

[28] Actually the semantics is stronger than that for it allows a mix of variables in the antecedent.

[29] For more on this point see my various discussions of nomological machines (to be found in the two references from footnote 21 plus further references in those).

[30] John Worrall, in referee's comments, suggests that many people think that the array of structures that could exist is not open-ended. I suppose they take a view of the world reflected in Wittgenstein's *Tractatus:* crudely, there are a fixed number of features in the world and the possible facts are exactly all the combinations of all the possible values of all the possible features. If I had to indulge in metaphysics, this is not one I would go for. But even if it were true, this does little in aid of establishing that there are random variables to represent this vast array since that requires reason to believe that there is a proper probability measures over it. And where does that come from?

One may wish for more. Indeed a referee for this paper expresses just this: 'We are told that we need to model the causal situation with tendencies but there is little detail on what such models would look like.' I am happy that sometimes such models will look like Pearl models, and that we could then use Pearl's semantics to generate counterfactuals with them. What I do not accept is that we can give much advice about how to build the model. I have spent a lot of time studying very successful models in physics – like models for lasers or for the gyroscopes that reveal precession due to space-time coupling in the Stanford Gravity-probe experiment, and also studying promising models in economics that are less predictively successful but are not disasters. The most I can say is that the modelling enterprise and importantly the enterprise of figuring out how good these models are *ex ante* – before they are used for prediction – seems to have no fixed rules and little good substance-neutral advice. But that I think is not only a fundamental fact about evidence; it is the human condition, better to be acknowledged and managed than denied or ignored.

3. Conclusion

My focus here has been on Mill's method-of-difference studies and what they can teach us about whether proposed interventions will have targeted effects when implemented as they would in fact be implemented (i.e. 'it-will-work-for-us claims). These methods, I have argued, can establish claims of the form 'It works somewhere.' But it's a long road from 'It works somewhere' to 'It will work for us'.

The central problem I raise is that we do not have very good methodological guides for how to traverse this road. I argue that 'external validity' is generally a dead end: it seldom obtains and, because it depends so delicately on things being the same in just the right ways, it is even rarer that we can have reasonable warrant that it obtains. Instead tendency claims are the chief conduits by which 'it-works-somewhere' claims come to be evidence that a proposed intervention will work for us. This narrows the problem but does not solve it. For we do not have good explicit methodologies for how to establish tendency claims. Nor do we have explicit methodologies for how to use them to build case-specific models for evaluating whether the proposed intervention will work. And if I am right about how predicatively successful models are usually built even in physics, we haven't much reason to think any such methodology will be forthcoming.

What then is the role of the highly vaunted Mill's method-of-difference studies, including the current favourite, the RCT, in providing evidence that T will work for us to promote O? The ideal RCT can show that T works somewhere; a real RCT is one fallible indicator in what is hopefully a far fuller evidence base that T works somewhere. That T works somewhere can be a part, albeit a small part, of an evidence base to support T's capacity to contribute to O. That T has a capacity to promote O can serve as part, again probably only a small part, of the evidence that supports the case-specific causal model that is the eventual base for our predictions about whether T will work for us. So it is indeed a long road and most often an insecure one. But it is better to understand and acknowledge that than to presuppose heroic assumptions without admission, without examination, without evidence and without all the hedging that responsible betting calls for.

Bibliography

Atkins D, Best D, Briss PA *et al* [GRADE Working Group] (2004) Grading quality of evidence and strength of recommendations, BMJ 328 (7454):1490 (19 June), doi:10.1136/bmj.328.7454.1490

Bohrnstedt, G.W., Stecher, B.M. (eds.) (2002), "What We Have Learned About Class Size Reduction in California", California Department of Education

Cabinet Office Performance and Innovation Unit (2000), *Adding It Up: Improving Analysis & Modelling in Central Government*, London: HMSO

Cartwright, N. (forthcoming), 'How to do Things with Causes', Presidential Address, *Proceedings and Addresses of the APA, 2009*

Cartwright, N. (2007a) 'Causal Laws, Policy Predictions and the Need for Genuine Powers' in N. Cartwright's *Causal Powers*, *What Are They? Why Do We Need Them? What Can and Cannot be Done with Them?* 2007, Contingency and Dissent in Science Series, London: Centre for Philosophy of Natural and Social Science, LSE. Also in Handfield, T (ed.), 2009, *Dispositions and Causes,* Oxford: Clarendon Press.

Cartwright, N. (2007b), *Hunting Causes and Using Them: Studies in Philosophy and Economics*, New York: Cambridge University Press

Cartwright, N. (1989), *Nature's Capacities and their Measurement*, New York: Oxford University Press

Cartwright, N. (1983), *How the Laws of Physics Lie*, New York: Oxford University Press

Epstein, Steve (2007), *Inclusion: The Politics of Difference in Medical Research*, Chicago: University of Chicago Press

Gigerenzer,G., Todd P.M., ABC Research Group (1999), *Simple Heuristics that Make us Smart*, New York, NY: Oxford University Press

Ludwig, J., J. Kling, G. Duncan, L. Katz, R.Kessler, L.Sanbonmatsu (2008), 'What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?', *American Journal of Sociology,* 114, 144-88

Mill, J. S. (1836 [1967]), "On the Definition of Political Economy and on the Method of Philosophical Investigation in that Science", reprinted in *Collected Works of John Stuart Mill,* Vol. 4, Toronto: University of Toronto Press

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

Reiss, J. (2007), *Error in Economics: The Methodology of Evidence-Based Economics*, London: Routledge

Sackett DL, Straus SE, Richardson WS, Rosenberg & Haynes RB (2000) *Evidence-Based Medicine: How to Practice and Teach EBM* (Second Edition), Edinburgh: Churchill Livingstone

Salmon, W. (1971), *Statistical Explanation and Statistical Relevance.* Pittsburgh: University of Pittsburgh Press

SIGN (Scottish Intercollegiate Guidelines Network) (2008), *SIGN 50: A Guideline Developer's Handbook (Revised edition, January 2008)*, Edinburgh; SIGN Executive

U.S. Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance (2003), *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide. http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf*

Woodward J. (2004), *Making Things Happen,* Oxford: Oxford University Press

Worrall, J. (2007), 'Why There's No Cause to Randomize', *BJPS,* 58, 451-88