Where's the Rigor When You Need It?

Nancy Cartwright

UCSD and Durham

I was charged with the question: When it comes to causality, 'What do social scientists know?'---and, presumably, 'How do they know it?' In expanding on that, Iván (Marinovic) in his invitation added: "Often the issue of causality and identification is ignored or 'resolved' by adding explanatory variables (given the large amount of data available). If you have some specific thoughts about this type of research I would really appreciate that you discuss them." There are issues here that need to be talked about. Many are technical, and you have here real experts to discuss them. But there are other issues that are not technical, that really matter, that do not get discussed, and we repeatedly get into trouble because we do not pay attention to them. To unearth some of these issues, there are two points I want to urge today:

1. Do not do inference by pun.
2. In general, well substantiated, reliable general claims in science ≠ generalizations, i.e. claims warranted by generalizing from individual instances.

The first must be totally uncontroversial. The second certainly should be.

1. Inference by pun

Behind the first is the obvious observation that there are a great many different kinds of causal claims one can hope to know---or better, since knowledge suggests a level of certainty that is generally beyond us---many different kinds of causal claims that we can hope to find compelling support for. Here are a few that will probably enter discussion today:

- X causes Y
- X(u) causes/d/would cause Y(u)
- The average treatment effect of X(u) on Y(u) is Φ
- The local average treatment effect of X(u) on Y(u) is Θ
- …

When we have claims like these there are three issues that need to be settled. The first concerns *meaning*. What are we claiming in making these claims? What is it for one quantity to cause another in an individual unit? What do we mean by the *effect* of a 'treatment' on an individual? And what is the scope of the claim: All units everywhere? All units of a certain kind (e.g. in a specific population)? Some units in a specific population? Some units somewhere in the world? The second issue concerns *method*. What are (reasonably) reliable methods for supporting these claims? The third issue is *use*. What further inferences could be drawn from these claims if they were established?

Not only must we be clear about these three issues, but *the three must mesh together properly*. The methods must be appropriate for establishing *the very thing* the claim asserts. And the inferences that draw from it must be licensed by *that* claim, the very claim that the methods support. Otherwise we are doing inference by pun. This is all too common. We employ some methods that, at least in the ideal, are very reliable for establishing one specific, usually very narrow, kind of causal claim. Then because the claim has the word 'cause' in it, we draw inferences that are licensed by some other similar sounding claim that also has the word cause in it, but one not supported by the methods we employed in the first place.

To see this more clearly, let's start with *methods.* There are a variety of methods that can provably provide solid support for causal claims, including:

- Controlled experiments
- Natural experiments
- Instrumental variables models
- Econometric models satisfying special conditions
- Qualitative comparative analyses that satisfy special conditions
- Causal Bayes nets methods
- Process tracing
- Derivation from a good theory.

Every method comes with a set of assumptions that must be met for it to supply reliable results. It is important for the issues I raise to recognize that though the very same assumption can play both roles, there are two distinct roles that assumptions play. The first is to *characterize* the notions involved; that is, to make clear what exactly it is that is being claimed. The second is to specify what features must be true about particular cases in order for the methods to provide reliable results about those claims.

For instance, discussions of randomized controlled trials (RCTs) or instrumental variable approaches often begin with what is sometime called a 'structural' or a 'potential outcomes' equation for a population of units u in U:

$$\text{POE: } y(u) c= \alpha(u) + \beta(u)x(u) + w(u)$$

where we suppose there is a probability measure *Prob* over (α, β, w) for U. We often then aim to discover β: how much oomph on average x supplies towards y; or, failing that, Exp[β]. If Prob(β = 0) = 1, then x never contributes to y for U; when β(u) differs from 0 for some u, we say that x is a cause of y for that u.

You will have noticed that I write this equation not with just an '=' sign but with the symbol 'c='. That's because we are not trying to estimate Exp[β] in just any equation relating x to y that is true for U; we want the right equation---the one with the causes of y on the right-hand side. For instance, suppose x and y are both joint effects of the common cause v:

$$1.\ x(u)\ c= \alpha_1(u) + \beta_1(u)v(u)$$

$$2.\ y(u)\ c= \alpha_2(u) + \beta_2(u)v(u)$$

Then

$$3.\ y(u) = \alpha_2(u) - (\alpha_1(u)/\ \beta_1(u)) + (\beta_2(u)/\ \beta_1(u))\ x(u) = \alpha_2(u) + \gamma(u)x(u)$$

It is not the coefficient of x in this equation that we are concerned about: finding that it is non-zero will not tell us whether x causes y for any units in U. To learn about the effects of x on y, it is not enough to show, say, that we have an unbiased estimate of the coefficient of x for y in an equation for y that holds in U. We must show that we have an unbiased estimate of the coefficient in the right equation---the causal equation. But what is the difference? What is meant by 'causes' when we say we want the equations that have causes on the right and effects on the left?

This is a question that is often not clearly answered. Economists talk of *structural* equations or *potential outcome* functions. I talk of a system of *causal equations* governing behaviors in U. This brings us squarely to the issue of *meaning.* What sense are we to give to the notion of *causal* or *structural* equations? I notice that Guido Imbens in a recent paper gives this brief explanation: "Functions are structural or invariant in the sense that they are not affected by changes in the treatment" (Imbens 2014, 12). This clearly needs elaboration. The usual way to do so would be to provide a set of axioms that a system of equations must satisfy if they are to be causal. *Minimally* I would expect these to include the assumptions that the causal relation is:

- asymmetric,
- irreflexive,
- time-ordered from earlier to later and
- causality is preserved if the causes of a right-hand-side variable are substituted for the variable,

plus, importantly, that any non-causal equations, like 3., are mere algebraic consequences of the causal equations.

With these axioms and an appropriate definition of *intervention* I can show that equations that are invariant under interventions on right-hand side variables must be causal. This makes a neat fit with Imbens' brief remarks and also with the 'invariance-under-intervention' criterion for causality championed by James Woodward and currently fashionable in philosophy.

I am not urging here that we should adopt this axiomatization, nor any other particular one. *Cause* is a loose everyday word that does many different jobs in many different contexts. But that won't do for social science, where rigor demands precise concepts. No single precise concept will match up with the ordinary one; there is no one axiomatization that is *the* right way to characterize causality. What matters is that we provide a precise characterization, that the scientific characterization we provide is up to the job we set it to and that we stick with the same characterization throughout.

Returning now to assumptions. Recall, there are two kinds: Some assumptions are required to make explicit the constraints we are adopting on what a system of causal equations is. What do we mean by 'causal', or 'structural'? Additional assumptions are required---facts about the population under study and about further features the causal equations must have---if our methods are to be trusted to teach what they are supposed to. For instance, we often use the net

outcome difference---Exp[y|treatment] – Exp[y|control]---to estimate the 'effect size' Exp[β] in a causal equation like POE. To do so we typically assume that x in the experimental population is probabilistically independent of each of α, β, and W in the POE. In that case Exp[β] = Exp[y|treatment] – Exp[(y|control)]. These independence assumptions about relations among variables in the study population are assumptions of the second kind; when these are satisfied, our RCT can be trusted to deliver what it is supposed to: an unbiased estimate of the coefficient of x for y in the POE for that setting and population. I stress the two different kinds of assumption because discussion is likely to focus on the second, the first is often ignored or the two get carelessly lumped together. But without being clear about the first, we do not know what it is that our methods have found out and we don't know what further facts we can infer from what we have found out.

That brings us to *inference.* There are three kinds of inferences we tend to draw from study results. Or---more carefully---three kinds of conclusions for which *study results can figure as part of a (sometimes very large) body of support*. These can be either qualitative or quantitative, causal or probability claims about:

1. The population and setting in the study.
2. A specific populations and setting outside the study.
3. A general conclusion about what holds widely or across a given domain of populations and settings.

One example of the first is the kind of conclusion we can draw from positive results in a good RCT: that Exp[β] = Exp[y|treatment] – Exp[y|control] for that study population and setting. Though there are issues with regard to conclusions of this kind, I shall not go into them since they are less controversial than for the other two.

So, 2. Consider an RCT, say for class size and reading scores for 8 year olds. When would you expect to get the same effect size in an RCT in California as in one conducted in Tennessee? We can tell exactly when by inspection. Recall: the effect size = Exp[y|treatment] – Exp[y|control] = Exp[β] in the POE governing the study population and setting. So you get the same effect size in both just in case a) x plays the same role in the POE for California as it does in the POE for Tennessee: it is either present in the POE for both or absent from both; and b) Exp[β] is the same in both.

So what does β represent? It represents the net effect of the whole team of support factors necessary for x to produce a contribution to y, the factors without which x cannot contribute to y. Epidemiologists represent these factors in causal pies, like the simplified one for class size reduction in Figure 1.

[Insert Figure 1 about here]

So you would get the same effect size in an RCT on a California study population and setting as in Tennessee just in case class size really could be a cause given the causal principles that hold in the California population---class size really figures in the POE for the California population, and the (net effect of) the support factors necessary for class-size reduction to produce its good effects in California have the same average as do the support factors necessary in Tennessee. That's a tall order (and note particularly that RCT evidence is not evidence for either of these assumptions).

This is just one example of how we might use facts about causes that can get nailed down in a study population as part of a case to support causal conclusions about a target population.

What more can be done? This is just what Elias Bareinboim and Judea Pearl try to answer in several recent papers on the transportability of causal effects (see e.g. Pearl and Bareinboim 2011). They consider two different populations. Suppose we have available both some causal results as well as some purely probabilistic information for population 1, while for population 2 we have only probabilistic information, but no causal information. We know certain probabilistic facts and certain causal facts that the populations share and some that they do not. Bareinboim and Pearl produce some theorems that describe what further causal conclusions about population 2 are fixed by this body of information. One of their overall lessons is to underline the obvious fact that exactly what conclusions about population 2 can be supported by information about population 1 depends on exactly what causal and probabilistic facts they have in common.

There are two things to note here. First, Bareinboim and Pearl do not say what they mean by 'causal'. But their work in these papers builds out from previous work by Pearl on causal Bayes nets, which starts with sets of equations which, as we can tell by inspecting their form and Pearl's use of them, satisfy the minimal set of axioms I described earlier. The second is another obvious fact that they do not underline: an argument, like a chain is only as strong as its weakest link. We can draw conclusions about population 2 from:

- Probabilistic facts and causal results on population 1,

plus,

- facts about probabilistic and causal commonalities between population 1 and population 2.

Our conclusion is no more certain than any one of its premises. What I find odd is the asymmetry with which we tend to treat the two kinds of premises. We currently put a vast effort into securing premises in the first category, both empirical work---more and more RCTs, better and

better instrumental variables models---and theoretical work---for instance, under what conditions can we estimate an average treatment effect and when can we only get what is called a 'local' average treatment effect. We invest far less in empirical work to establish premises of the second kind. There are broadly speaking two kinds of warrants that could support these assumptions: methodological and causal-empirical. A methodological warrant consists in there being good reasons---based on the way the study was sampled and conducted---to support the assumption that the study is representative of the target. Causal-empirical warrant consists in knowing enough about the support factors and their distribution in the study and the target to support the conclusion that the net effect would be the same.

The theoretical work is even more wanting. We have little clear, widely accepted methodology for how to go about establishing premises of the second kind. And I see little investment in trying to fill this gaping hole in our methodology.

2. You can't get general claims by generalizing

Let's turn now to inferences to the third kind of conclusion: general claims. We see these everywhere nowadays, in the expression 'It works', in the rise of hundreds of 'What Works' centers and in the admonition in evidence-based policy to rely only on interventions that have been shown to work. In practice it seems that what is supposed is that we can infer that an intervention works---works in general---from successful results on RCTs in some number of different populations. Here are two instances from evidence-based medicine, borrowed from Jonathan Fuller (2015):

> We suggest that guideline panels deal with the issue of generalizability by accepting that results of randomized trials apply to wide populations unless there is a compelling reason

> to believe the results would differ substantially as a function of particular characteristics of those patients. (Post et al., 2013, 5).

> [I]n trying to judge whether interventions studied in research 'will work for us', Cartwright, like many others, conceptualises the challenge as being to demonstrate that the characteristics and circumstances of the research are sufficiently similar to those to which extrapolation is being contemplated. But why should the challenge be conceptualised that way round? Why not instead ask 'Are there any good reasons to believe that the research is not relevant to us, that "It won't work for us"?' If there are not, and considering the undesirable alternative ways of reaching a decision, the default position should be that the result should be regarded as applicable. (Petticrew and Chalmers 2011, 1696).

Supposing that we can infer that an intervention works---works in general---from successful results on RCTs in some number of different populations, or that we should take this as the default conclusion if we have no reason to the contrary, is nonsense. That mode of inference is what we call 'induction by simple enumeration' and we know we cannot trust that: Swan 1 is white, swan 2 is white…, so all swans are white. Study population 1 does x, study population 2 does x...., so all populations do x. And with the studies we face the additional drawback that we would usually be generalizing from a very small inductive base indeed, not tens of thousands of British swans but 1 or 2 studies, or in the best of cases, a handful.

It is as if we have forgotten the lessons about simple induction that have been rehearsed generation after generation for eons. Recall Bertrand Russell's chicken. She infers, on very good basis, that when the farmer comes in the morning, he feeds her. That inference serves her well till

Christmas morning when he chops off her head to serve her for Christmas dinner. Of course the chicken did not base her inference on a randomized controlled trial. But had we conducted one for her we would have obtained exactly the same results that she did. Her problem was not her study design but rather that she was studying surface relations. She did not understand the underlying socio-economic structure that gave rise to the causal relations she observed. So she did not know how widely or how long they would obtain. We often act as if the methods of investigation that served the chicken so badly will do perfectly well for us.

Why do we do this? I think there is a tendency to suppose that causality by its very nature must be general. Philosophers have certainly contributed to this supposition. The very influential philosopher from the end of the 20th century Donald Davidson for instance argued that it can be true that the event reported in column 1 of the first page of the *New York Times* can cause the event reported in column 2 on page 15 but only if there is some other descriptions of the first and second events---say the first is a C-type event and the second an E-type event---such that it is true generally that C-type events cause E-type events. Suppose we allow here that some kind of 'generality' is required: we suppose at least that if the circumstances were just the same again, the same outcomes would be produced; so, the same thing would always happen in circumstances just like this. The issue then is: How general must these circumstances be? How widely must they obtain?

Socioeconomic causal principles are not 'universal', as we suppose the law of gravity to be, but depend upon underlying 'local' structure. Rube Goldberg's pencil sharpener (look it up!) is my favorite example of the kind of thing I mean. There are two different kinds of reasons illustrated here for why we should not expect the causal claims we can nail down in our studies to hold generally. The first is that they link the wrong kinds of features. They link features that

can be described and operationalized very concretely. But general principles tend to need more abstract concepts. Rube Goldberg's pencil sharpener involves a number of general principles but they do not use concepts like 'lever' and 'pulley' and 'fluid escaping from a breach in a closed container'. Second, the causal claim that flying a kite sharpens pencils is not even an instance of a general principle, even if we look for more abstract descriptions to give to the cause and the effect. The kite flying sharpens the pencil through a series of steps: the kite flying opens a little door, the opening of the door permits the moths to fly out into the room, etc. Each of these steps is an instance of a general principle: the kite flying opening the door is in this situation an instance of the law of the pulley; the opening the door permitting the moths to fly into the room is an instance in this situation of the principle that a breach in a closed container allows fluids within to escape. And so forth. So: though each step instantiates a general principle, there is no general principle to connect the start to the finish, no descriptions, *C* and *E*, under which the initial cause and the final effect fall such that it is true that C-type events cause E-type events.

In his recent 'Risk Relativism and Physical Law' (2014), Alex Broadbent likens the assumption that relative treatment effects are universally transportable to the idea that we discover something like a physical law through an epidemiological study. Along these lines, I think that we treat the effect size that we discover in a study as a universal constant, like G in Newton's law of gravitation, or as economists treat a structural parameter in a structural equation, as invariant across circumstances. Doing so is absurd once you consider that the effect size is the net effect resulting from the distribution of the various support factors. To assume that the effect size is a constant is to assume that the distribution of support factors in a population is either (i) invariant across populations or (ii) varies but generally in such a way as to produce the same net effect. (i) is almost always untrue, since we can generally identify causally relevant

differences between two populations. We should only expect (i) to be true if we routinely design studies so as to be representative of the target situations to which we wish to generalize (which is seldom the case, as other considerations prevail). (ii) is simply fanciful.

Given these kinds of worries, the hope that we can arrive at general claims by simple generalization of study results is chimerical. How then do we support general claims? That's the rub. It cannot be done by recipe. To establish a general claim it takes a great deal of to-ing and fro-ing: an interwoven complex of conceptual development, theories---big and small, observation, experiment, analysis, modelling, reasoning, antagonistic assessment and severe testing. A credible general claim will rest on a *tangle of support*. There is a long history of vigorous debate about just what it takes to confirm general hypotheses but there is wide agreement that we should require a good mix of at least the following:

- Falsifiability.
- Observation of genuine instances of the hypothesis.
    - Including strong reason to count these *as* instances.

    What are the rules for linking abstract descriptions and more concrete ones, and how well substantiated are these rules. E.g. why in the case observed can we count the seesaw (or the tire jack) as a lever so that weighing your end of the seesaw down and thus raising your children in the air is an instance of the law of the lever? Why can we count a great circle as a geodesic on a sphere so that a body travelling along a great circle on a sphere subject to no forces is an instance of the general claim that bodies subject just to inertia travel on geodesics? Why can we count money or status as utility in cases where we see behavior as actions maximizing expected utility?

- A good fit with other established claims.
- Use of well characterized concepts with clear links to other well characterized concepts that have been shown to be theoretically and empirically useful.
    - With the aid of auxiliary hypotheses, the concepts used should be measurable in some circumstances.
- A good body of predictive successes when coupled with a variety of different sets of credible (plausible) auxiliary hypotheses.
    - That they are *different* sets of auxiliaries matters. The more different auxiliaries there are the less likely that the successful result is a consequence of compensating errors.
    - Most everyone looks for some of these to be *novel* predictions (the principal exceptions are certain Bayesians).
- Reasonable grounds against the truth of alternatives.

It also a considerable boon if, with plausible/credible auxiliaries, the hypothesis can be derived from a reasonably well established theory; and also if we have direct reasons to back up that the features generalized are (what philosophers call) 'projectible', i.e. that the features generalized are the right kind to hold generally. (For instance, in the case of the swans, we would want to know enough about the biology to know whether the connection between species and color is robust. What supports the assumption that all birds of the same species have the same color?)

This looks nothing like simple induction. Instances of the general claim are an ingredient in this mix and in many cases we may want to insist on the observation of genuine instances of the hypothesis before accepting it. Nevertheless, they make up only a small part of the tangle it takes to confirm the hypothesis. And they count for nothing at all unless we have sufficient

reasons to back up the assumption that they *are* instances of the general claim. In particular we need to establish that the concrete concepts in the claims we can nail down genuinely are instances in the case at hand of the abstract features that figure in the general claim we aim to support. If the effect of the kite flying on the little door in Rube Goldberg's pencil sharpener is to be taken as an instance of the law of the pulley, we need good reasons to suppose that in this case flying the kite *is* pulling on a pulley rope.

The process is iterative and complicated and it draws on a great variety of different resources. That is why I call it a tangle. This is just contrary to the impression you get from most evidence-based recipe guides, which do not advocate supporting general claims with a robust body of evidence, but rather with the best evidence of a certain kind (e.g. an RCT). Their image is of a hierarchy, rather than a tangle, of evidence.

None of this is simple, and none of it is recipe-like. But it is no good averting our gaze from the tangle to focus on the nails we can drive in by recipe. As I already remarked, a chain of inference is only as strong as its weakest link.

3. In Sum

I have stressed that study results can't do much of a support job on their own. But they are an ingredient, and are often taken to be an essential ingredient, in a mix of very different kinds of considerations that together can confirm a hypothesis. But which hypotheses can they help support? With respect to general claims:

1. Results can be instances of general conclusions that use different---often more abstract---concepts from those in terms of which the study is framed. In a spherical geometry, for

instance, motion along a great circle of an object subject to no forces is an instance of the law 'bodies moving by inertia alone travel on geodesics'.

2. A study result that depends on a chain of events supports any general claim instantiated at any step.

Both principles play a central role in confirming general claims in physics, where progress has depended on development of abstract, theoretical concepts and on tying these to more concrete experimental concepts, and where the bulk of the evidence for general principles comes from their display in complex settings, not in 'isolating' experiments.

With respect to local claims now:

1. Modularity: What happens in a specific setting in a specific process can sometimes support claims about what happens when parts of that process are inserted into other processes. This is of central importance in the natural sciences and in engineering, including social engineering.

There is an important lesson here for both general and local conclusions. When it comes to the questions, 'What conclusions can a result support?' Unfortunately, *you can't tell just by looking at the result itself.* My major point is that rigor matters. And where rigor matters,

- We need to know what we are talking about.
- A chain of support is only as strong as its weakest link.
- No result wears on its sleeve what it is evidence for.
- Rigor ≠ nailing down.

A tight tangle of diverse evidence is what delivers the kind of rigor we need for general claims.

References

Bareinboim, Elias and Pearl, Judea. 2013. 'A General Algorithm for Deciding Transportability of Experimental Results'. *Journal of Causal Inference*. 1(1): 107-134.

Broadbent, Alex. 2014. 'Risk relativism and physical law'. *Journal of Epidemiology and Community Health*. Published Online First: 14 August 2014.

Fuller, Jonathan. 2015. 'Myths and Fallacies of Simple Extrapolation in Medicine'. Ms.

Imbens, Guido. 2014. 'Instrumental Variables: An Econometrician's Perspective'. *Statistical Science*. 29(3): 323-358.

Petticrew, Mark and Chalmers, Iain. 2011. 'Use of Research Evidence in Practice'. *The Lancet*. 378(9804): 1696.