

**EVIDENCE-BASED POLICY:
WHAT'S TO BE DONE ABOUT RELEVANCE?**

For the 2008 Oberlin Philosophy Colloquium

Nancy Cartwright

INTRODUCTION

Evidence-based policy is all the rage now. It is mandated at the international, national and local levels and much money and effort has been devoted to providing advice and institutional structures to ensure that we do it and that we do it well. But mandates need policing, which creates a serious job for philosophers. If decision-makers are mandated to consider evidence seriously in their deliberations, guidelines are needed for what counts as evidence for policy and how it is to be used.

A wealth of such guidelines is now available in practice. That provided by SIGNS, the Scottish Intercollegiate Guidelines Network, is among the best. Here is another from the What Works Clearing House set up by the US Department of Education. Notice the similarities in the two. RCTs are the gold standard in both cases.

In this paper I want to make five basic points:

1. Practical guides are supposed to evaluate evidence for evidence-based policy. They tackle only half the job. They provide conditions under which evidence will be *sound*,ⁱ i.e. very likely to be true. But we are not interested in just any old true facts. The question of *relevance* is equally significant and little attention is given to that.

2. Even with respect to *soundness* the practical guides go astray, and on two levels. First they are unduly demanding, mistrustful and wasteful. Second they are exceedingly narrow in what they treat. Essentially they consider only claims about the *efficacy* of a treatment, where efficacy is a technical term having to do with the power of the treatment to produce an effect. But this is only a small part of the story when it comes to deciding what results will occur when the treatment is introduced on the ground.

3. Philosophers are better at relevance, Sherri Roush being one very good example. But most philosophical advice is too abstract to be of genuine use.

4. In any attempt to make matters more concrete, it is especially important to consider the *whole story* about what might happen given the policy, where, when, and as it will be implemented, which

5. This involves evaluating all the likely scenarios and their outcomes.

Before proceeding I should note that the advice guides I am discussing aim to help in the evaluation of the *effectiveness* of a proposed policy – Will the policy have the desired outcome? – and not with a variety of other questions, such as ‘Is it morally/socially/politically acceptable?’ or ‘How much does it cost?’ or ‘Does it have deleterious side effects?’ So I shall also confine the discussion here to questions of effectiveness.

TWO CRITERIA FOR GOOD EVIDENCE

For policy, indeed for any conclusion we are thinking of betting much on, we want sound evidence that speaks for the conclusion. Note that there are two criteria here. First, the evidence must be sound: Evidence claims should be pretty likely to be true. Second, the full body of evidence should make the conclusion probable, or probable enough given the size of bet.

Sherrie Roush's work on evidence stands out here in its insistence that we attend to both criteria at once. Roush requires that e have high probability if it is to be evidence, evidence for anything at all, and that the likelihood ratio of e for h be high if e is to be evidence for h , which is what in her view best ensures relevance.

SOUND EVIDENCE

In the ideal, evidence that supports adopting a policy should itself be very likely to be true. That's what I mean by *sound*. This naturally gives rise to the demand that $P(e)$ be high. We do not after all want to build our conclusions on shaky premises.ⁱⁱ

Evidence ranking schemes are designed to help answer the question, When will $P(e)$ be high? The RCT is the golden-haired boy here. But why?

Before turning to that question I want to note two limitations of the strategy the guides adopt for deciding what evidence claims will have high probability. First, they admit only claims produced by off-the-shelf test procedures that can be described in a subject neutral way. This ignores both claims that are backed up by a wealth of different successful

applications and those whose test procedures require subject specific assumptions, thus ruling out most of our most revered physics claims.

Second is in the *form* of the evidence claims. The aim is to assess claims of **effectiveness**: “Treatment T will result in outcome O when implemented when and how it will be in the target situation/population”.

What kinds of claims are needed as evidence for an effectiveness claim?

A very great many, I shall argue when I come to discuss relevance. But evidence-ranking schemes consider only evidence claims of one particular form, essentially, “T causes O in particular circumstances X in particular population Φ ”, ignoring both the kinds of information it would take to show that this kind of information is relevant to results in the target as well as all the other kinds of information needed about the target population to judge T’s effectiveness there.

To see the other limitations embodied in the ranking schemes let us now turn to the consideration of what is so good about RCTs.

In the first place RCTs are what I call *clinchers*. Given the right definition of ‘ideal’, it is possible to show that in an ‘ideal’ RCT a positive result deductively implies the conclusion under test:^{iiiiv} If there is a higher probability of O in the treatment group than in the control group then it

follows deductively that T causes O in the experimental population under the experimental conditions. This is great for soundness. If only the experiment has been carried out in an ideal way, something of course difficult to ensure, the probability of the evidence claim is very high indeed.

So RCTs are clinchers. But a large number of other methods are clinchers as well and that's true even if we restrict attention to conclusions of the form 'T causes O'. Among these is our own favourite in philosophy, the hypothetico-deductive method when used to show that a causal hypothesis is false. There are also a large number of econometric methods that when applied in the ideal deductively imply causal conclusions^v; there is process tracing, which can be deductive, and so forth. What is wrong with these other clinching methods?

The reason, I take it, is that RCTs^{vi} have another characteristic that seems to be enormously highly prized in the evidence-based policy community. They are what I call *self validating*. All methods have assumptions that must be met if the conclusions drawn from them are to be trusted. Econometric methods tend to require that we start with a full set of possible causes; the methods then tell which are actual causes. One central assumption for an ideal RCT is that confounders are equally

distributed between the treatment and control groups. Manuals for proper conduct of an RCT^{vii} list a variety of tactics to ensure that the requisite assumptions are met, including randomisation and quadruple blinding. The methods themselves, as laid out in these manuals, provide a check that the assumptions that make the RCT valid are met. By contrast in the case of econometrics there is no checklist to make sure that all the possible causes are in the model to begin with; this knowledge must be brought in from elsewhere. That's what I mean by saying that the RCT is self validating but econometric methods are not.

So RCTs are nice. They are both clinchers and self validating. But methods that don't clinch their results can still confer high probability on them. What is needed is an understanding and assessment of how to calculate the probability. Nor need a method be self validating. There is a lot we know and that we have worked hard to find out, indeed found out at vast expense of money and effort. To insist on self validation is to throw away hard-won knowledge.

So when it comes to helping evaluate the probability of evidence claims, the ranking schemes on offer have a number of deficiencies:

1. They consider only methods with a general structure that is repeatable and can be readily characterised without subject specific concepts.

2. They are most keen on clinchers and not very helpful about how to think about methods that merely vouch for their conclusions.
3. Among clinchers they consider only self validating methods, thus throwing huge amounts of hard-won knowledge straight into the bin.

RELEVANCE

Suppose though that we are happy with the advice offered about the soundness of evidence-claims. What are we meant to do with this evidence once we have it? The advice about that is thin. The US Department of Education website for instance explains that you have acceptable evidence for introducing a new programme into your school if the programme has passed two good RCTs in “schools like yours”. SIGNS is more informative. See slide 3. But not very much so. Can we do better?

RELEVANCE : FROM EFFICACY TO EFFECTIVENESS

The methods recommended by typical evidence ranking schemes are very good at establishing *efficacy*: Whether a treatment causes a given outcome in the selected population under the selected circumstances.^{viii} In evidence-based policy we are interested in *effectiveness*: What would happen were the treatment to be introduced as and when it would be in

the population of interest. How can we move from efficacy to effectiveness?

To do so we need an inference ticket. For a long time I have been selling inference tickets underwritten by an ontology of capacities. To see why, consider what might be considered a metaphysically cheaper inference ticket: *induction*. The RCT can establish a claim of the form ‘T causes O in population Φ in the circumstances of the experiment’. Perhaps we should just do an induction: ‘T caused O in population Φ administered in accord with experimental protocol, so T will cause O in the target population in the way in which it will be administered there’. That will almost certainly be a mistaken inference, even in cases when the results of the RCT are clearly relevant to what happens when T is implemented in the target situation.

One problem is with the *way* T is administered in the target and concomitantly with the outcome O itself. In the ideal experiment T and T ‘alone’ is changed but outside of the experimental situation T is likely to be implemented in a way that simultaneously introduces other factors that also bear on O. So it is unlikely that O will obtain. This doesn’t mean, though, that what T produces in the experiment is irrelevant; we expect it will still be part of the story about what happens outside the experiment.

It shows rather that we should not do a simple induction from ‘T caused O here’ to ‘T will cause O there’.

Consider the California class-size reduction programme^{ix}. The plan was backed up by evidence that class-size reduction is effective for improving reading scores from a well-conducted RCT in Tennessee. Yet in California when class sizes were reduced across the state reading scores did not go up.

There’s a conventional explanation.

- First, **implementation**. California rolled out the programme state-wide and over a short period creating a sudden need for new teachers and new classrooms. So large numbers of poorly qualified teachers were hired and not surprisingly the more poorly qualified teachers went to the more disadvantaged schools. Also classes were held in spaces not appropriate and other educational programmes commonly taken to be conducive to learning to read were curtailed for lack of space.

- Second, the **distribution of confounding factors** already in place may have been different in California from Tennessee. That is widely recognized to be likely.

In both cases we might still expect that the difference in scores seen in the RCT in Tennessee will still contribute in California even if the final outcome is not the same in the two states.

- Third, there may be something **structurally different** about California and Tennessee students that makes the two respond very differently to reductions in class size.^x

In this case, unlike the former two, the results in Tennessee will not be relevant to California.

My own solution of capacities to underwrite the inference ticket from efficacy to effectiveness solves the problem of the relevance of Tennessee to California. The conclusion to be transported from the RCT to the new situation is not that T causes O but rather that T *contributes* O^{xi}. This contribution is generally not what would happen were T to be introduced into the new situation but what O will ‘add’^{xii} in the new situation.^{xiii}

The paradigm is forces in mechanics. The force of gravity exerted by a large mass causes a smaller mass to accelerate towards it. Well, not really. Other forces might be at work as well, e.g. electromagnetic forces that are trying to accelerate the object in a different direction. The actual acceleration is a vector sum of the accelerations contributed by all the forces at work.

The logic of capacities – when applicable – thus solves all three problems in one fell swoop. Regarding problems of confounding and implementation, it accounts for the fact that we would not normally expect the same outcome outside the experimental setting as inside. *T contributes O* implies that what happens will have some systematic relationship with *O* but not necessarily be *O*. Regarding structural similarity, if the conclusion of an RCT can be cast as ‘*T contributes O*’ the issue has already been settled. Nothing is properly labelled a contribution unless it is ‘contributed’ in a systematic way across different populations and as other causal factors are added in different combinations (or at least across the range of populations and implementations presupposed).

Though capacities can solve these problems they are both epistemically and ontologically expensive. With respect to ontology it is obvious what

kinds of objections one might have, particularly if one is a diehard Humean. Epistemically we have just shifted problems from one place to another. The RCT can reveal T's contribution to O but only assuming that there is something properly labelled a 'contribution' in the first place. The information that T has the capacity to produce a stable contribution must come from somewhere else and backing up this claim will take a large body of different kinds of evidence.^{xiv}

Perhaps then we should find some other infrastructure to provide inference tickets from efficacy of effectiveness. My bet is that anything that works will be extremely expensive, at least epistemically. What matters for my concerns is that we come to understand how these inference tickets work and what it takes to support them. When we say that the results in Tennessee are evidence for claims about outcomes in California, *which* claims about outcomes are they relevant to, *in what way* are they relevant, and what does it take to support these relevance? Whatever inference ticket is used, a great deal more evidence than RCTs themselves will be needed to answer back up the answers.

FROM EFFICACY TO EFFECTIVENESS: THE WRONG ISSUE

It is easy to get sucked into the problem of how to get from efficacy to effectiveness – as I just did. But putting the question this way is back to

front. Finding a causal relationship under different conditions in a different population may not be evidence at all for effectiveness for your policy; but if it is, it is only one very small part of the argument. First for the reasons just described: Efficacy is no evidence whatsoever for effectiveness unless and until a huge body of additional evidence can be produced to show that efficacy can travel, both to the new population and to the new methods of implementation. But second and even more often ignored, efficacy is only one small piece of one kind of evidence. The last section on telling a variety of what-if narratives and assembling evidence for assessing their probability should point up one argument for that.

The general lesson, however, is easy to put. The focus on efficacy→effectiveness adopts exactly the wrong perspective. This is the narrow perspective of the experimenter, the experimenter who has worked extremely hard and produced a beautiful result – a ‘high quality’ claim one can be fairly sure is true. Now she needs to figure out where can she sell it. But the policy deliberator has no special concerns for this golden nugget. The experimenter asks, 'To what is my experiment relevant?' The policy maker instead needs to ask --- 'What is relevant to my policy hypothesis?'

RELEVANCE: A 3-PLACE RELATION

The issue of relevance should be separated into two pieces, which I have not so far done. The first concerns which facts or claims have a bearing on the truth of a hypothesis. The second is what probability the hypothesis has in the light of all those facts or claims that have a bearing on its truth.^{xv}

Before proceeding it may be helpful to lay out how I think about the problem of relevance in a policy setting. Suppose we are deliberating about a particular policy, *T*. We are looking for evidence to judge how probable it is that *H*: *if T were implemented – as and when it would be – outcome O would ensue*. Gathering facts and considering them are both costly. So we would like to assemble for consideration only facts that bear on the truth (or probability) of *H*. What criteria can help in deciding what facts to ‘buy’ to put on the table?

Many of our best philosophical criteria of relevance are probabilistic, e.g. *e* is relevant to *H* iff the likelihood ratio of *e*,*H* is high [$P(e/H) \gg P(e/-H)$], or iff *e* is probabilistically relevant to *H* [$P(H/e) > P(H/-e)$]. These are not so helpful in practice. We already know that we want as evidence facts that are connected with the probability of *H*. What we don’t know is what kinds of facts those will be. This is one advantage at least of Peter

Achinstein's account that e is relevant to H iff H and e are explanatorily connected (so long as 'explanation' does not itself reduce to some facts about the probability relations of e , H). It gives us a clue about what kinds of facts to look for. For instance, to tell if someone has a disease, we look for facts about the presence or absence of a cause of the disease, or about effects the disease produces or about other features one might have if one had been exposed to the disease.

Besides this practical problem, there is also a more principled problem to worry about with many probabilistic conditions for relevance. They make it look as if relevance is a two-place relation when in fact it seems in many cases to be 3-place: e is relevant to H assuming A . Think for instance about falsificationism and the hypothetico-deductive method, an excellent method for ruling out hypotheses. If the hypothesis up for evaluation implies e and someone offers to sell us very cheaply information about whether e obtains or not, then we ought to accept their offer. For if, when we open the report with their results, it turns out that e is false, we will know that our hypothesis is false as well. So in the sense of relevance as what facts we should like on the table, deductive implications of H are relevant.^{xvi} But now we have the notorious problem of Duhemian wholism. Interesting hypotheses do not by themselves imply facts that we can fairly readily learn about. More usually $H \& A \rightarrow$

e. So, $\neg e$ shows H is false only assuming A. It looks then as if e's relevance to H via the h-d method is *conditional* on A.

NARRATIVES

This conditional nature of relevance is all the more visible when we consider one straightforward way to assess the probability of our 'what if T happens' hypotheses – by working through the steps of the processes that might occur once T is introduced. First we evaluate the situation as it stands. Next we try to judge what changes T and its implementation might produce. Then we try to figure out what follows from that, and next from that, and so on till the point at which O should occur. Even at the first step we have three problems:

- We don't know all the features of the actual implementation that matter to the eventual occurrence of O.
- We don't know all the other factors relevant to O that won't be changed during the implementation.
- And even if we did know both of the above, we aren't sure what would follow from them.

So we begin to construct a variety of different narratives, some more plausible or more probable than others.

From this view point the probability of H is the probability of the set of narratives that start with T and end with O. So much of the relevant evidence will be claims^{xvii} that support one step or another in one of these narratives. But this leads to huge complications. A claim that supports a step in a narrative is relevant to H only if the narrative starts with T, leads to O or \neg O, and itself has sufficient probability to be taken seriously. But whether this last is true will depend on how well supported other steps in the narrative are, and that depends on what other claims support these steps. So a claim that is relevant to a step in a narrative is relevant to H only assuming the narrative has reasonable probability and whether that is true is relative to what other evidence supports other steps. So...how then do we talk coherently about the mutual dependencies in trying to develop a reasonable, and hopefully reasonably useful, account of relevance?

Finally, if relevance does become 3-placed, we not only have the problem of how to regiment that in advising what evidence to put on the table in deliberation. We also multiply problems at the last step – when we try to assess the probability of H in light of ‘all’ the evidence. The nice picture would have it that we finally end up with a set of evidence-claims of varying probability, each of which speaks either for or against H, where maybe we can and maybe we cannot say how strongly each claim

separately speaks. Our problem then is what kind of voting, or weighing, or amalgamation scheme to use to arrive at a final judgment of $P(H)$. Now, however, we are allowing that whether a given fact is evidence or not, and in what way, is conditional and that different claims on the table have different conditions on them. Indeed the very same fact may speak strongly both for and against H depending on different conditional assumptions. Amalgamation now seems a total nightmare.

SO....what can be done to make this more manageable?

IN SUM

What makes for good evidence for a policy hypothesis? I propose the question breaks into three others:

1. What are criteria for sound evidence: evidence that is likely to be true?
2. What practicable criteria can be provided for relevance: when does an evidence claim bear on the truth of the policy hypothesis?
3. How probable is the policy claim in light of 'all' the evidence?

The first two are questions about what should get on the table for consideration in policy deliberation. Help in evaluating not only yes-no

answers but also, *how* sound and *how* relevant will matter for deciding how much to pay to get an evidence claim into deliberation.^{xviii} Then given a fixed body of evidence, how shall we arrive at a final judgement of the probability of the policy hypothesis in light of it, especially keeping in mind that much of the evidence will only be conditionally relevant? I know we have good work already available on these issues, but I am hoping with this paper to generate interest in a great deal more work on them. For policy deliberation we need theories of evidence that deal with both soundness and relevance and that are at the same time both principled and practicable. That's what I hope philosophy can provide.

ⁱ Notice that I use 'sound' in a different way than Julian Reiss discussing similar issues in his *Error in Economics*.

ⁱⁱ I ignore here questions about whether the right kind and quantity of low probability evidence will suffice instead.

ⁱⁱⁱ Cf NC's Nature's capacities and Their Measurement, my BIOS paper, and references therein.

^{iv}

^v Cf. NC's HC&T and Julian Reiss...

^{vi} CUT: (and the other members in these lists – with the exception of the anomalous 'expert judgement')

^{vii} There are many of these – checklists running to 40 or more pages.

^{viii} For a formal definition of efficacy and a discussion of it see my paper on efficacy. Note that efficacy is really a three term relation, the efficacy of T for O relative to a given population and set of circumstances.

^{ix} CUT: which was instituted to improve California's notoriously bad schools at the height of the dotcom boom when the state had a lot of money

^x I think of this as a difference in the causal laws governing the two: T+specific arrangement K of confounders causes O in Tennessee; T+K cause O' ≠ O in California.

^{xi} More cautiously, we learn something about T's contribution.

^{xii} I put 'add' in scare quotes because in order for the language of capacities and contributions to be appropriate, O must contribute in some *systematic way* in new situations; 'addition' – e.g. the vector addition familiar from mechanics – is only one example. (For others, see NC's HC&UT and NC&tM.)

^{xiii} So far this is not taking into account other changes we also make in the situation in implementing T nor ways in which the causal factors already present in the target may have a different distribution than

they did in the experimental population nor that a different set of causal laws altogether govern the two different populations.

^{xiv} Note though that the backup needed for the inference on any particular occasion is weaker than a capacity claim. For any one inference we need only assume that what T produced in the experimental situation will ‘add’ in the way we suppose when T is present in the new circumstances and population. But as is usual in science, this weaker conclusion may be deemed implausible without the stronger to back it up. (Compare deflationary accounts of scientific realism. To back up any particular prediction we don’t need to accept the whole paraphernalia of theoretical claims and entities; we need only accept the consequence of those that directly underwrite the specific prediction.)

^{xv} CUT: The second is enormously complicated issue when it comes to understanding the force of ‘all’ in the last sentence. Does this mean all known facts, or all available facts, or all facts that we happen to have on the table, or all facts that we could get on the table had we world enough and time, or all facts that we could get on the table for some reasonable price, etc? I lay aside this issue for now and focus instead on the simpler, and probably antecedent, problem of understanding what facts are *relevant* to the truth of a policy hypothesis. But I should note that independent of questions about what counts as ‘all’, the issue of assessing the probability of the policy hypothesis in the light of the evidence will be complicated, particularly complicated, by an odd fact about relevance itself, which I shall raise here.

^{xvi} It will also be counted relevant on many more formal accounts as well. I bring this example up here to illustrate the point about relevance often requiring assumptions.

^{xvii} Throughout I am using a very general sense of ‘facts’ that includes general facts – like causal laws – as well as singular ones.

^{xviii} CUT: (since gathering and considering evidence are costly in both money and effort). The last is terribly complicated. How much is ‘all’? Presumably that will depend on the cost of additional evidence somehow balanced with its soundness and relevance against the cost of mistaken judgements without the new evidence somehow balanced with the probability of the mistaken judgement. This is no simple cost-benefit analysis.