

Evidence-based policy and its ranking schemes: So, where's ethnography?

Nancy Cartwright with Sophia Efstathiou

ABSTRACT

Evidence-based policy is widely mandated now throughout the UK, the USA and increasingly in Europe. Mandates invite policing and policing calls for standards for compliance. So there are now on offer a host of advice guides about what counts as good evidence. Most of these are ranking schemes, ranking schemes that, not surprisingly, do not rank individual evidence claims but rather rank methods for producing evidence. Ethnographic methods seem never to appear in these rankings, presumably dropping off the bottom (along with physics' favourite, derivation from theory, and biology's tracing of causal pathways), beneath even the lowly 'expertise', which is generally the only non-statistical method on the lists. This talk will pinpoint some genuine virtues of the statistical methods listed, more basic than the usual grounds ('guarding against bias') for endorsing them, as well as concomitant vices, and raise the question of how well ethnography can compete on the virtues and how much it can help make up for the vices.

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: if decision-makers are mandated to consider evidence seriously in their deliberations we need guidelines for what counts as evidence, evidence for policy, and how it is to be used.

A wealth of such guidelines is now available in practice. Look for instance at this scheme for ranking methods for producing good evidence provided by SIGNS, the Scottish Intercollegiate Guidelines Network, which we take to be among the best because it allows more flexibility than usual. [SOPHIA: slide here] Here is another from the What Works Clearing House set up by the US Department of Education. [SOPHIA: slide here] Notice the similarities in the two. RCTs are the gold standard in both cases. They are the basis for evidence of the very highest rank. What falls below are similar kinds of methods but ones putatively less good at controlling for bias. The only exception is the last and lowly 'expert opinion' – which doesn't appear at all on many lists.

We have two worries about this. First notice all that is missing. Notably for this conference – ethnographic methods are nowhere on the list. But you are not the only ones not invited to the party. Where are philosophy's favourites – the hypothetico-deductive method, which is widely believed in philosophy to be the method for testing physics theories, like Newton's and Einstein's? Or Bayes-nets methods? Or inference to the best explanation? Then there are the methods of econometrics, which are often

lauded over ethnography by advocates of quantitative methods. Notice they are not there either.

Second, 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. [SOPHIA: slide here] But not very much so.

We are not at all expert on ethnographic methods, which makes it odd for us to be addressing the issue of how ethnography might provide evidence for evidence-based policy. There is something relevant we think we can do however, which may be a help. We think we can isolate

- the special virtues of methods that are on the list, virtues that single them out from other methods as well as
- a gaping hole in the advice standards, or more a kind of deep canyon, that the advice guides do not tell us how to cross.

That should help show up just where ethnography might contribute, and how. In aid of this we shall argue two basic points:

1. Practical guides are supposed to evaluate the quality of evidence for evidence-based policy. They tackle only half the job. They provide conditions under which evidence will be *sound*,¹ i.e. very likely to be true. But policy deliberators are not interested in just any old true facts. The question of *relevance* is equally significant and very little attention is given to that.
2. Even with respect to *soundness* of evidence, the practical guides go astray, and on two levels. First they are unduly demanding, mistrustful and wasteful. Second they are exceedingly narrow in their conception of what kinds of results are relevant for judging the effectiveness of proposed policies. 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 ‘active ingredient’ in the treatment to produce an effect. But this is only a very small part of the story when it comes to deciding what results will occur when the treatment is introduced on the ground.

TWO CRITERIA FOR GOOD EVIDENCE

For policy, indeed for any conclusion one is 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 likely to be true; Second, the full body of evidence should make the conclusion probable, or probable enough given the size of the bet.

¹ Notice that we use ‘sound’ in a different way than Julian Reiss discussing similar issues in his *Error in Economics*.

Though they don't make the distinction explicit, advice guides concentrate on the first criterion. They tell when evidence is sound and they have much good advice to offer. But it is exceedingly narrow, which in the end can make for very bad advice. What they say about what they talk about is thoughtful and helpful; but there is the clear suggestion that what they don't talk about is beneath mention. This leads to ignoring – throwing out – lots of good evidence, evidence that could well change, or ameliorate, our judgements. So we shall start with these issues of soundness.

Before proceeding we should note that the advice guides we are talking about are aimed at helping 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 we shall also confine the discussion here to questions of effectiveness.

SOUND EVIDENCE

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

Evidence ranking schemes are designed to help answer the question, 'When will P(*e*) be high?' They do so by assessing methods for producing evidence. Some methods they claim are better than others, indeed much better. The RCT is the golden-haired boy here. Why?

Before answering that question we should like to remark on one general limitation of the evidence-ranking schemes: that is in the *form* of the evidence claims themselves. Ultimately the aim of this kind of evidence for evidence-based policy is to assess claims of *effectiveness*: claims of the form "Treatment T will result in outcome O when implemented". In order to assess such claims of effectiveness we are enjoined to look at sound evidence. But what kinds of claims might be evidence for such an effectiveness claim? A very great many, we shall argue when we come to discuss relevance. But evidence-ranking schemes consider only evidence claims of one particular form, labelled *efficacy* claims. The form is essentially, "Treatment T causes outcome O in special circumstances – those of the RCT – in some studied populations". It seems that only claims to the effect that T has caused O somewhere can be offered in favour of the claim that T will cause O somewhere else implemented in some different way. And that of course is just not the case.

Let us turn now to the question of why RCTs have pride of place in establishing efficacy claims – claims that T causes O in special circumstances in some studied populations. First, RCTs are *clinchers*: if the method is properly applied a positive result deductively implies the conclusion under test. That is, it is possible to characterise an ideal RCT in such a way that given some fairly natural assumptions

[†] We ignore here questions about whether the right kind and quantity of low probability evidence will suffice instead.

about connections between causes and probabilities, 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 of course when one wants to ensure that this causal claim has high probability. 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. That is the beauty of the RCT. It leaves no wriggle room.

Lower down the ranking lists there are methods that are not so tight. Or so it seems. In an ideal RCT, as 'ideal' must be defined in order of the proofs to go through, the full set of confounding factors for the effect of T on O must be equally distributed between the treatment and control group. That essentially guarantees that when a difference in the probability of O appears between the two groups no conclusion is possible except that T causes O. All other accounts for the difference in probability have been ruled out. In characterising an observational study or a quasi-experiment, however, even an ideal one, it would not be natural to write in as part of the formal definition that confounders are equally distributed. Rather the usual characterisation would build in only the requirement that all *known* confounders be distributed equally. In this case a positive result does not deductively imply a causal conclusion since not all possible alternative accounts for the probabilistic difference are ruled out.

So RCTs seem to have this nice advantage over all the other test methods considered in the list: They are clinchers. There are however a large number of other methods that 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 definitely false. There are also a large number of econometric methods for which we can show that they deductively imply causal conclusions when applied in the ideal. So the list in typical ranking schemes is strange. It ranks the RCT at the top, presumably because it is a clincher for causal conclusions. But it omits a number of other methods that can equally clinch causal conclusions. Why?

The reason we take it is that RCTs have in principle another characteristic that seems to be highly prized in the evidence-based policy community, namely that they are what we call *self validating*. All methods have assumptions, assumptions that must be met if the conclusions drawn from them are to be sound. One we mentioned for an ideal RCT is that confounders must be equally distributed between the treatment and control groups. [SOPHIA: note this is necessary for the conclusion to be properly *causal* though Fisher claimed something that sounds contrary – it's too complicated to go into here] Econometric methods tend to require that we start with a full set of *possible* causes; the methods are then good at telling which are *actual* causes. For an ideal RCT when we lay these assumptions out we see that in good manuals for how to conduct an ideal RCT (and there are many of these – checklists running to 40 or more pages) there are a large variety of tactics listed to ensure that the requisite assumptions are met. This includes randomisation and quadruple blinding. The point is that, as laid out in these guidelines, **the methods themselves provide a check that the assumptions are met**. The assurance that the assumptions are met does not have to be brought in from elsewhere as it does for instance in the case of econometrics, where there is no checklist to go through to make sure that all the possible causes are

represented in the model in the first place. That's what we mean by saying that the RCT is *self validating*.

So RCTs are clinchers and they are self validating. That is what is so nice about them. 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. And of course a method need not be self validating. Assumptions can be brought in from elsewhere; they needn't be just checked on the spot within the method itself. The demand that the only assumptions one is allowed to use in evaluating an evidence claim are assumptions that can be checked on the spot is a hugely wasteful one. There is a lot we know and that we have worked very hard to find out, indeed found out at vast expense of money and effort. To insist on self validation is to throw all this work straight into the rubbish bin.

So when it comes to evaluating the probability of claims about the efficacy of a proposed policy, the favoured method of the ranking schemes – the RCT – then has two special nice features: It clinches its results and it is self-validating. We doubt whether ethnography can compete along these dimensions. But the schemes also have two important deficiencies:

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

RELEVANCE : FROM EFFICACY TO EFFECTIVENESS

Look down the list of grades recommended by typical evidence ranking schemes. Notice first that these methods all look at probabilistic differences in outcomes in particular populations under particular circumstances; and second, as we have stressed, they all aim to establish the same kind of conclusion: a conclusion about the *efficacy* of the treatment with respect to the outcome.[†]

We are not however in evidence-based policy interested in the outcomes a treatment will produce in an experimental population under special experimental or quasi-experimental conditions. We are interested in what would happen were the treatment to be introduced as and when it would be in the population of interest, a question which, as we have noted, usually goes under the heading of the *effectiveness* of the treatment in the proposed setting. So how does one move from efficacy to effectiveness?

This is the very familiar issue of the *external validity* of the RCT result, so we will not say much about it, except to underline that two distinct moves need to be made:

- From the experimental population to the 'target' population. This can be problematic because the underlying social structure or mechanisms that

[†] For a formal definition of efficacy and a discussion of it see NC's paper on efficacy. Note that efficacy is really a three-place relation, the efficacy of T for O relative to a given population and set of circumstances.

support both the probabilistic and the casual relation in the experimental population may differ from those for the target. So efficacy in the experimental population need not translate to efficacy in the target. In addition, the actual outcomes we see in the experimental population may have little bearing on those that would occur in the target because the distribution of causal factors relevant to the outcome in place in the experimental and in the target may differ, even if the underlying social structure is the same.

- From the experimental implementation to the implementation that will actually occur. When we implement a treatment in an experiment, it is supposed to be the treatment and only the treatment that gets introduced. But when we implement a social policy we will certainly change a great deal more than just the specified treatment. We will have to do a lot to get the treatment into place and to ensure uptake. These very procedures can affect the outcome.

Though familiar these are not trivial worries; rather the reverse – and they both must be met if efficacy is to serve as evidence for effectiveness.

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 we almost 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 we 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 – and one clear place where ethnographic skills might matter.

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, a claim one can be fairly sure is true. But now she needs to figure out of what use this result can be, where on earth she can sell it? In thinking of evidence for evidence-based policy we need instead to take the view of the policy deliberator, who has no special concerns for this golden nugget and where it can be used. The policy maker instead needs to ask --- ‘What is relevant to my policy hypothesis?’ not the question of the experimenter who wants to learn, ‘To what is my experiment relevant?’ [SOPHIA: Use here slides on the car part scavenger with the mint condition part v the car designer gathering a huge variety of parts to build the car.]

RELEVANCE: A 3-PLACE RELATION

When it comes to relevance, we would like to separate the issue into two pieces which we have not so far done. The first issue concerns *which facts or claims have a bearing on the truth of a hypothesis*. The second issue is *what probability the hypothesis has in the light of all those facts or claims that have a bearing on its truth*. The second is enormously complicated issue when it comes to understanding the force of ‘all’ involved. 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 etc? We propose to 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 we 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 we raise here.

First it may be helpful to lay out how we think about the problem of relevance in a policy setting. For simplicity we narrow the context dramatically. Suppose we are deliberating about a particular policy, P. We wish to judge how probable it is that H: *if P were implemented – as and when it would be – a certain desirable outcome O would ensue*; and we wish to do so on the basis of evidence. But gathering facts and considering them are both costly. So we would like to assemble on the table 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?

The problem is much exacerbated by the fact that relevance is not really a 2-place relation – ‘e is relevant to H’ – but rather a 3-place relation: ‘e is relevant to H assuming A’. Think for instance about Popperian falsificationism and the hypothetico-deductive method, a very good 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 our sense of relevance as what facts one should like on the table for deliberation (pace costs of buying and considering them), deductive implications of H are relevant.⁴ But now we have the notorious problem of Duhemian wholism. Very often interesting hypotheses will not by themselves imply facts that we can fairly readily learn about. And this seems especially likely in the case of hypotheses like the ‘what-if’ hypotheses we have under discussion here. More usually $H \& A \rightarrow e$. So, e shows that 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.

This conditional nature of relevance is all the more visible when one considers one very straightforward highly reasonable way to assess the probability of a ‘what-if’ claim – by working through the steps of the processes that might occur on the way to the occurrence of O once P is introduced. First we think about the situation as it stands and about what changes P and its implementation will produce. Then we try to

⁴ It will also be counted relevant on many more formal accounts as well. We bring this example up here to illustrate the point about relevance often requiring assumptions.

figure out what follows next from that, and next from that, and so on till the point at which O should occur. Even at the very first step we have two problems:

- We don't know all the features of the actual implementation that matter to the eventual occurrence of O nor all the other factors relevant to O that won't be changed during the implementation
- Even if we did 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 P and end with O. So, much of the relevant evidence will be claims that support one step or another in one of these narratives. But this seems to lead to huge complications. A claim that supports a step in a narrative is relevant to H only if the narrative itself starts with P, 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 one 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 claims can be established that are relevant to 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 buy to put on the table in deliberations. 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 the probability of H. Now, however, we are allowing that whether a given claim 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 claim may speak strongly both for and against H depending on different conditional assumptions. Amalgamation now seems a nightmare.

TO CONCLUDE

SO...what can be done to make all this more manageable? This is really the question that we bring to you. Ethnographic methods do not have the special virtues possessed by the methods endorsed in standard ranking schemes: They neither clinch conclusions nor are self-validating. But have your methods other, more than compensating, virtues:

- Can they provide evidence of what outcomes would occur in the actual situation as the policy will actually be implemented? Perhaps by having a better understanding of the target population, of what factors are at work in it,

of what changes might occur as the policy is implemented, or what might result from the complex of causal factors eventually at work?

- Are there special skills and methods in ethnography for how to manage the complicated conditional evidence relations needed to construct plausible narratives of what might happen if the policy is adopted and evaluating how likely they are relative to one another?

Hopefully the answer to these questions is 'yes' because there are not many alternative fields that have much to offer here.