Evidence, Argument and Prediction

Nancy Cartwright

Abstract In this paper I propose a theory of evidence – which I call *the Argument Theory* – for domains where it is appropriate to demand high standards of rigor, explicitness and transparency, as in evidence for scientific conclusions and especially for evidence-based policy, which is where the need for such a theory first became apparent to me. I then apply the Argument Theory to answer a question that is too seldom asked, and never properly answered, in evidence-based policy where randomized controlled trials (RCTs) are taken as the 'gold standard' for evidence for predicting policy effectiveness: What does it take to makes positive RCT results evidence for policy predictions? The answer it turns out is quite a lot: information is required both about the causal role of the policy in the local circumstances and the helping factors required for it to work there.

1 The Context and the Problem

This paper is about evidence, specifically about evidence for *effectiveness predictions*: predictions that a well-described programme, policy or treatment will work for us, i.e. that the programme will result in an improvement in a well-specified outcome if we were to implement it in a targeted situation in a specific way – the way we would in fact implement it. Evidence-based policy advocates have invested a great deal of effort over the last few years in evaluating and providing warehouses for storing what they offer as evidence for hypotheses of this form in various areas of concern, warehouses to be visited by 'ordinary' policy makers and analysts. There includes for instance the Cochrane Collaboration for medical

N. Cartwright, LSE and UCSD (✉)
Department of Philosophy, Logic and Scientific Method, LSE, Houghton Street, London WC2A2AE, UK
e-mail: n.l.cartwright@lse.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_1,

studies, the Campbell Collaboration for general social policy, the US Department of Education's What Works Clearinghouse, the George Mason University Centre for issues in criminology and the greater London Authority's new Project Oracle for 'Understanding and sharing what really works' against youth violence.

These warehouses advertise that they store programmes that 'work' to produce targeted results. We as philosophers know to be wary of sloppy language like that. What they store are programmes for which there is very good evidence that they work somewhere, and, if we are very lucky, in a few somewheres. The warehouse keepers police certain kinds of scientific studies, studies that aim to establish causal connections between a programme and a targeted outcome. Programmes that make it onto the shelves in the warehouse are ones that have been tested in what the warehouse regulations regard as very good studies. In particular the warehouse purchasing rules strongly favour RCT study designs – that is, randomized controlled trials.

What an RCT can evidence directly is that the programme worked there, then, in the study population. What makes that evidence for the effectiveness claim of concern to policy analysts: 'It will work here, now, as we would implement it?' What does it take for the RCT result to play a part in a support structure that argues for the truth of the effectiveness prediction? That's my question.

I propose the same answer I urge for claims in any domain where the demands for rigor and explicitness are high, as in warranting conclusions in science or for evidence-based policy, namely what I call *the Argument Theory of Evidence*: conclusions are warranted by good arguments, arguments that are both valid and sound. It is surely trivial to remark that a conclusion is warranted by a good argument. But this reminder helps underline two important facts that are not currently at centre stage in discussions about evidence for effectiveness predictions:

1. Evidence is a 3-place relation: e is evidence for h *relative to* a specific argument A for h. Failing the rest of the premises in A, or relative to a different argument A′, the very same fact, e, can be totally irrelevant to the very same hypothesis h.
2. Arguments are like chains: they are only as strong as their weakest premise. Focusing on the argument forces the premises to the fore. Often it is just the ones that aren't generally stated that turn out to be most dicey.

2 The Argument Theory of Evidence

2.1 The Theory and the Reasons for It

What is evidence? More specifically, under what conditions is one empirical claim *e* evidence for a second empirical claim *h*? I note from the start that evidence is not a natural kind. There is no 'correct' theory of what evidence is, as there might be a correct theory of what an electron is. When this is the case our account of what makes for a good theory should be responsive to what needs the theory addresses. The theory I propose started as a theory of 'evidence for use', in particular for use

in making reliable predictions about what results will be produced by actions we consider taking. The Argument Theory is not confined to this context, however, but should fit anywhere we face the same needs.

A central problem I see everywhere that evidence-based policy is on the tapis is that the way the term 'evidence' is usually used lets in far too much. And it does so while at the same time purporting to be very restrictive by subscribing to the highest standards of rigour. In response my theory of evidence is demanding. That is because I agree with a common supposition. It is commonly – and I think reasonably – supposed that

Desideratum

A piece of evidence for a hypothesis should speak for the truth of the hypothesis.

It is with this in mind that I offer stringent criteria. I want criteria such that, once a fact meets those criteria, we should be happy to allow it to weigh in.

Here is what the *Argument Theory* demands of evidence:

> A well-established empirical claim *e* is evidence for hypothesis *h* relative to a good argument A (or A, A′, A″ . . .) if and only if e is a premise in A, which is itself a good argument for *h* (or, is a premise in A′ which is a good argument for a premise in a good argument A for *h*, etc.), where a good argument has true premises and is deductively valid.

The Argument Theory is akin to Clark Glymour's bootstrapping theory of confirmation (Glymour and Stalker 1980) in which we bootstrap from evidence to hypothesis using background assumptions and inductive logic. On the Argument account, given the other premises (which are like Glymour's background assumptions), h follows deductively from e. For Glymour, by contrast, the conclusion we derive from e and the background assumptions is an *instance* of h. Then we must use inductive principles to get from 'instance of h' to h.

For me h itself is the fixed point that we wish to arrive at. The Argument Theory requires that we do so by a good deductive argument. So I need far stronger background assumptions than Glymour. This, I urge, is all to the good. Science and evidence-based policy gain their high status in large part because they lay claim to being rigorous, public and explicit. These were the demands of Popper and of the Positivists and ones that we should insist on adhering to. There is in principle no objection to inferring h from instances of h in particular cases – so long as it is clear what it is about h and these instances that warrant this inference in this case. Are all the instances the same always? Are they at least all the same in this situation? Does the instance in question have special features that make it characteristic, so that if it holds, h holds? Or . . . ? The Argument Theory demands that the assumption that warrants the inductive leap be explicit in each case so it too can be subject to scrutiny. That's because hiding what it takes for the conclusion genuinely to follow from the evidence is both morally and intellectually culpable in any enterprise that sails under the flag of science or of evidence-based policy.

Two parallel lines of defence support the Argument Theory, one ontological, the other epistemological. That's because evidence is Janus-faced. On the one hand it has to do with truth and truth trackers: with what facts of Nature there are and what other facts can ensure they obtain. I should note that here I take a generous view

of what the facts of Nature include. In particular, facts can be expressed by general claims, like Maxwell's equations or the claims of general equilibrium theory in economics, as well as by singular claims, like 'The cat is on the mat.' On the other hand, evidence has to do with our attempts to arrive at truths: with our hypotheses about what facts obtain and the further hypotheses that provide warrant for them. The two lines of reasoning are obverse sides of the same coin, one expressed – to use Carnap's terminology – in the material mode, the other in the formal mode.

Begin with the material mode. Some facts or sets of facts are sufficient for others: if the first obtains, the second cannot fail to obtain. One fact, f_e, is evidence for a second, f_h, then if f_e is a necessary member of a set of facts sufficient to ensure f_h obtains. Note that 'sufficient to ensure X obtains' is not the same as 'brings X about'. It means just what it says: the one set cannot obtain if the other fails.

If you were brought up in the tradition of Hempel and Nagel you may well be more comfortable with the formal mode version of the parallel lines of defence. Evidence for a claim is supposed to contribute to warrant for the truth of the claim. What contributes to warrant for the truth of a claim are reasons, and what makes some claim e a reason for another h is that e figures in a good argument for h. 'Good' here = valid and sound; the premises are true and the conclusion genuinely follows from them. Deduction provides a clear sense to what it means for a conclusion to follow from a set of premises. It is the formal mode counterpart to one set of facts being sufficient in Nature to ensure that a second obtains.

Beware the formal mode though. We are looking for a formal mode counterpart of the relationship in Nature where one set of facts is sufficient for another to obtain. Then evidence can satisfy the Desideratum that a piece of evidence for a hypothesis genuinely speaks for its truth. That is the sense of 'warrant' involved in the formal mode account of evidence. Alternatively 'warrant for h' sometimes means 'justifying a belief in h'. That is not the sense at stake here. Belief is an attitude or an action, and, I would argue, there is no context-independent sense of justification for it. Whether it is justified to hold a belief in h depends on what is consequent upon believing it. Will God send me to hell for it? Will I build a bridge supposing h is true, which bridge will fall down if h is false? Will I teach it to my graduate students who might then win a Nobel Prize by taking it as the basis for their research or alternatively, fail to get their PhD because their research went nowhere? Still what I think about justifying belief is an aside since belief is irrelevant to my topic.

Evidence in the sense supposed in the evidence-based policy literature and in the sense required for establishing scientific hypotheses has nothing to do with belief. It has to do with the truth of empirical claims and with what facts ensure that truth. So inductive logics and subjective probabilities have no place in the characterization of evidence for these purposes. Of course they may, if you believe in them, play a legitimate role when it comes to our estimates of whether one claim is evidence for another.

The demand that an evidence claim figure in a good argument – both valid and sound – may seem excessively strong. I actually make a stronger demand. Not only should there be a good argument from e to h if e is evidence for h, but we should not

count e as evidence until that argument is displayed. I sometime express this in the slogan 'It's not evidence till there's evidence it's evidence.' C.G. Hempel's account of explanation also demanded validity and it also majored on deductive arguments. Hempel though allowed that many good explanations in science are enthymematic, in particular they are often not completely laid out. When it comes to evidence for scientific claims or policy predictions, I think it can be a bad mistake to allow this. Both science and evidence-based policy get their status in part from their claims to rigor. As a way to ensure rigor nothing beats laying out the arguments and looking to see how good they actually are.

2.2 Some Objections and Answers

There are a few objections philosophers may have right away to the Argument Theory of evidence. None, I urge, undermines the account.

- On this account of evidence we never know that a claim is evidence because that would require knowing that the claim is a necessary part of a good argument. To know the argument is good you need warrant for the other premises. To warrant those premises you need good arguments; to warrant that these arguments are good you need warrant for the premises in them. Etc, etc. That does not seem to me a problem: it's what good honest evaluation requires. Of course we stop somewhere; we have to. In the best of cases we stop with claims that can be taken as well established. To the extent that our stopping points are not ones we can take for granted, to that extent we should be cautious about our supposition that a proffered evidence claim really is evidence after all. We know from Otto Neurath that in reasoning we are like sailors who must repair our boats at sea without ever putting in to dry dock to build from firm foundations. So we will always have to trust to some claims we take as true, at least for the nonce. But we should not make our situation worse by neglecting our arguments: without laying out all the premises in all the arguments we don't know how leaky our boat is.
- The Argument Theory implies a number of what might be thought oddities, to all of which I have the same answer. Yes these facts are indeed evidence but they are not usually very useful pieces of evidence for us.
  - Anything true is evidence for a logical truth since anything – any claim at all – is a premise in a good argument for a logical truth. Yes, and so, I maintain, it should be. Anything is evidence for a logical truth. Still I wouldn't advise spending much to buy information about other facts to warrant a logical truth. If you know a claim is a logical truth you don't need to buy information about other facts to warrant the claim. And if you don't know that the claim is a logical truth, you will have trouble warranting that the claim is implied by the fact you buy. Still, if I don't know h is a logical truth but I am assured that if e then h, then e is surely worth learning.

  - *A&B* is evidence for A; $A > B$ & *A* is evidence for B; etc. Yes, they are. But we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here in evaluating the warrant for the conclusion.
  - Everything is evidence for itself. That's ok. Any claim does speak for itself. Again though, we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here either.

- The Argument Theory employs a flawed theory of relevance. It lets in as evidentially relevant just the kinds of things philosophers have been at pains to rule out. Consider the canonical example: 'John Jones takes birth control pills.' Surely this is not evidence for his non-pregnancy. But I think, to the contrary, that it is excellent evidence:

  1. Nobody who takes birth control pills gets pregnant.
  2. John Jones takes birth control pills.

  Therefore: John Jones does not get pregnant.

  Given (1), (2) speaks – and speaks compellingly – for the truth of the claim that John Jones does not get pregnant. What better basis could the truth of this claim have? To suppose that John Jones's taking birth control pills is not evidence for his failure to get pregnant is to confuse the task of providing evidence that a fact obtains with the task of explaining why it obtains.
- If all arguments are deductive then on the Argument Theory

  - There can't be both evidence for a claim and evidence for its negation since evidence claims must be able to participate in good deductive arguments and there can't be good deductive arguments for a hypothesis and its negation. That's okay too. It can still be reasonable to say 'We have evidence for h and evidence for not-h' when there are results that can figure in plausible arguments for h and results that can figure in plausible arguments for its opposite. What matters is that we recognize that the results only count as evidence relative to some good argument so that we don't just let the result weigh in without commitment to the existence of these arguments.

    Of course if there are good arguments that are not deductive and hence the truth of the premises does not guarantee the truth of the conclusion, then it can be literally true that there is evidence for both h and evidence for not-h on the Argument Theory of evidence. But that is as it should be.

  - There can be no evidence for false claims. As soon as you know there is evidence for h by lights of the Argument Theory, you know that h is true. But that seems to me no problem. The problem is coming to know that e is evidence for h. This is a serious job and one of my concerns with the evidence-based policy literature, as I shall explain tomorrow, is that it does not take the job seriously enough, while all the while boasting that that is just what it does.

> Although I don't think our ordinary locutions count for much in efforts like mine here to make precise an everyday concept like evidence so that it can serve specific scientific purposes, I'll just note that often we do use the term 'evidence' in a way that supposes that there's no evidence for false claims. If I am accused of cooking the books or murdering Ackerly, I might very well respond, "But you couldn't have evidence for that. I didn't do it."

2.3 An Alternative Account of Objective Evidence and Why I Do Not Adopt It

My insistence that in science and policy we want a sense of evidence in which evidence for a hypothesis speaks for its truth echoes views of Sherrilyn Roush, who has done a great deal of very instructive thinking about evidence. In her book *Tracking Truth* (Roush 2005), Roush links a theory of evidence with her theory of knowledge – where the latter has to do with what we are entitled to claim for ourselves as knowledge. Her very first sentence in the chapter 'What is Evidence? . . . ' is on the knowledge side: 'It is a truism that the better one's evidence for a claim p the more likely one is to have knowledge that p.' [p. 149]. But like me Roush is keen to keep the enterprises of theory of knowledge and theory of evidence separate:

> . . . the notions of evidence that I am aiming for are objective in the following sense. That e is evidence for h is understood as holding in virtue of a factual relation between the statement's being true and the statement h's being true, not in virtue of anyone's believing that this relation exists. [p. 156]

Her basic idea is this: 'Intuitively, good evidence for a hypothesis is a discriminating indicator of the truth of the hypothesis,' [p. 154] where 'discriminating indicator' means some appropriate probabilistic analogue of 'h is true if e is true and false if e is false'.

Formally Roush's account of evidence requires that for good evidence:

- P(h/e) be high.

In order to satisfy what she calls the *leverage condition*, Roush in addition requires:

- The likelihood ratio [P(e/h)/P(e/−h)] be greater than 1.

Moreover it is highly desirable that

- P(e) be high.

There are a number of reasons that I do not adopt Roush's account, hinging primarily on the fact that it is still too much of a hybrid between a theory of evidence and an account of how to justify our claims to knowledge.

Where our accounts part company at the start is over what Rush calls 'Bayesianism'. For her this does not mean a subjective interpretation of probability. Rather –

'the Bayesian makes the idealizing assumption that all statements of the language in question possess probabilities. This is in contrast to the approach of classical statistics in which it is denied, for instance, that hypotheses have probabilities.' [p. 155] For the objective notion of evidence that Roush and I both have in view, though, it cannot be probabilities of statements that matter but rather probabilities of facts. I do not see that there generally are such probabilities. Probabilities for facts arise from chance set-ups, which are a special kind of nomological machine (Cartwright 1999), and while nomological machines are not all that rare, those that count as chance set-ups appear to be a small subset.

Then I disagree with each of her conditions in turn.

- P(e) is high. Roush insists on this in a debate about whether evidence should be surprising, which many, Bayesians especially, require. Her discussion at this point repeatedly refers to degrees of belief despite the fact that she means to be embarked on an objective theory of evidence. And I think that's a clue. If we are thinking about a license to 'accept' h, there are a variety of reasons to value observing consequences of h that were not expected beforehand: like worries about accommodation rather than novel prediction, or the demand that h have content that goes beyond summarizing what's already known.
- Notice I say here 'expected' – that has to do with subjective probabilities which are not relevant to the objective notion of evidence. On the objective side, I urge that e should be true, not objectively probable. High probability of e only comes in as a demand when we consider whether we should 'accept' that e is evidence.

  Consider an example where we might all be willing to suppose there are objective probabilities. We have three coins:

  – For C(1), P(h) = .2
  – For C(2), P(h) = 1 . . . .it is two-headed.
  – For C(3), P(h) = .2

  Imagine that the following chance-se-up is in place from time t(1) through time t(3):

  – At t(1), flip coin1
  – At t(2), if C(1) = h, at t(2) flip C(2)
    At t(2) if C(1) = t, flip C(3)
  – At t(3) either h occurs on C(1) or either heads or tails on C(2).

  Now consider e = 'C(1) = h at t(2)' and h = 'heads occurs at t(3)'. P(e) = .2. That is low. But e is compelling evidence for h. What I want to underline is that it is compelling evidence not despite its low probability but regardless of its probability. It would be evidence no matter what its probability. Even though e has an objective probability, that objective probability is irrelevant to its status as evidence. This claim is true in general I maintain.

What I would say about e is this: 'C(1) = h at t(2)', if true, is evidence not for h but for h' = 'At t(2) the objective probability of heads at t(3) is .2'. This I think is the right thing to say and it is what follows on the Argument Theory.

- The likelihood ratio is high. This is in aid of leverage. Roush tells us: '. . . evidence provides leverage on the truth of claims about the world. Specifically, knowing that the evidence statement is true is usually a lot easier than knowing that the hypothesis statement is true, and we use the former to help us make progress on the latter where we could not have made progress directly.' [p. 158] Damien Fennell and I have elsewhere (Cartwright and Fennell 2009) explained problems we have with thinking the likelihood ratio can provide leverage in the way Roush wants. I won't rehearse those worries here but rather make a more general point. I don't see how to justify any condition that demands leverage in this sense for an objective notion of evidence. Leverage clearly makes sense when we are in the business of justifying our claims to knowledge or trying to estimate what to expect in the future. Suppose e, if true, is evidence for h. There is no point in spending a lot of money to learn whether e is true or not as an aid to deciding whether h is true when it is a lot cheaper just to learn h directly. But that has nothing to do with whether e is evidence for h or not.
- P(h/e) is high. Suppose this is so and P is an objective probability and e is true. Then the objective probability of h is P(h) = P(h/e) and on the argument account e is good evidence for this – and that is so whether P(h) is high or not. For Roush it is also evidence for h. One could make this stipulation as part of an objective account of evidence but I think it is misleading. We don't have evidence that h will obtain, just that it can, or might or might well; more precisely, that it has probability P(h) of obtaining. There may be no harm in adding Roush's requirement to the argument account but it will mean that there can be good evidence – in the fully objective sense – for false h's, not just evidence we mistakenly thought was good. Evidence does not provide the same assurance as it does on the basic Argument Theory.

  Also, note that if it is added as an allowance on the Argument Theory it would play a different role than in Roush's. For Roush this is what secures the relevance of e to h. On the Argument Theory, that is secured by arguments linking e and h. And that demand should be enforced here as well. We should still demand a good argument – valid and sound – for the claim that P(h) = φ.
- There is one other feature on which Roush and I differ but not, I think, disagree. That is on *discrimination*. For Roush e should track h; bracketing issues about probabilities, e should be true if h is. The Argument Theory requires only that h be true if e is. One could perfectly well add this. 'Evidence' even 'objective evidence' is not a natural kind with a fixed criteria or a fixed extension. I do not wish to opt for this stronger notion since it is far stronger than what seems supposed in the evidence-based policy literature and in the bulk of scientific cases I am familiar with. In particular it would undercut the claim that positive

results in ideal RCTs are evidence for causal claims since positive results imply a causal connection between treatment and outcome but negative results do not show there is none.

- I also have a worry about probabilistic characterizations of evidence like Roush's even when the topic is not objective evidence but rather our entitlement to hold some cognitive attitude to a hypothesis or to use it in some way: probabilistic characterizations put the cart before the horse. Subjective probabilities, at least when we employ them in serious decision making, should have reasons behind them. Like what? Conditional probabilities generally play an important role, like P(h/e). How do we set that? One standard way is look to see if e is evidence for h and how strongly it speaks for h's truth, then set the probability of h given e accordingly. But to do that, we need some independent way of characterizing evidence that does not depend on our subjective probabilities.

3 What Makes RCTs Evidence for Effectiveness?

I have rehearsed the Argument Theory of Evidence because it can provide us with an answer to this question and an answer that matters to getting our predictions right in evidence-based policy.

The current evidence-based policy literature rates positive outcomes in well-conducted randomized controlled trials as gold standard evidence for predictions that the treatment in the trail will work if we implement it in our setting. So, what's the argument?

RCT results are normally *effect sizes: ES* = df the difference in the expectation of the outcome (y) in treatment group and in the control group ($\mathrm{Exp(y)_T - Exp(y)_C}$). Causes do not, we suppose, produce their effects willy nilly, at least not where prediction is possible. Rather these effects are generated in accord with causal principles. We can without loss of generality suppose that these principles are of this form[1]:

$$\mathrm{CP: y(i) = a + b(i)x(i) + z(i)}$$

where $y(i)$ is the outcome for individual i in the population where the principle holds, $x(i)$ is the treatment variable, a is a constant and $z(i)$ represents all the other casual clusters that contribute linearly with x to produce the value of y in i. It is apparent from this principle that x is a genuine contributor to y for at least some individuals i in this setting if $\mathrm{b(i)} \neq 0$ for some i. A well-known argument – which I shall call the *RCT Argument* – shows that, under usual assumptions about ideal RCTs,

$$\mathrm{ES = Exp\,(b)\,\left(X - X'\right)}$$

[1]The results I shall describe are essentially the same for more complicated functional forms.

where X = the value of the treatment variable in the treatment group and X′, the value in the control group.

RCT Argument

1. $y(i) = a + b(i)x(i) + z(i)$
2. $ES = Exp\,(y(i)/x(i) = X) - Exp\,(y(i)/x(i) = X')$
   $= Exp\,(a/x(i) = X) - Exp\,(a/x\,(i) = X')$
   $+Exp\,(b(i)/x(i) = X)X - Exp\,(b(i)/x(i) = X')\,X'$
   $+Exp\,(z(i)/x(i) = X) - Exp\,(z(i)/x(i) = X')$
3. x is probabilistically independent of b and w.

Therefore $ES = Exp(b(i))(X-X')$.

Premise (3) is supposed to be guaranteed by random assignment of individuals to the treatment and control groups and by masking, quadruple masking if possible. I shall suppose that it holds by definition in an *ideal RCT* and henceforth consider only ideal RCTs. We should remember of course that real RCTs are generally far from the ideal and that randomization only assures the independence assumptions in the long run were the same experiment repeated indefinitely.

So for an ideal RCT, if the effect size is positive, so is Exp(b) which means that b is positive for at least some i. So x is a genuine contributor to y for some individuals in a population subject to CP. This shows that there is a good argument, A', that has among its premises the evidence claim

> e = df 'The effect size of x for y in the population in a well-conducted RCT is ES > 0.'

and has as its conclusion

> h_1 = df 'x is contributes to the production of y for some individuals in the population in that study.'

So e is evidence for h_1 *relative to* the RCT Argument and thereby relative to the other premises in that argument (including especially the assumption that conducting the experiment well – randomizing, masking, etc. –delivered the features an ideal RCT is supposed to have). To establish e's evidential relevance to effectiveness prediction h, we now need to find an argument – a good argument – that I shall call the *Effectiveness Argument*, in which h_1 figures essentially as a premise and h as conclusion.

Before I propose one, I want to point out something about CP, which is often subject to a grave misunderstanding, one that I hope the reader won't have been led into because I was careful with the notation. Often CP is written with the reference to the *i*'s implicit, so it looks like this:

$$CP' : y = a + bx + z.$$

In this case it is easy to suppose that b is a constant. But there are few treatment variables x for which this is likely be the case. After all, the treatment is usually

only the salient factor, or the factor of focus, in a cluster of factors that together are sufficient to produce a contribution, that is, sufficient *when they all take the right values at once*. To use the terminology of JL Mackie (1965), x is a cause, yes; but it is an INUS cause of contributions to y: it contributes to y, but only when operating in cooperation with helping factors and often a great many of these. In CP, b(i) represents in one fell swoop the values for i of all the helping factors that are necessary along with x to ensure a contribution to y.

Now to the argument. First we need to formulate a conclusion properly. One version would be

> h_{ES} = df 'If x = X were introduced in our setting, as opposed to x = X′, keeping fixed all the other causes of y in our situation [except those downstream from x], the effect size would be ES for us too.'

So, will x make the same average contribution; that is, is the efficacy, which is measured by the treatment effect in the study situation, the same there as here. Certainly if the same principle holds there as here, *a* will be the same since it is constant. But *b* is not a constant; and the effect size is its expectation – that is, the effect size is an average over x's supporting factors. The average in each situation depends on the distribution of these in that situation. Even if the same principles govern the two, that is no reason to suppose the distributions of support factors would be the same. To the contrary in fact, this distribution very often heavily depends on local circumstances so it is unlikely to be the same.

Anyway, the same distribution is not really what you hope for. What you'd really like is that you have – or can arrange to have – a distribution that favours the good values of b – the ones that provide the largest contribution from the programme. At the least, you will want to have some values for which x's contribution is positive and these should outweigh the effects of those that make x's contribution negative; and if getting negative contributions in some individuals in your setting is to be avoided, then you don't want any of these at all.

Suppose though we can lay aside worries about negative contributions in some individuals. Suppose we want to predict simply

> h_{cont} = df 'If x = X were introduced in our setting, as opposed to x = X′, keeping fixed all the other causes of y in our setting [except those downstream from x], a positive contribution would result for some members of our population.'

What does it take to make ideal RCT evidence relevant? I am going to talk, for short, about whether x *can play a causal role* in the production of y – is it genuinely there in the principle for the production of y for some individuals? Here then is what I take it is the weakest valid argument that uses the results we can get from an RCT there as a premise and concludes that the programme or treatment will contribute positively for some individuals here.

Effectiveness Argument

1. x can play a causal role in the principles that govern y's production there.
2. x can play a causal role in the production of y here if it does so there.
3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.

Therefore, x can play a causal role in the production of y in some individuals here and the support factors necessary for x to make a positive contribution are present for at least some individuals here (i.e. x contributes to the production of y for some individuals here).

Where then does the RCT come in? It enters in a different argument, an argument that supports premise (1). That is why I talked earlier about what a study can evidence *directly*. As I use this term, a well-warranted empirical claim *e* is *direct evidence* for a hypothesis *h* if e figures essentially in a good argument for h – a valid argument with well-warranted premises. Now the RCT Argumnet is a valid argument that takes as premise a positive effect size in an experiment and as conclusion, that the programme contributes to the targeted outcome there in the study situation (post implementation). The other premises in the RCT Argument have to do with further features of the study; for instance that confounding factors are independent of x. The keepers of the evidence warehouses police these premises for particular studies: they judge how well-warranted the other premises in an argument like the RCT Argument are, mostly on the basis of the study design. So if we find a programme in a conscientious warehouse, we have good reason to think there is a good (valid and sound) argument like A′ to warrant the claim that x plays a causal role somewhere – there in the study setting. And that is the first premise in the Effectiveness Argument.

So the RCT result can be evidence for effectiveness here, but it is only *indirect*. It is not a premise in an argument for effectiveness but rather a premise in an argument for a premise. Moreover, its relevance is conditional, highly conditional, since it depends on the validity and the soundness of both the RCT and the Effectiveness Arguments. As in this picture, a positive effect size in an RCT is leveraged into evidence that the program works there (in the RCT setting) by the RCT Argument; and 'it works there' is leveraged into evidence for 'it works here' by the Effectiveness Argument; if either argument fails, the lever drops and evidential relevance disappears with a thud.

Both the RCT and the Effectiveness Arguments are valid, so what really matters is their soundness. We may take it for granted that the RCT Argument is pretty good if we find the programme in a reputable warehouse. What about the Effectiveness Argument? What ensures that its premises are well-warranted? Recall, the two additional premises necessary are:

2. x can play a causal role in the production of y here if it does so there.
3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.

What further arguments support these premises? That's the problem. There are no warehouses for information like this, and the kind of information needed is really hard to come by. I don't see how (2) can be supported without a great deal of theory; so too with (3), in order to identify what the requisite support factors *are*. Then, in addition, (3) will require a good deal of local knowledge to determine if we have here even some of the right values for the support factors, let alone a desirable distribution of them.

Before returning to my overarching message, let me take up two objections to my account of what can count as warrant for an effectiveness prediction beyond the earlier objections to the Argument Theory in general.

First: RCTs are often advocated by people who don't like theory – they think our claims to theoretical knowledge are too slippery; they just don't want to trust to them. That means they don't like my view about how (2) gets warranted. They have an alternative proposal: more and more RCTs, with as much variation in circumstances as possible. I agree that more RCTs, and especially across a variety of circumstances, can improve the warrant for an effectiveness prediction. It does so by supporting a premise like (2): the program plays a causal role here. How? That's the rub. The argument could be by simple enumerative induction: swan 1 is white, swan 2 is white . . . ; x can play a causal role in situation 1, x can play a causal role in situation 2, . . . And how good is that argument? For induction we need not only a large and varied inductive base – lots of swans from lots of places; lots of RCTs from different populations. We also need reason to believe the observations are projectable, plus an account of the range across which they project. Electron charge is projectable everywhere – one good experiment is enough to generalise to all electrons; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

I urge that rather than some weak inductive argument, we need a rigorous deductive argument. Then we know just what we are betting on when we bet on the conclusion. So I would add a premise to the effect that x can play the same causal role here as in all those other places, add it so that the challenge is clear: just what is the warrant for this very strong claim? That matters because of the weakest link principle: the conclusion can never have any more warrant than each of its premises individually.

The second objection is this. Surely the best evidence that the program will work here is an RCT here. I agree this would be good evidence – let's not quarrel about 'best'. *Would be* were it possible. But we never do an RCT here really, here on the same population at the same time. And both matter. A sample is almost never going to be a representative. Representative: that means governed by the same causal principles and having the same probability distribution over the causally relevant factors. And time certainly cannot be ignored. Are the causes the same now as they were when the study was done? That's a particularly pressing question for socioeconomic programme since economists from JS Mill to the distinguished British econometrician David Hendry have worried that past regularities are a poor guide to the future in economics, just because the background arrangement of causes shifts so often, and so unpredictably. Of course the experimental population could be representative enough and the causes at work stable enough. Let's just get this stated explicitly as one of our premises. Then we can think about what warrant there is for these assumptions in our case.

4 Conclusion

That returns us to my overarching point. Evidence is a 3-place relation; e is evidence for h only relative to some argument or other. That is not a new idea at all, and it may not be very controversial. But taking it seriously matters. It is altogether too easy, when we do not keep the arguments to the fore to overestimate the warrant that our studies can deliver. The RCT is a good example. It is widely taken in the evidence-based policy literature as gold standard evidence for effectiveness claims. Though perhaps with a caution. The US Department of Education, for example, warns that trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students. This kind of warning simply conceals what needs to be exposed. What is the argument that makes a particular RCT result evidence for a particular effectiveness prediction? As we have seen, if evidence, it is indirect evidence – there are layers of arguments to get from the study result to the effectiveness conclusion. And they all have additional premises, every one of which, along the way, is essential for the security of the final conclusion. No matter how firm the RCT result is, the effectiveness conclusion – for which it is supposed to be gold standard evidence – can have no greater claim to knowledge than the shakiest of these.

Nor is this unusual. Most of our knowledge claims, even in our securest branches of science, rest on far more premises than we would like to imagine, and far shakier. This recommends a dramatic degree of epistemic modesty. Most of us have adjusted to Neurath's lesson that we are like sailors rebuilding our boat at sea. The conclusions I draw about evidence and the amount of warrant it can confer point to his less familiar warning: the boat is far leakier than we like to think.

Acknowledgement I want to thank Alex Marcellesi and the members of both the EPSA audience and the audience for my Pufendorf lectures who participated in the discussion for help with the immediate contents of this paper and the AHRC, the British Academy, LSE's Grantham Research Institute on Climate Change, the Spencer Foundation and the Templeton Foundation for support for the research for it.

References

Cartwright, N. (1999). *The Dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, N., & Fennell, D. (2009). Does Roush show evidence should be probable. *Synthese, 175*(3), 289–310.

Glymour, C., & Stalker, D. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*, 245–264.

Roush, S. (2005). *Tracking truth*. Oxford: Clarendon.