To Appear in a volume in honor of Peter Achinstein edited by Gregory J Morgan

Evidence, External Validity and Explanatory Relevance

Nancy Cartwright[1]
LSE and UCSD

1. Introduction

When does one fact speak for another? That is the problem of *evidential relevance.* Peter Achinstein's answer, in brief: Evidential relevance = explanatory relevance.[2] My own recent work investigates evidence for effectiveness predictions, which are at the core of the currently heavily mandated evidence-based policy and practice (EBPP): predictions of the form 'Policy treatment T implemented as, when and how it would be implemented by us will result in targeted outcome O.' RCTs, or randomized controlled trials, for T and O are taken to be the gold standard for evidence for effectiveness predictions. I question this: Not just whether they are gold-standard evidence, but more, How can they be evidence at all? What makes them relevant to the truth of the prediction that T will work for us?

I am going to follow Achinstein's lead here and suppose that evidential relevance = explanatory relevance, where A is explanatorily relevant to B just in case A is an ineliminable part of a correct explanation of B, or the reverse or A is indirectly relevant to B: There is some common fact that is an ineliminable part of correct explanations for A and B. I shall argue:

1. It's not evidence for us without evidence that it's evidence.
2. Evidential relevance is a conditional relation: E is evidence for H conditional on the non-shared factors that fill out explanations for E and H. Finding these involves a *horizontal search*.
3. To get shared explanatory elements we need a *vertical search*, up and down the ladder of abstraction. If we haven't climbed the right ladder in the right way, an RCT may not show what we think it does.

It follows from my discussion that RCTs cannot play anything like the central evidential role for effectiveness predictions that they are standardly awarded in EBPP literature.

[1] Research for this paper was supported by grants from the British Academy to study evidence for use, the LSE ESRC Centre for Climate Change, Economics and Policy and the associated Grantham Centre and a UK AHRC grant to study evidence related to child welfare policies. I am grateful to all three for financial support and to collaborators on them for intellectual support. I am also grateful to Eileen Munro for help with the child welfare example and to Adam Spray and Ravit Alfandari for help in editing.

[2] The ideas of Peter Achinstein I draw on here are primarily from Achinstein 2001, 1983. But of course also from his long series of works over three decades from Achinstein 1978 onwards.

I begin with some terminology, some assumptions, and some simplifying procedures. First, the fact that effectiveness predictions are predictions should not put us off an explanatory relevance account. Just suppose that the predictions are true. Then look for explanatory relevance.

Second, I adopt the probabilistic theory of causality. I suppose that for each effect-type at a time t, O^t, and for each time t' before t, there is a set of factors $\{C_1^{t'},...,C_n^{t'}\}$ – the causes at t' of O at t – whose values in combination fix the objective chance at t' that O takes value o for any o in its allowed range. A *causal structure*, $CS^{t'}(O^t)$, for O^t is such a set along with the related objective chances for all values of O^t for all combinations of allowed values, $L_j^{t'}$, of the causes in the set: $Prob(O^t = o/L_j^{t'})$. For simplicity I will usually suppress time and other indices and also restrict attention to two-valued variables. So a causal structure looks like this: $CS^{t'}(O^t) = <\{C_1^{t'},...,C_n^{t'}\}, \{Prob(O^t/L_1^{t'}),...,Prob(O^t/L_m^{t'})\}>$.

Third, I follow the EBPP literature and concentrate on the effect size of T for O in a population: Prob(O/T) – Prob(O/-T).[3]

Fourth, I restrict attention to predictions about the effects of policies on populations and not on single units.

Fifth, I consider only positive relevance since that fits in a simple way within Achinstein's explanatory account.

Sixth, I concentrate on cases where E is indirectly relevant to H because these are the most complicated cases.

Finally, for simplicity I assume that the evidence claims in question are well-confirmed – we can reasonably take them as true.

2. Relevance is conditional on unshared factors

The relevance relation I focus on is objective: One fact (E) *bears on the truth of* another (H). This relation holds between facts because of the way nature and society operate; it does not depend on our knowledge of this operation. There are corresponding epistemic notions – like our reasoned judgments about what is relevant to what. These do depend on the state of our knowledge and a variety of other factors as well, such as time and resource constraints or level and type of expertise. Objective relevance is important for policy deliberation predictions: Gathering, discovering, and surveying facts are all costly. We'd like to confine our attentions to facts that matter to the truth of the policy prediction.

Deleted:

'Bears on the truth of' can seem hopelessly vague. So there are various well-known attempts to explicate it with more familiar notions. One takes relevance to be some kind of causal relation. That's too narrow. So too are various kinds of probabilistic relations: There just aren't enough of these in the world to account

for all the obvious evidential relevance.[4] Moreover, relying on probabilities puts the cart before the horse when it comes to the needs of estimating if a policy will work. Achinstein's explanatory relevance by contrast fits the bill nicely.

Why should explanatory relevance be a good stand-in for the more abstract concept 'bears on the truth of'? My answer is a mix of views of Achinstein and of my own. Just as the relevance relation aimed for is objective, so must the explanatory relation be in order to serve as a marker for relevance. 'Explanation' as I use it, then, doesn't mean something that has the right form and is proffered as an explanation; it means something that *is* an explanation. There will be many of these, some of them nested, which is why, as I argue in Sections 4 and 5, we need good vertical searches to find the widest scope of evidential relevance a result can have.

Achinstein has been criticized for using explanatory relevance because this concept itself, it is argued, is in need of explication. I disagree that we need an explication for the task at hand.[5] There are a host of different 'thick' relations in nature we label 'causal' (like pushing, feeding, lapping up, mailing,...). So too there are a host of relations that we lump together under the label 'explains' when explanation serves as a guide to 'bears on the truth of'. The fact that we cannot give an interesting non-circular explication of 'explains' as an objective relation does not mean that we cannot recognize it when we see it – Newton's laws explain Kepler's and my taking an aspirin explains my headache getting better. Nor does it mean that we cannot take certain claims to be generally true of it.

There is good reason why the Achinstein slogan should work for EBPP. To start with, a correct explanation is always evidentially relevant to its explanandum and vice versa. The first follows trivially if one adopts a deductive nomological account of explanation since the explanans cannot hold without the explanandum doing so as well. But, even if one follows GEM Anscombe (1993) in maintaining that an explanans can be enough – it can be as full an explanation as nature allows – without the explanandum obtaining, nevertheless the occurrence of the explanans is undoubtedly evidentially relevant to the occurrence of the explanandum. The converse is trivial since 'explanation' is meant to be 'correct explanation'.

Indirect evidence is harder. E is (indirectly) relevant to H if there is a correct explanation for H that shares a common element, *X*, with some correct explanation for E. $X + X_u^E$ correctly explains E and $X + X_u^H$ correctly explains H.[6] E is evidence that X obtains. But obtaining X cannot be part of a correct explanation for H unless X_u^H obtains. If X_u^H is not the case, then X and X_u^H *cannot* be a correct explanation for H – it doesn't matter how well-confirmed X is. The relevance of E's truth to the truth of H flows through X and it can only do that given X_u^H. E's truth is of no matter at all to H's where X_u^H fails.

[4] For Anchinstein's views on why purely probabilistic characterizations of evidence do not work, see inter alia Achinstein 2004, 1996, 1981.
[5] For Anchinstein's views on this issue see especially Anchinstein 1981.
[6] Subscript 'u' marks the unshared elements of the explanations.

Suppose your interest is in whether H is true. But you know that X_u^H is false.[7] Would you pay to learn E? No. Or take a stock philosopher's case: You are asked to predict the color of a bird in the river. Is the bevy of observed white swans relevant? It is if 'All swans are white' is part of the explanation of both your bird's color and theirs. But if you are told that your bird is certainly not a swan, all those observations of swan color are worthless to you.

So: When the topic is evidence for policy predictions, the relevant concept of relevance is a conditional one: The relevance of a fact E that would have a shared explanatory element with H were H to be true is conditional on the obtaining of the unshared portion of the explanation H would have. Moreover, the epistemic probability awarded to E being relevant should be no higher than the epistemic probability that appropriate unshared factors obtain.

3. External validity and the need for horizontal search

An ideal RCT is a study in which the population in the study, φ, divides into two groups that are identical with respect to all features casually relevant to the targeted outcome, O, except for the policy treatment, T, and its downstream consequences. Suppose the probability of O is greater in the T group than the –T group. Where can we go from there?

Under the probabilistic theory of causality, the values of a full set of O's causes fix the objective chance that O takes any value in its range. That's what prompts the attention to the conditional probabilities from the causal structure for φ, Prob(O/K&T) > P(O/K& –T), where K is an assignment of values to all the members of $CS_\varphi(O)$ with the exception of T and its downstream effects. Whether T has a positive effect size in φ depends on the relative weights in φ of subpopulations in which T acts positively and those in which it acts negatively.

A study is said to be externally valid when 'the conclusion established in the study holds elsewhere'. Consider an ideal RCT for T,O on a large study population φ that has a positive result:

Study Conclusion (*SC:*) Prob(O/T) > Prob (O/–T) in φ.

The study has external validity for target population θ if

Target Conclusion (*TC*): Prob(O/T) > Prob (O/–T) in θ.

(Recall, θ describes the target population supposing the implementation that would in fact occur given the policy in question.)

When is SC evidence for TC?

Since neither SC explains TC nor the reverse, if SC is to be evidence for TC there must be some shared part in their separate explanations. The explanation for

[7] I suppose here that X would not figure in any other correct explanation for H were H to obtain.

the successful RCT results in φ under the probabilistic theory of causality look like this for some specific causal structure, $\mathcal{CS}(O)$, and some specific set of causally homogeneous subpopulations from $\mathcal{CS}(O)$, $\mathcal{K}$ = {...,K_j,... },

> Study Conclusion Explanation (SCE):
> SCE1: The causal structure for O of φ is $\mathcal{CS}(O)$.
> SCE2: For K_J in $\mathcal{K}$ Prob(O/K_J & T) > Prob(O/K_J & –T) according to $\mathcal{CS}(O)$.
> SCE3: The possible negative effects of T on O in other subpopulations are not enough to outweigh this increase.[8]

The explanations for the predicted hypothesis TC are the same in form and must refer to the very same causal structure and the very same causally homogeneous subpopulations if there are to be shared factors:

> Target Conclusion Explanation (TCE):
> TCE1: The causal structure for O of θ is $\mathcal{CS}(O)$.
> TCE2: Some member(s), K_J,$K_{J'}$,... of $\mathcal{K}$ are subpopulation of θ.
> TCE3: For these K_J, Prob(O/K_J & T) > Prob(O/K_J & –T) according to $\mathcal{CS}(O)$.
> TCE4: The possible negative effects of T on O in other subpopulations of θ are not enough to outweigh the increase due to these.

Since most of the claims in both explanations are indexed to the population, the only shared element is the claim that $\mathcal{CS}(O)$ implies that Prob(O/K_J & T) > Prob(O/K_J & –T) for the K_J of TCE3. It is this– and only this– one shared explanatory element that makes the RCT result relevant to the policy prediction. But it is shared only supposing that TCE is a correct explanation for the prediction about θ. That is, the RCT is explanatorily relevant, and thus evidentially relevant, only *relative to* the truth of TCE1,2,&4. What then should be required for the RCT to be accepted as evidence? My dictum: It's not evidence for us unless we have evidence that it's evidence. That means having evidence for TCE1,2,&4. And what reasons do we have to accept these?

To start, what supports TC1 – that φ has the same causal structure for O as φ? Common causal structures are not all that typical. The refurbished Cuisinart Classic 4-slice toaster that I almost bought for £41.46 has a different causal structure than does the Dualit 3-slice stainless steel toaster at £158.03, which has a different structure again from the new Krups expert black and stainless steel toaster at £44.99. Perhaps you think – as many economists and medical RCT advocates seem to – that your two populations are more likely to share causal structure than are the toasters on offer in Oxford. That's fine. But for EBPP you should have good evidence-backed reason for that.

Deleted:

Supposing that the two populations do have a common causal structure, what assures that some of the very subpopulations K_J of φ in which Prob(O/K_J & T) > Prob(O/K_J & –T) are subpopulations of θ? The mix of causal factors that obtain shifts all the time, both across situations and across time. Worse, no matter

[8] One can express this more formally, but that seems needlessly complicated for our purposes.

what mix was there before, in implementing policy we all too often alter that mix. Consider the California class-size reduction program. Reduced class sizes did not improve educational outcomes because the program was rolled out over a short time; the need for teachers doubled within a year but the availability of trained teachers did not. Teaching quality went down, offsetting the good influence of class size.[9]

Finally, why suppose that were T to increase the probability of O in θ as predicted, that would be due to the positive effects in the shared subpopulations rather than in some subpopulations of θ not shared with φ?

These questions need answers, and for EBPP, answers reasonably underpinned by empirical and theoretical support. One cannot just plop SC on the table and *say* that it is relevant to TC. Whether it is relevant depends on common explanatory factors, and presuming that common factors obtain requires good evidence. 'It can't count as evidence unless there's evidence that it's evidence'.

Clearly this dictum can create a regress. That, however, is the human condition. We have to stop somewhere. But it should be somewhere reasonable and defensible. Consider CCTV cameras.[10] Are they working? A glance at the monitor is generally enough to be reasonably certain, despite the fact that in heist movies elaborate techniques are undertaken to make the monitors lie. For relevance, too, we need reasonable and defensible stopping points for the chain of evidence that shows that evidence offerings are evidence. Consider for a moment not the relevance but the credibility of evidence offerings. Where detailed scientific argument and experiment are involved, this is going to be hard for policy analysts and practitioners to judge. That is why institutions like the Cochrane and Campbell Collaborations or the What Works Clearing House have been set up. If they give a study result high marks, it is generally reasonable for a practitioner to take that on faith.[11]

What then of the evidence for the relevance of SC for TC? Sometimes we can assemble some body of facts that are reasonably well attested and that provide good reasons in favor of claims like TCE1,2,&4. But it is hard. And the very cases in which one most wants to perform an RCT are the cases where there will be least evidence that a positive RCT result for the policy treatment is evidence that the policy will work for us. RCTs are touted as gold standard because only they 'control for unknowns', for the factors in the causal structure for O that we don't know are there and hence can't check explicitly are distributed the same in the two groups.

---

[9] Elsewhere I describe this case in terms of capacities. The same kinds of problems arise in both cases.

[10] See Pawson and Tilley 1997 for a good use of the example of CCTV cameras in parking lots discouraging car theft to argue the need for what I here call 'horizontal search' and to show how understanding the mechanism at work can help with that.

[11] I think, however, that negative judgements by these organizations are often made on bad premises. They tend to presume that trusting to pure method is always better than supposing substantive knowledge claims. That, for example, is why RCTs are gold standard and econometric modelling doesn't get a look in. See Cartwright 2007 for more details.

So RCTs come into their own when we suspect that a good many factors in the causal structure for the study population are unknown. But then how are we supposed to produce evidence that those very unknown factors are causal factors according to the causal structure for θ? And that θ has some of the same causally homogeneous subpopulations in which T is positive for O as does φ? Finally, how do we estimate that in other subpopulations of θ, T won't have enough negative effects to decrease the chance of O there? The very same epistemic gaps that make the RCT the method of choice also make results practically useless for prediction.

The problems discussed in this section demand *horizontal* search. T can increase the probability of O in some mixes of causal factors and not in others; it can even decrease the probability in some while increasing it in others. A positive RCT result is relevant to a policy prediction only relative to assumptions about the mixes of factors operating in the study population and in the target population. To be justified in taking the RCT as evidence we need to gather information about what other factors operate with T in the two populations. That's what I mean by a 'horizontal search'. To increase the range of relevance of the RCT we also need a 'vertical search', which reviews causes across levels of abstraction.

4. External validity and vertical search

The causes in a causal structure can be more or less abstract; and structures involving factors at different levels of abstraction can all obtain at once. "The trajectories of bodies moving on a sphere subject only to inertia are great circles" is true; so too is "The trajectories of bodies moving on a sphere subject only to inertia are geodesics (i.e. the shortest distance between two points)". They are equally true because on a sphere, a great circle *is* a geodesic.[12] Generally the higher the level of abstraction of a causal structure, the more widely it is shared across populations. For example, bodies on Euclidean planes subject only to inertia follow geodesics but not great circles. This matters for explanatory relevance.

An easy way to get a grip on how it matters is to consider some examples. The first is from climate-change modeling, where development economists argue that many of the policies that can help alleviate harmful effects of climate change are things that should be done in developing countries anyway. This is the case of the Bangladesh Integrated Nutrition program (BINP) for providing pregnant women with nutritional counseling, with the idea that poor nutrition is not only due to poverty but also to ignorance, for instance to belief in 'eating down' during pregnancy. (White 2009) Of course knowledge by itself is not enough, resources are required too, so the counseling was joined by a

[12] I shall here be relatively cavalier about the metaphysics of properties. I treat abstract features and concrete ones both as real and I treat them as different features even if having one of these (the more concrete feature) is what constitutes having the more abstract one on any occasion. I take it that claims like this can be rendered. appropriately, though probably differently, in different metaphysical accounts of properties.

supplementary feeding program. This is the kind of factor that comes up in a horizontal search.

An analysis by the World Bank's Operations Evaluation Department found no significant impact on infants' nutritional status. This despite the fact that the program had 'worked' elsewhere. What went wrong?

A number of reasons suggest that the results elsewhere were not evidentially relevant to the success of the policy in Bangladesh. They might have been. It is natural to expect that explanations for the results elsewhere and for Bangladesh success would share an important common element: A general principle

> Principle 1: Better nutritional knowledge in mothers plus supplemental feeding improves the nutritional status of their children.

In fact the two populations did not share this principle.

The first reason for the lack of impact, it seems, is that there was 'leakage': In Bangladesh the food was often not used as a supplement but as a substitute, with the usual food allocation for that child passing to another member of the family. (Save the Children 2003) The principle 'Better nutritional knowledge in mothers plus supplemental feeding improves children's nutrition' was true in the original successful cases but not in Bangladesh. A better candidate for a shared explanatory element is

> Principle 2: Better nutritional knowledge in mothers with *sufficient resources* to use that knowledge improves children's nutrition.

This principle uses concepts at a *higher level of abstraction*. In the successful cases the more concrete description 'food supplied by the supplementary feeding program' counted as an instance of the more abstract concept 'sufficient resources', but not in Bangladesh. Not getting this straight is a failure of *vertical search*: A failure to identify the right level of abstraction to find common explanatory elements.

A second reason for the lack of positive impact is also a problem with vertical search.

> The program targeted the mothers of young children. But mothers are frequently not the decision makers, and rarely the sole decision makers, with respect to the health and nutrition of their children. For a start, women do not go to market in rural Bangladesh; it is men who do the shopping. And for women in joint households – meaning they live with their mother-in-law – as a sizeable minority do, then the mother-in-law heads the women's domain. Indeed, project participation rates are significantly lower for women living with their mother-in-law in more conservative parts of the country. (White 2009, 6)

This suggests yet another vertical move to secure a shared principle:

Principle 3: Better nutritional knowledge results in better nutrition for a child in those who
1. Have sufficient resources to use that knowledge to improve the child's nutrition,
2. Control what food is procured with those resources,
3. Control how food gets dispensed, and
4. Hold the child's interests as central in performing 2. and 3.

Just as supplementary food did not count as sufficient resources in the BINP, mothers in that program did not in general satisfy the more abstract descriptions in 2. and 3.

The previous successes of the program are relevant to predictions about the BINP only *relative to* the vertical identification of mothers with the abstract descriptions in 2., 3., and 4. But not all of these identifications hold. So the previous successes are not evidentially relevant. For an RCT to be relevant, and to be justifiably taken as such, we need good reasons to back up the claims that the characteristics referred to in study conclusions, which are often fairly concrete, are the same as the characteristics appearing in principles shared across study and target populations, which are often relatively abstract.

Consider another possible example, this from UK child-welfare policy. In many cases a child's care-givers, though not legally compelled, are heavily encouraged, perhaps even badgered, into attending parenting classes. Consider in this context making fathers attend parenting classes.

First, is 'father' to be instantiated by 'biological father' or, e.g. 'male partner of the mother who lives in the household with the child', or maybe 'male care-giver'? It may well be that the policy would be effective if the male care-givers or men living with the mother are the target but not biological fathers who are neither on site nor care-givers. If so, to focus on 'being a father' would be to move to too high a level of abstraction since only the more specific feature, 'male care-giver' or 'male partner of mother who shares the child's household', enters into a reasonably reliable principle.

On the other hand 'compelling father' or 'compelling male care-giver' can simultaneously be too concrete. Different cultures in the UK have widely different views about the roles fathers should play in parenting. Compelling fathers to attend classes can fall under the more abstract description, 'ensuring care-givers are better informed about ways to help the child', in which case it could be expected to be positively effective for improving the child's welfare. But it may also instantiate the more abstract feature 'public humiliation', in which case it could act oppositely. And of course it can fall under both at once. In any case, if the two more abstract features pull in opposite directions, there will be no reliable principle to formulate at the more concrete level involving 'fathers'. Nor is this pull in opposite directions an unrealistic hypothesis. We know from empirical research that there are varying outcomes associated with compelling/strongly encouraging parents to attend parenting classes and also that these are correlated with varying motivations. (Barlow et al. 2006) Unfortunately we do not yet have sufficient theoretical probing to explain the variation and the correlations.

5. Troubles for vertical search

To secure explanatory relevance in cases like the BINP, it is necessary first to find and defend a shared explanatory principle. This involves finding the right ladder of abstraction to climb and knowing just when to stop.[13] But a principle can only be shared between study and target if it applies to both. So it is equally necessary to defend that what happens in the study and what is predicted to happen in the target instantiate the abstract concepts in the putatively shared principle.

This is no easy matter since what in the concrete an abstract property consists in often differs dramatically from circumstance to circumstance. This problem arises regularly in economic climate mitigation and adaptation models (and many other economic models as well). Consider studies of how to change American insurance schemes to provide financial incentives for those living in high risk areas, like the chic Florida coast, to make their homes less prone to risk, for instance by changing the roof construction. (Cf., Kunreuther and Michel-Kerjan 2009 plus references therein.) The models often rely on game theory assumptions that rational agents act to maximize their expected utility.. Here we have to worry about misplaced concretization of the abstract feature 'utility'. The models typically take money to instantiate utility. But there is a good chance that the targeted agents – say rich owners of beach-front residences – will be more moved by the disruption to their domestic arrangements of having builders at work for months than by any contrary financial incentive that could realistically get built into an insurance scheme.

The same problem of context dependence resurfaces when it comes to measurement, where we see a familiar trade-off: Shared principles require higher levels of abstraction; good measurement, lower. For good comparable measurements, we want specific operational procedures that are carried out in the same way each time the measurement is performed. By contract, the methods for measuring an abstract feature generally differ depending on what more concrete features it consists in, which is not the same from case to case.

We are pulled in two directions here. One: Plump for a false universal concretization in order to secure a universal measure. For instance, measure 'educational value added' in new British inner-city academies by counting the number of GCSE's passed at a grade of C or better. Or, devise a measurement definition that more correctly captures the abstract feature of interest across its various concrete instantiations. The danger then is that the definition will be so

[13] Stopping matters. Increased abstraction generally goes along with increased generality. So the more abstract the principles you embrace, the more so-far-unexplored concrete predictions you are committed to. My own advice has always been: Don't commit to anything more than you need. That is why I have always urged sticking to the numerous more concrete, detailed laws that explain – and explain in proper detail – the various natural and experimental results we observe rather than committing to the super-abstract laws of high theory.

abstract that we don't know what it consists in from situation to situation. For example, what constitutes human flourishing differs dramatically according to individual circumstances and abilities, natural resources, availability of public goods, need, and the like. So the capability approach of Amartya Sen (1985, 1999) proposes as a measure 'the number of lives worth living open to the individual'. Or, some propose to measure the economic freedom individuals enjoy by the size of their choice sets. Neither of these provides much of a clue about what we are actually to do to assign numbers or ranks to the individuals to be measured.

For EBPP we look to science for advice. Unfortunately when it comes to fixing what constitutes abstract features in the concrete, science offers at best rules of thumb that are highly defeasible. In particular they are beset by what John Perry (2010) dubs 'the failure of enrichment': That A consists in M in circumstances C does not imply that A consists in M in circumstances C & C' for every C' consistent with C.

The moral particularism literature is rife with examples where A is a moral feature.
Stuart Hampshire (2000), for instance, describes telling stories to philosophical audiences. The stories involve a young intellectual French Fascist, a reader of Celine, held by the Free French, whom Hampshire is sent by the British to interrogate. The French will execute the young man; but they tell Hampshire that he can certainly promise the prisoner – falsely –that he will not be executed in exchange for information. Is it acceptable, or even required, for Hampshire to lie to the young man? Hampshire tells the story differently on different occasions. Often the descriptions can be nested, the more detailed descriptions containing the previous, plus more. Depending on how Hampshire tells the story, the audience is in general agreement about what he should do – but the verdict changes as he shifts from level to level. Enrichment fails.

Hampshire's stories involve highly abstract features – *morally acceptable*, *morally required*. Perry's own example involves specific motions that may or may not instantiate his eating a Brussel sprout at his Dewey lecture, depending on the level of detail of the description of the circumstances. So the abstract feature need not be very abstract at all for the failure of enrichment to appear.

Where then can we find help in science either with the problem of settling on the right level of abstraction to find shared explanatory principles or of ascertaining what the abstract features in these principles consist in for both study and target populations? I don't know an answer. But I am sure it takes both theory and local knowledge, neither of which are much in favor in EBPP communities. Without these, scientific studies like RCTs, which are so highly prized for the credibility they confer on their results, will not be explanatorily relevant to the predictions about what will work for us that we need for practice and policy. And I am sure Achinstein is right for these kinds of cases: If explanatory relevance goes, so too goes evidential relevance. Then we have no scientific evidence to bring to bear and evidence-based policy and practice is out the window.

References

Achinstein, Peter. 2001. *The Book of Evidence*. New York: Oxford University Press.

———. 1981. "Can There Be a Model of Explanation?" *Theory and Decision* 13: 201-27.

———. 2004. "A Challenge to Positive Relevance Theorists: Reply to Roush." *Philosophy of Science* 71: 521-24.

———. 1978. "Concepts of Evidence." *MIND* LXXXVII: 22-45.

———. 1983. *The Nature of Explanation*. New York: Oxford University Press.

———. 1981. "On Evidence: A Reply to Bar-Hillel and Margalit." *MIND* XC: 108-12.

———. 1996. "Swimming in Evidence: A Reply to Maher." *Philosophy of Science* 63: 175-82.

Anscombe, Gertrude Elizabeth Margaret. 1993. "Causality and Determination." In *Causation*, edited by Ernest Sosa and Michael Tooley, 88-104. Oxford: Oxford University Press.

Barlow, Jane, Isabelle Johnson, Denise Kendrick, Leon Polnay, and Sarah Steward-Brown. 2006. "Systematic Review of the Effectiveness of Parenting Programmes in Treating Abusive Parenting." *Cochrane Database of Systematic Review* 3: 1-20.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

———. forthcoming. "The Long Road from RCTs to Effectiveness." *The Lancet*.

———. 2009. "What Is This Thing Called Efficacy?" In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, edited by Chrysostomos Mantzavinos, 185-207. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Eileen Munro. 2010. "The Limitations of Randomized Controlled Trials in Predicting Effectiveness." *Journal of Evaluation in Clinical Practice* 16: 260-66.

Dekkers, Olaf, Erik Von-Elm, Ale Algra, Johannes Romijn, and Jan Vandenbroucke. 2010. "How to Assess the External Validity of Therapeutic Trials: A Conceptual Approach." *International Journal of Epidemiology* 39: 89-94.

Embry, Dennis, and Anthony Biglan. 2008. "Evidence-Based Kernels: Fundamental Units of Behavioral Influence." *Clinical Child and Family Psychology Review* 11: 75–113.

Hampshire, Stuart. 2000. *Justice Is Conflict*. Princeton: Princeton University.

Kunreuther, Howard, and Erwann Michel-Kerjan. 2009. *At War with the Weather: Managing Large-Scale Risks in a New Era of Catastrophes*. New York: MIT Press.

Mackie, John Leslie. 1980. *Cement of the Universe*. Oxford: Oxford University.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. London: Sage.
Perry, John. 2010. "Dewey Lecture: Wretched Subterfuge." American Philosophical Association: Pacific Division.
Save the Children. 2003. *Thin on Ground*. London: Save the Children.
Sen, Amartya. 1985. *Commodities and Capabilities*. Oxford: Oxford University.
———. 1999. *Development as Freedom*. New York: Knopf.
White, Howard. 2009. *Theory-Based Impact Evaluation: Principles and Practice*. New Delhi: The International Initiative for Impact Evaluation (3ie) Working Paper 3.