

A Theory of Evidence for Evidence-Based Policy
Nancy Cartwright (with Jacob Stegenga)

- I. The preliminaries
 - I.1. The project
 - I.2. How to think about the problem
 - I.2.a. Viewpoint
 - I.2.b. Effectiveness
 - I.2.c. A structure for the problem
- II. Evaluating effectiveness
 - II.1. How philosophy can help
 - II.2. Causes and counterfactuals
 - II.3. Causal models
 - II.3.a. What's a causal model?
 - II.3.b. Had we world enough and time
 - II.4. INUS conditions
 - II.4.a. Philosophers' talk
 - II.4.b. Epidemiologists' talk
 - II.4.c. Four examples
 - II.5. Two central principles for a theory of use
- III. The neglected questions
- IV. Making life somewhat easier
- V. Mechanisms: A principle in aid of practical advice
 - V.1. Tracing the causal process: an example from economics
 - V.2. Identifying the means of production: a criminology example
- VI. In Sum

Part I: The preliminaries

I.1. The project

I aim here to outline a theory of evidence for use; more specifically, to lay foundations for a guide for the use of evidence in predicting policy effectiveness *in situ*, a more comprehensive guide than current standard offerings, such as the Maryland rules in criminology, the weight of evidence scheme of the International Agency for Research on Cancer (IARC), or the 'What Works Clearinghouse'. The guide itself is meant to be well-grounded but at the same time to give practicable advice, that is, advice that can be used by policy-makers not expert in the natural and

social sciences, assuming they are well-intentioned and have a reasonable but limited amount of time and resources available for searching out evidence and deliberating.

I go into the project with some assumptions. The first is a delimitation of the topic. The guide for which I aim to lay a theoretical base is to be concerned with the use of evidence to estimate, if only roughly, whether if a proposed policy were implemented, as it would in fact be implemented, a specific, identified outcome would be produced.

The second is that the project needs to be approached from the point of view of the evidence user, not the evidence producer.

Third, I assume that rigor is a good thing, so that the advice should be firmly rooted in sound principles; but we must not be pseudo-rationalistic. A rigorous argument with 9 well-grounded premises and one weak one does not make for a rigorously established conclusion. For the most part, estimates of whether a policy will be successful made by real people in real time will be both rough and uncertain. That is important to keep in mind as policy decisions are made. But it is also important to keep it in mind as advice guides are devised. If advice is to be practicable, it may well not be hugely reliable, even if it is ultimately well-grounded. We should aim for advice that improves decisions even if we cannot do the job perfectly. The best should not be the enemy of the good.

Fourth, and closely connected with the third, is that we should not expect policy effectiveness judgments to be very reliable. There are a variety of different reasons conspiring to make these judgments especially difficult, including the obvious difficulties of doing what I propose here as necessary for reasonably reliable judgments. I shall not rehearse these reasons but just offer one remark to make vivid how difficult the task is. Asking if a policy of a specific design will achieve a targeted result is structurally just like asking whether a laser of a specific design will produce a coherent beam when we plug it in. We know how difficult it is to answer that question reliably before actually plugging it in – and how complicated it would be to produce advice about what counts as evidence for or against a yes answer and about how to marshal that evidence to settle on a prediction. Social effectiveness will be even harder since the systems under study are more open, our theories and knowledge of the materials are less secure, and the choice of targeted outcomes is generally dictated by social need, not by an assessment of how achievable they are.

1.2. How to think about the problem

1.2.a. Viewpoint

When it comes to evidence-based policy, viewpoint matters. Whether wittingly or not, typical advice guides focus on the *production side* of scientific evidence and not on the *use side*. They tell us what counts as good science, not how to use that science to arrive at good policy.

Most available guides, like the Maryland rules, the IARC scheme, and What Works, provide ranking schemes for the ‘quality’ of evidence. These schemes police the credibility of results that can be counted as evidence. Evidence claims are ranked according to the methods by which they are tested. High quality means that the tests are stringent: Results that pass the tests are very likely to be true. RCTs are necessary for strong evidence according to the dominant guides. Many object on the grounds that this can mean throwing out a lot of good evidence that we ought to be attending to. This issue is not my concern here. The central concern I raise here is that these rankings focus on too narrow a range of *claims that need evidencing*, not that the kinds of evidence admitted are too narrow. Why?

Truth is a good thing. But it doesn’t take one very far. Suppose we have at our disposal the entire encyclopedia of unified science containing all the true claims there are. Which facts from the encyclopedia do we bring to the table for policy deliberation? Among all the true facts, we want on the table as evidence only those that are *relevant* to the policy. And given a collection of relevant true facts, we want to know how to assess whether the policy will be effective in light of them. How are we supposed to make these decisions? That is the problem from the *user’s* point of view, and that is the problem of focus here.

1.2.b. Effectiveness

There are a great many things we need to evaluate in considering whether to adopt a policy or not. Will the policy work? Does it have unpleasant side effects? Does it have beneficial side effects? How much does it cost? Have we made the correct choice of target outcomes? Is the policy morally, politically and culturally acceptable? Can we get the necessary agreement to get it enacted? Do we have the resources to implement it? Will enemies of the project sabotage it in various ways?

Every one of these questions needs answering and in each case evidence will help get the right answer. I shall confine my discussion, however, to the *question of effectiveness*:

Question of Effectiveness. Will the proposed policy produce the targeted outcomes were it to be implemented in the targeted setting in the way it would in fact be implemented?¹

I.2.c. A structure for the problem

Start then from the point of view of the policy deliberator trying to estimate whether a proposed policy will be effective. **For a reliable decision one wants credible evidence that, all told, speaks for (or against) the policy.** This simple observation suggests that from the point of view of the user three different issues need addressing:

1. *Quality*: When are evidence claims credible?
2. *Relevance*: When does an established result bear on a policy prediction and how does it do so?
3. *Evaluation*: How should predictions about policy effectiveness be evaluated in the light of all the evidence?

The first is an issue about the production of knowledge by the social and natural sciences; it is the meat of evidence-ranking systems. The latter two are the more neglected questions I focus on.

The fact that the three questions are distinct should not suggest that their answers are unrelated. Despite the common emphasis on question 1, it seems *prima facie* as if the natural starting point is with question 2. First establish what kinds of evidence are relevant to effectiveness. Then, for question 1, provide guidelines that police the quality of evidence of those kinds; and for question 3, propose some scheme for amalgamating or combining evidence.

In aid of this approach one could adopt one or another of the characterizations of relevance on offer from philosophy and methodology of science, where the topic has been explored and debated for years; then follow on with one or another of the schemes available for combining evidence or adapt weighing schemes with known characteristics from

¹ Of course there will seldom be a highly certain yes or no answer. So at some point an assessment of the probabilities will have to be made in light of the evidence, even if only roughly. But reasonable probability assessments depend first on understanding the structure of the problem, which is the topic to be tackled first.

other areas, like those for amalgamating preferences or expert testimony. This is one approach that we are looking at in this workshop.

I adopt a different strategy. I propose to start with an account of how to evaluate claims of effectiveness and work backwards from there to figure out what kinds of evidence would be relevant for the evaluation, finally returning to the first issue of how to assure that the kinds of evidence claims needed are sufficiently credible to enter into deliberation.

Before beginning with this account, I want to stress the importance for the success of evidence-based policy of covering all three questions. Question 1 is a question for knowledge producers: What is necessary in order to ensure that a claim entered as evidence is likely to be true? Users have in addition to face questions 2. and 3.² Yet most of the rigor and most of the attention is to question 1. We are urged to extreme rigor at one stage, then left to wing it for the rest.

But: a chain of defense for the effectiveness of a policy, like a towing chain, is only as strong as its weakest link. So the investment in rigor for one link while the others are left to chance is apt to be a waste. To build the entire chain one may have to ignore some issues or make heroic assumptions about them. But that should dramatically weaken the degree of confidence in the final assessment. Rigor isn't contagious from link to link. If you want a reasonably secure conclusion coming out, you'd better be careful that each premise is secure enough going in.

Part II: Evaluating effectiveness

II.1. How philosophy can help

I propose to borrow the three central principles of the theory of evidence for use from philosophy. The first two provide the basis of the theory and the third, some practical help in implementing it.

- Truth values for counterfactuals are fixed by causal models.

² Is relevance really, as I say, a question for the user rather than the knowledge producer? Many think not. Indeed it is a common criticism of studies in the social sciences that they do not say what they show, what the results bear on, at a practical level. I don't think they can. Perhaps they can do better, but there will always be a great number of relevance judgements that must be left to the user. Whether a given fact is relevant as evidence for a given claim depends on a host of other assumptions, both theoretical and local to the situation. (This is the lesson of the famous 'Duhem-Quine' problem in philosophy of science.) For causal counterfactuals of the kind we assess in effectiveness evaluations, relevance will depend in addition on *how* the cause is supposed to produce the effect. (See *Part V* here.)

- Causes, as JL Mackie explains, are INUS conditions.
- In understanding how causes operate together, mechanisms matter.

II.2. Causes and counterfactuals

For sound policy we need to evaluate whether if the proposed policy were implemented as it would in fact be implemented, the targeted outcome would occur in consequence. We are looking for the probability of what in decision theory is called *a causal counterfactual*.

There is good reason to expect an intimate connection between causes and these special kinds of counterfactuals. Nature forges it. Consider: How does nature decide what effects to produce in a particular situation? First she surveys the causes that will be operating. Next she consults her rules of combination to calculate what should happen when they all act at once. Then she produces the prescribed effects. We can't lose by imitating nature.

That is my proposal. To predict what will result if we introduce some new policy or program, we should follow Nature's lead. We should reconstruct Nature's list of causes and mimic Nature's calculation. This provides us with a surefire way to predict the effects of our policy implementations.³

II.3. Causal models

I propose then to adopt standard philosophic advice as the first principle of the theory of use: To evaluate causal counterfactuals, build a causal model. But the term 'causal model' should not carry a lot of baggage with it, either from philosophy or from the sciences, where various different kinds of specialized causal models are on offer.

II.3.a. What's a causal model?

For our purposes a **causal model** has two essential ingredients, where I separate the first into two parts to highlight issues about implementation that we know policy makers need to take into consideration.

³ Later (Part IV) we can consider 'cheap heuristics' that might get the same conclusion enough of the time.

1. A list of the causes relevant to the targeted effect that will operate in the target situation. This includes
 - 1.a. the causes present in the situation independent of the policy action
 - 1.b. any changes in this set of causes introduced in implementing the policy.
2. A rule of combination that calculates what should happen vis-à-vis the targeted effect when those causes operate together.

Consider a simple case. Later we shall look at both some real and some pastiche social policy cases. But for now I illustrate using everyday physics knowledge. I do so because the reasoning is simple, well-understood, and I am not likely to get involved in subject-specific debates in education or criminology or health policy. More importantly, I choose this kind of case to start out with because it is one where our knowledge of the principles and of the aptness of the concepts is secure, so that we can focus on the structure of the reasoning needed.

The case of the desk magnet versus the industrial magnet. I have access to a desk magnet, alternatively to a large industrial magnet. I know the exact strengths of these with a very high degree of certainty – claims about their efficacy for lifting objects have passed far more than two good RCTs; they have centuries of study behind them. Shall I use one of them to lift an object in my driveway? That depends on the other features of the target situation.

First, magnets need helping factors to be effective at all. My desk magnet is useless for lifting a matchstick; it is only the *combination* of a magnet and a metal object that produces a magnetic force. Then the acceleration caused by the magnet is still only one part of the story, often one very small part. To know what happens when we apply the magnet we need to know the other forces as well. Here, especially gravity. The desk magnet may lift a pin but it is hopeless for my car, where we need the industrial magnet. We also need to tend to what other forces we introduce in the course of getting the magnet in place. Perhaps the industrial magnet would have lifted the car if only we hadn't thrown the heavy packing case for the magnet into the trunk.

Finally, we need to know how all these factors combine to produce a result. Often in social contexts we assume additivity: add a good thing and the results can only get better. But that doesn't work in even this simple physical case. We get so used to vector addition

that we forget that it isn't simple addition of effect sizes. Add a magnetic acceleration of 42 ft/sec/sec to that of gravity's 32 ft/sec/sec and you won't usually get 74.

The point is that whether the magnet will be effective at all in the target situation and to what extent depends on nature's causal model of the situation. So the most direct way of predicting its effects is to construct our own causal model in imitation of nature.

I know no-one wants to hear this since it seems difficult. But consider: We know industrial magnets would pass any number of RCTs, of any degree of stringency. But that's not anywhere near enough to know. None of us would rent an industrial magnet to remove a load of rubbish without looking at the rubbish. Knowledge that magnets just like this *can* lift is only a small part of what we consider when we evaluate whether renting the industrial magnet will be effective in removing our rubbish. If this is so in everyday calculations and in applied science and engineering, why should we expect it to be substantially different – and substantially easier – in social engineering?

Of course constructing causal models is hard, even if the models are rough and we have figured out ways to tolerate uncertainties. Sometimes there are shortcuts, 'cheap heuristics' that get us, more-or-less, well-enough, the same conclusions that the causal model generates. As decision makers, we can opt for a heuristic if we want. But there is no avoiding the fact that the choice of the right heuristic depends on the right causal model. We may not wish to build a causal model; we may not know how to; we may think it takes too much time or money, intelligence or attention. That does not alter the fact that when we buy a policy we are betting on a causal model, willy-nilly, whether we wish to think about it or not.

II.3.b. Had we world enough and time

A great deal more can be said about causal models. But it is subject and discipline specific and almost always requires expertise and training to do at all properly. Moreover, many scientific models do less than what I demand of a causal model, though they provide more detail and zero in, usually very precisely, on specific features of interest.

Consider a joint effort to explore the causes of delays in emergency rooms.⁴ The modeling expertise was provided by the Department of Operational Research at LSE, while orientation to the problem area, judgments on design choices, and introductions to stakeholders were supplied by Casualty Watch, a project organized as a response to public concern that cuts in the NHS were producing an inadequate emergency service and harming patients. System dynamics was selected as the appropriate modeling medium and the model was calibrated with information from an inner London teaching hospital. Here's what the model looks like:

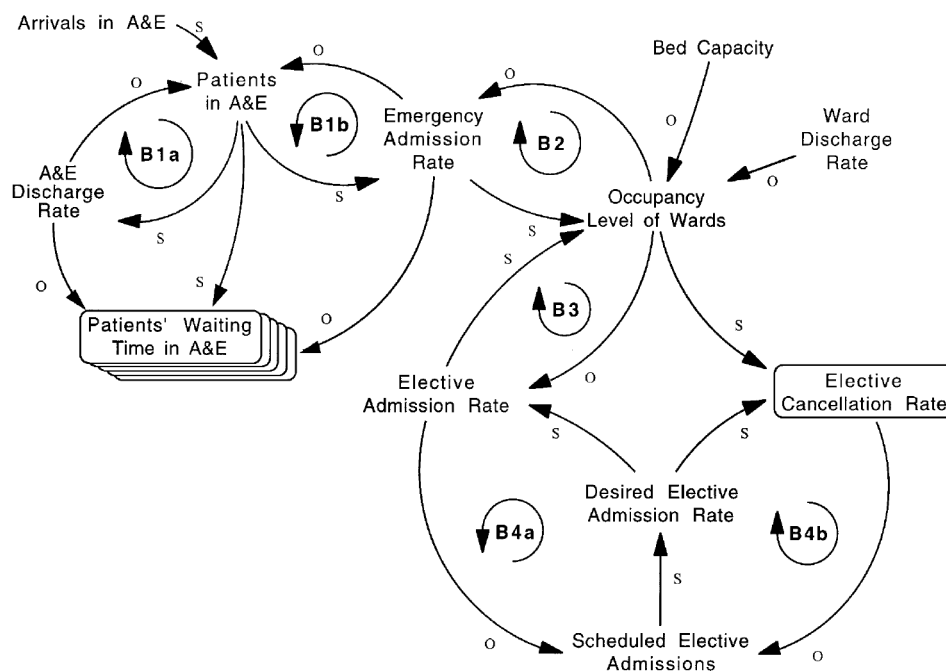


Figure 1. Model of delays in emergency rooms.

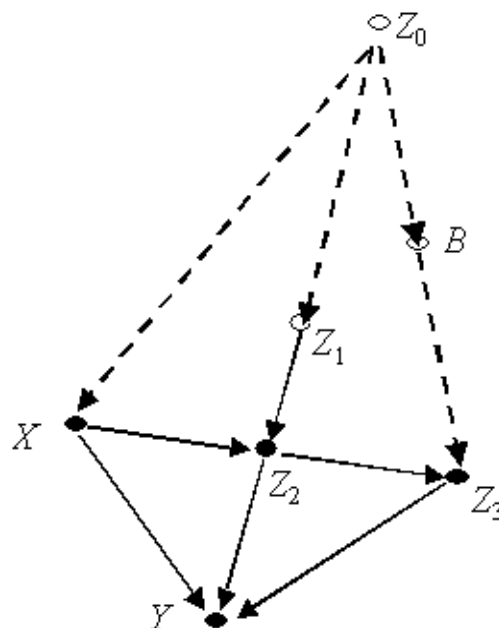
What's important about this model is its ability to detect and represent feedback loops and its dynamic structure. It shows what happens between the initial cause, arrivals at the Accident and Emergency Department, and the final effect, patient waiting time at A&E. As I shall explain in *Part V*, tracing through the dynamics like this, step-by-step, can be a big help in constructing a significant part of the second component I demand in causal model: an account of how causes act together to produce the targeted effect, because it focuses on what auxiliary causes are needed at each step if the salient cause is to produce the next step in the process.

⁴ Ref to David Lane's paper

Notice, however, that this information is not explicitly represented in the model since the model treats causes singly. At the head of the arrow – at the causes end – is a single variable; e.g. bed capacity, ward discharge rate, and emergency admission rate are all pictured as separate causes of the ward occupancy rate. There is no information encoded about how these different causes combine, in particular which causes must act together before they can contribute to the effect at all. Thus this model, like most professional models, does less than I require, though what it does, it does more precisely and in more detail.

Here is another example, this one from Judea Pearl.

Figure 2: A causal Bayes net:



Variables: X : fumigants; Y : yields; B : the population of birds and other predators; Z_0 : last year's eelworm population; Z_1 : eelworm population before treatment; Z_2 : eelworm population after treatment; Z_3 : eelworm population at the end of the season.⁵

In this model, as in the last, causes are at the top of the arrow, effects at the tip. By calling it a causal 'Bayes net' special assumptions are made

⁵ Pearl (1995) 669-70

about the relations among the variables that may not hold in every causal model; for instance causes and effects pictured in the graph are all supposed to be probabilistically dependent. So this kind of model contains more information than is required by my two conditions for a causal model, information peculiar to particular kinds of causal systems. But like the dynamic-systems model for emergency room admissions and hospital beds, it also contains less since the model does not show how the causes interact among themselves in affecting yields.⁶

This kind of missing information is readily supplied by models presented in the form of equations, if they can be constructed. Here for instance is the final equation from a causal model I shall discuss in *Part V*:

$$y_t = \theta\beta[p_t - p_{t-1}] - \theta\beta\pi + y_{pt} \quad \dots \quad (*)$$

Here y_t is output at t and p_t is price at t , so $[p_t - p_{t-1}]$ is a measure of inflation. This equation yields as a next step the classic Philips curve representing a trade-off in which rising inflation causes decreasing unemployment. Once the parameters, θ , β , and π , are filled in the equation shows exactly how the two causes represented – inflation, $[p_t - p_{t-1}]$, and earlier output, y_{pt} , combine to produce later output, y_t : in this case, simple linear addition.

In section II.4.c I will present a simple physics example where a complete set of causes is also laid out in an equation, but the rules of combination for the causes are more complicated, involving not simple addition but also multiplication and vector addition.

Equations for calculating the exact result of a given set of causes are wonderful when you can get them. But they may not be possible, even in principle, for many cases; Nature herself may proceed with less quantitative precision. Whether she does so or not, this level of precision is generally well beyond the ability of normal policy deliberators. Also, as my colleagues at a recent conference on causality urged me to remind you: Our list of causes will almost always be incomplete; the very best we can hope for is a probabilistic assessment of the outcomes and even that

⁶ Many of those developing the theory of causal Bayes nets describe them as a method for ‘causal discovery’. I think that’s right. They are tools on the knowledge production side; a way to sidestep the need for RCTs by establishing efficacy with the same degree of rigor as an RCT but using population, not experimental, data. They may even be of far more immediate relevance to policy than an RCT if the data comes from the very same population as the target population. Still, without further additions, they are not enough to evaluate causal counterfactuals. (Though see Judea Pearl’s beautiful work on how to use them to evaluate the probability of casual counterfactuals, given input probabilities for exogenous factors and given that the special Bayes-nets axioms hold in the system under study.)

should generally not be too precise. So don't get hung up trying to produce equations.

But that is not advice to ignore the need to get a grip on the dominant causes that will be affecting the outcome or the need to bet on what they do in combination. It is just advice not to expect a degree of precision or a degree of confidence that neither the subject nor our capabilities can support.

II.4. INUS conditions

II.4.a. Philosophers' talk

To evaluate a causal counterfactual we need to consider the major causes at work and how they combine. One characteristic of causes widely accepted in philosophy can help with both enterprises. As JL Mackie argued, causes are INUS conditions.⁷⁷ I propose to adopt this at the second basic principle in the theory of use.

An *INUS condition* is an **I**nsufficient but **N**on-redundant part of an **U**nnecessary but **S**ufficient condition.

Let me give several illustrations. The first has been artificially constructed by Charles Ragin to illustrate his own methods for identifying INUS conditions. Consider Ragin's example of a hypothetical study of the causes of defection-related turnover in HMO's using qualitative comparative analysis to isolate INUS conditions.

Social science example. Defection-related turnover in HMO's can be caused by two different factors, each of which is unnecessary but sufficient:

- A change in ownership or management combined with a speed-up of the patient flow
- Management appropriation of the power to veto all referrals to medical specialists combined with the use of outside specialists.

So here we have an effect with four causes, four INUS conditions: change in ownership or management, speed-up

⁷⁷ That is, all causes are INUS conditions. But not necessarily the reverse.

of patient flow, management appropriation of the power to veto all referrals to medical specialists, and use of outside specialists.

I introduce this odd technical term, *INUS conditions*, from philosophy because usually when we discuss policy we focus on a single cause, a single INUS condition. But we won't be able to predict the effect of that cause without considering *all the other INUS conditions and the relations among them*.⁸ Thinking in terms of INUS conditions then serves several purposes:

- It focuses attention on the fact that there are usually a number of distinct causal complexes that contribute independently to the effect.
- It focuses attention on the other factors that are necessary along with the policy variable if the policy is to have any effect at all.
- It focuses attention on the functional form of the relations of the variables within a single causal complex.
- It focuses attention on the overall functional form: How do the separate causal complexes combine? Recall my earlier remark. Often in social contexts we assume additivity. But that doesn't work in even simple physical cases. The vector addition of classical mechanics is after all a long way from the simple linear addition of effect sizes.

All four of these focuses played a role in my tale of the desk magnet and the industrial magnet. So readers may wish to look back to that discussion for illustration.

II.4.b. Epidemiologists' talk

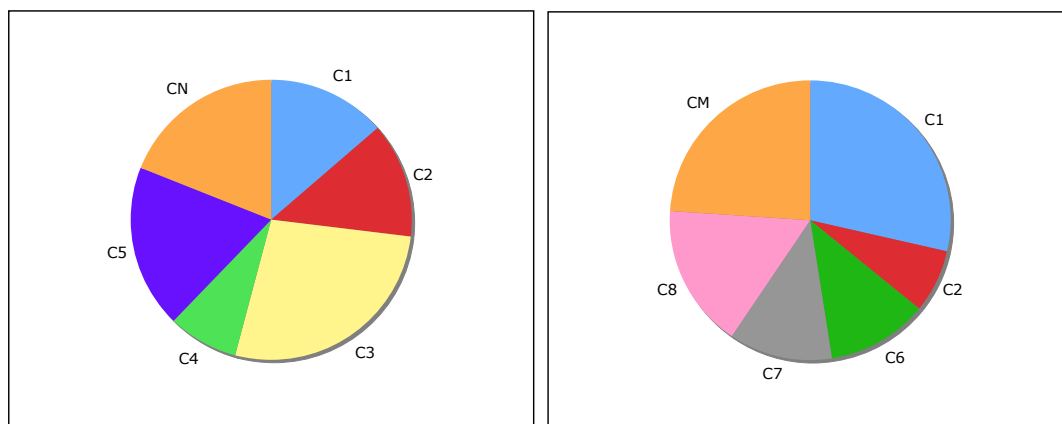
INUS conditions are not just a topic for philosophers. They have been useful in epidemiology for a while now. Looking at how epidemiologists describe and use them may help get a better grip on them. Epidemiologists define a sufficient cause as a constellation of component causes that together is sufficient to cause a disease. They use pie-chart diagrams to represent sufficient and component causes. Each slice in a given pie is a component cause, and a whole pie is a sufficient cause. A pie slice on its own is insufficient to cause disease; the whole pie is

⁸ Sometimes we are only interested in estimating what difference the policy will make and even then sometimes only the direction of change so that we can get by without an estimate of size. For that we clearly need somewhat less information. To be discussed in *Part IV*.

needed. So, in the philosopher's vocabulary, a pie slice is an INUS condition.

Below are two sufficient causes for a disease, with the component causes shown as pie slices. There are some shared component causes (C1 and C2), but some unique component causes (C4 and C8, for e.g.). Also, I have indicated the unknown component causes as CN in the left pie and CM in the right pie.

Here's an example. We say that smoking causes lung cancer, but not all smokers develop lung cancer. There are other factors, perhaps genetic factors and other environmental factors, that contribute to one's predisposition to develop lung cancer. So in the pie charts below, Sufficient Cause A would be the constellation of factors, including smoking, that together cause lung cancer; smoking could be C3. But we also know that people develop lung cancer without ever smoking. So in the pie charts below, Sufficient Cause B would be the constellation of factors, not including smoking (C3 is not present), that together cause lung cancer. Working in a coal mine, for example, could be C8.



Sufficient Cause A.

Sufficient Cause B.

Figure 3. Two sufficient causes and their component causes.

II.4.c. Four Examples

In this section I provide four examples from very different subjects to illustrate the importance of INUS conditions and causal models. The first is an example about the effectiveness of laws mandating the use of bicycle helmets.

Bicycle Helmet Example. Vigorous debate regarding the efficacy of bicycle helmets to reduce head injury has been published in the pages of the *British Medical Journal*.⁹ Case-control studies suggest that cyclists wearing helmets have fewer head injuries than cyclists not wearing helmets, whereas time-series studies in jurisdictions that have passed helmet laws do not show a clear decrease in the rate of head injuries after helmet laws have been implemented, and in some cases these studies suggest an *increase* in head injuries after the law is implemented.

At first glance this is paradoxical. Our intuitions, supported with evidence from case-control studies, say that helmets should reduce head injuries, whereas helmet compulsion laws fail to show much benefit and in some cases possibly show an *increase* in head injuries.

There are methodological reasons that could partly explain the differences between these studies. A worry about confounders in the case-control studies could exaggerate the estimated efficacy of helmets: There is some evidence suggesting that helmet wearers are overall safer bicycle riders, are involved in less severe accidents, are richer, and more likely to be white. A worry about confounders in the time-series studies could dampen the result of introducing helmet laws. In some jurisdictions, helmet laws were introduced concomitantly with safety measures, and over the periods of these studies there have been more cars on roads, and these cars have increased in size and speed.

Leaving aside a discussion of the methodological quality of case-control studies versus time-series analyses, this paradox can be understood by thinking about INUS conditions. The case control studies give one piece of a causal pie: Helmets can cause a reduction in head injuries. But those studies don't tell about the other pieces of the pie, that is, other factors that are causally relevant to a cyclist's head injury; things like driver behavior, cyclist behavior, and road conditions. Now, there is evidence to suggest that at least some of these things change with helmet wearing.¹⁰ Drivers give less space to cyclists who are wearing a helmet, and cyclists take more risks (a 'false sense of security' phenomenon). So helmet compulsion laws don't just change one piece of a causal pie,

⁹ See especially *BMJ* 2006;332:722-725 and numerous letters in response.

¹⁰ This naturally suggests that a feedback model as with the A&E study above would be a good one to try if one wants to lay out the steps in the causal process in aid of producing what is called a causal model here.

they change several pieces. And that could partly explain the differences between the two kinds of studies.

The nice thing about this bicycle example is that it illustrates two lessons at once. First, the importance of identifying the other INUS conditions that go into a sufficient cause, i.e., the other slices in the same pie – which one can think of as ‘helping factors’ necessary in order for the policy lever to work: Helmet wearing in combination with usual driver behavior will decrease head injuries from bicycle accidents; helmet wearing with more dangerous driving may increase head injuries.

Second, it reminds us that in thinking about INUS conditions we need to pay attention to the unintended consequences of our actions. In implementing a policy we may not only produce unwanted side effects; we can, as in this case and in the Lucas example to be discussed, introduce factors that undermine the effectiveness of the very policy lever we employ. Of course we will always be plagued by uncertainty. We are in no position to predict many of the unintended outcomes of our policies. But some we can predict, if only we think about them in the right way.

The failure of the California class-size reduction program may well be a case in point. The reduction in class-size was rolled out state-wide over a very short period of time. That necessitated the hurried hire of a large number of new teachers and in consequence, teaching quality went down.¹¹ But teaching quality is a slice of the same pie as small class size: Reducing class size cannot be expected to increase reading scores without the cooperation of good teaching. The point is that this unintended consequence of the policy implementation is the kind that might well have been foretold if careful thought had been put towards it. So in producing a practicable guide based on the principles here, we will have to figure out ways to remind users to think about the unintended consequences of their policies and implementations, and to help them do so.

Homework example. Harris Cooper will tell us at this conference about evidence on how effective homework is. Let me then illustrate INUS conditions with his case. In a systematic review of the effect of homework on achievement (usually measured by a standardized test), Cooper and colleagues conclude that homework has a positive influence on achievement. This finding is fairly consistent across

¹¹ Bohrnstedt, G.W., Stecher, B.M. (eds.), 2002, “What We Have Learned About Class Size Reduction in California”, California Department of Education

multiple study designs, despite methodological flaws in all reviewed studies. The effect of homework on achievement can be usefully characterized within the INUS framework. Homework is one INUS condition (one slice of a causal pie) that contributes to higher test scores. As Cooper notes, other conditions are necessary to ensure that homework is maximally effective (the whole pie has to be in place). Cooper calls these “moderator conditions”. These include student motivation and student ability: The beneficial effect of homework on achievement will be mitigated if the student is unmotivated or unable to do the assignment. Other helping factors include having access to a proper study space, a supportive family, getting a consistent message from teachers and parents, and receiving teacher feedback on assignments – the maximal effect of homework on achievement will be when these conditions are in place. The first pie below is a visual representation of this set of INUS conditions for the outcome of higher test scores (but note that the size of the pie slices are arbitrary in this depiction).

There are other practices that might achieve the same primary outcome as assigning homework (as Cooper noted in his published review), and so these would be different pies altogether. For example, directed in-class tutorials, while resource intensive, could achieve the aim of higher test scores (though wouldn't achieve some of the other purported benefits of homework, such as a dedicated time for family involvement). The second pie below is a depiction of this (speculative) set of INUS conditions for the same outcome. We could depict more pies for all the other complexes of factors that we expect to affect test scores (like smaller classroom sizes).

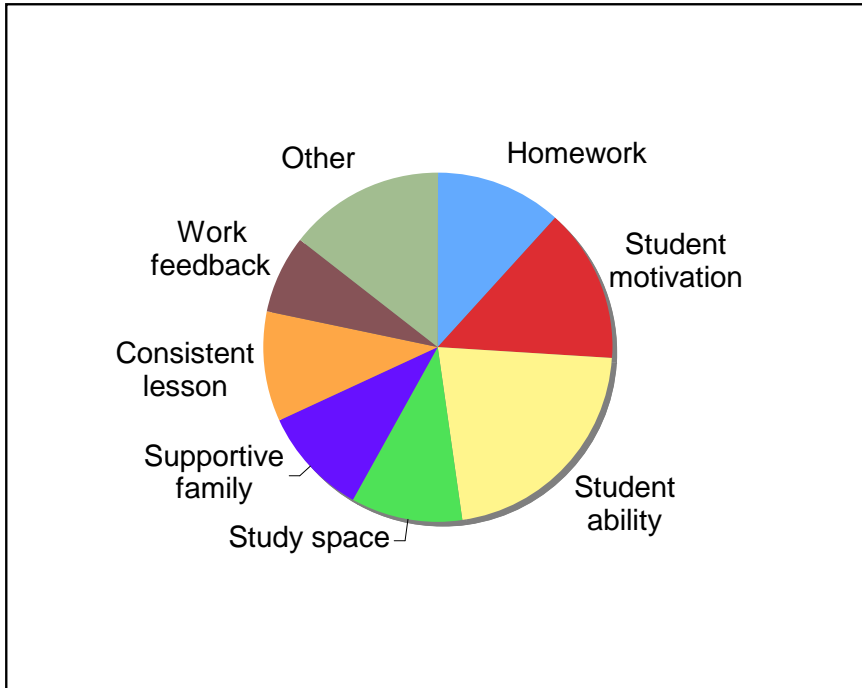


Figure 4. INUS conditions (including homework) to improve test scores.

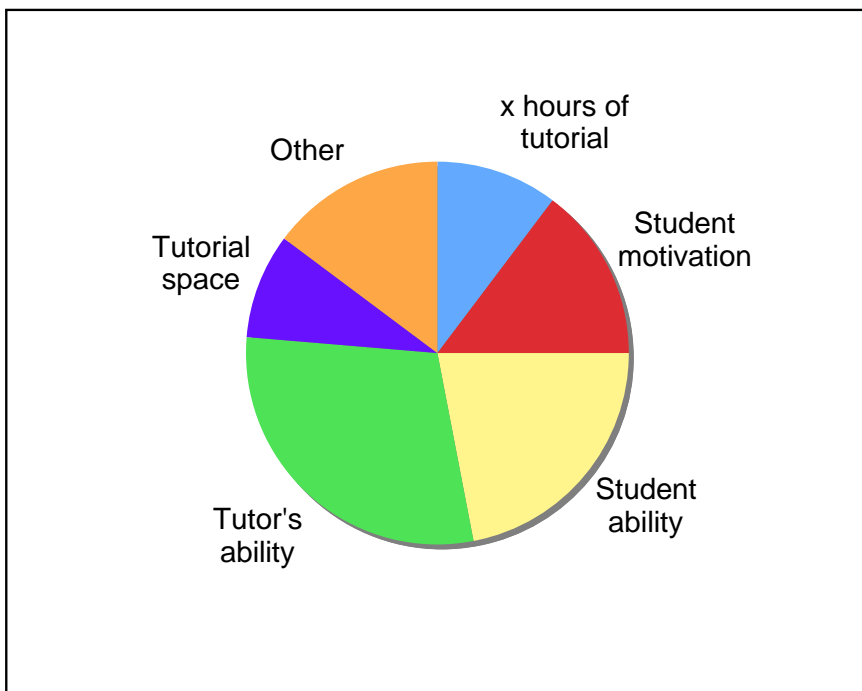


Figure 5. INUS conditions (NOT including homework) to improve test scores.

To give some meat to the idea of causal model, as I use the term here, let me describe to you another example, this a real life case that a friend of

mine (Jeremy Hardie) has been discussing with me. The case provides a nice illustration because in the course of deliberating about a policy decision he has to make – to invest in a company producing a new device or not – he was led to ask for a lot more information, and, as you will see, the information he was naturally seeking is just the kind of thing that constitutes a causal model.

Business policy example. Videoconferencing has been around for some time now. It is a technically quite mature industry in most respects. There will be improvements, particularly in cost and speed. But the betting is that in ten years' time what is on offer will differ from what we see today only as a consequence of steady incremental changes – more but better of the same.

There is one exception to this. Many meetings – whether they are no more than between the high street banker and the mortgage client, or at the other extreme two rooms full of lawyers and executives finalizing a deal between New York and Tokyo – need to end with a legally binding real signature or signatures, which both sides can see. There are fairly satisfactory ways of dealing with this, e.g. electronic signatures. But if we want the real thing, it may not always be enough, e.g., for me to sign at my end, and for my lawyers to say I have done so, and send the original document to the other remote party. At present, if I am in Tokyo and you are in New York, we cannot mimic providing a wet signature to the same document in the same place at the same time. More generally, I cannot sign in New York a document that has to be in Tokyo in an hour.

A Canadian company claims to be well on the way to solving this problem. It says that it has a working prototype which enables me in London to sign a document with a real pen that is linked via the internet to a robotically controlled pen in Tokyo, which writes a wet signature on the Tokyo document, in ink, with the same pressure and in every respect identically as has written the New York pen.

So that's the story.

How did my friend set about deciding whether to invest in this business?

First he made a list of questions that need to be answered.

1. Are such signatures legally binding?
2. Is the technology protected by patents?
3. How good/bad are the alternatives?
4. Does the pen work?
5. Who are the actual and potential competitors?
6. Have the inventors enough money and management resources to make a go of it?
7. What are the needs this is meant to satisfy?
8. Do the people with those needs have any money?

And so on. As he thinks about it and starts to get answers, the list will get longer, because new lines of enquiry will occur to him – is there about to be a world recession so we ought to wait? – and because the existing questions will break down into sub questions – can they get a better finance director?

His aim is that when he gets to having to decide, he will have the best answers he can to the best list of questions he can formulate.

Let us say that the above list is a good first go.

As I remarked in section *I.1* all of the questions like these require *evidence* to help answer them.

But for us today the important point is that only one of these questions – will the pen work in situ? – is to do with evidence for *effectiveness*, the effectiveness of the product in context as people will actually use it.

And the way he went about tackling this question maps well onto the schema which I am presenting.

He started by seeing it work. He went into the Tech Garage, as it is called, and saw the demonstration. One of the technicians signed a document with the real pen and sure enough at the other end of the garage the robot pen produced an indistinguishable wet signature.

But all this shows is that in this context this device succeeded, once.

Everybody knows that prototypes don't always work in the field.

So he now started thinking about what complications there might be in practice. Because we need the device for transactions which are very important for the parties, and must be legally binding, absolute confidence that it will work in many specific contexts is critical. Any suspicion that the signature may not appear, or may be distorted, undercuts the legally and psychologically necessary belief that the device can do no wrong.

So he set about asking for more information, thinking about possible complications, including:

1. If you have not used it before, you may be nervous, and do a bad signature. Even if it is perfectly reproduced, nobody including you will think that it is your signature.
2. The demonstrator is used to signing on the slightly slippery pad used for the original signature. Most people find it hard to sign well on that surface.
3. Does it work if you press very hard, or write fast, or use a ball point....
4. Does spilling coffee on it matter?
5. Across the room is one thing, New York to Tokyo is another. How much does distance matter, why?
6. Does temperature matter?
7. How good does the internet connection have to be?
8. Is the device sensitive to use? Does it get out of alignment, are the key components robust, does it break down, is it easy to fix?
9. A rubber band plays a surprisingly important part in regulating the movement of the robotic pen. How good is that in the field?
10. What happens if you drop it?
11. Computers freeze, have to be rebooted. Is this software system like that?

Again, this list will get longer and longer, and its components will subdivide.

Again, when it comes to deciding, he will hope that the list is complete and that the answers are as good as he can get.

But for today the key points are first, that questions about effectiveness are only a subset of all the questions which have to be asked about a product – or a policy; and second, that the decision maker has to sieve his long list to get to the subset of questions that bear on effectiveness. Then when he has, his list fills in the blanks in the causal model.

Thinking about INUS conditions might, at first glance, seem daunting, overly technical, and difficult to actually execute. What this example suggests, however, is that we think like this all the time. When we want to know if an intervention, decision, product, or policy will be effective in implementation, the best thing to do, as a start, is think about all the possibly relevant factors. The decision maker should think about what pies can cause the outcome and what the slices **are** in each of the pies. This is the first step when a building a causal model.

My friend thinking about investing in this Canadian company was doing just that. He wanted to know if, in the actual roll-out and use of this new product, the product would be effective in the specific kinds of situations it was likely to be used in. He asked the developers about some common problems that we're all familiar with in an office setting: spilled coffee, bad internet connections, poor reliability of electronic tools, and so on. What he was determining was: What are the relevant factors, or component causes, that must necessarily be in place for this new device to work sufficiently in the real world. That it, he was asking about INUS conditions.

One relevant factor might be that the device must be dry, so no coffee can be spilled on it and no sweat dripped on it from nervous signatories. Another relevant factor might be that the internet connection must be consistent and high bandwidth, so if a server crashes, we in London can't sign a document in Tokyo. Another relevant factor might be that all

the component parts (including the rubber band that regulates the movement of the pen) must be intact. We'd expect these, and others. The point is that the more of the slices that we can determine, the more we'll be able to predict the effectiveness of the product, **and** what is required for the product to be effective.

My friend may get evidence, for instance, that a properly functioning device plus a signatory who presses hard and does not sweat plus use of a felt-tipped pen produces a reasonably good signature at the other end across even a long distance so long as nothing untoward happens to the surface even after prolonged use of the machine and rubber band, independent of the quality of the internet connection and the temperature. But if the signatory drips sweat on the surface or presses lightly or messes up the surface, it generally won't work. The first causal complex is positive for results; the second not. And so on.

Notice that to evaluate effectiveness he needs to get a grip not only on what factors are relevant to a good signature at the far end, but also how they must combine. If there are too many too ideal conditions that must be met at once before a good result is reasonably likely, investing in the device may not be such a good idea.

I also want to give an example from physics. I give this example because it is well understood and not controversial. It also shows what an ideal end product of inquiry can look like: Knowing the relevant factors (the pie slices) **and** knowing precisely how they relate allows us to make accurate and extremely precise predictions of what would happen if we changed one of the factors. Knowing what slices make up a pie is less helpful than knowing the functional, formal relationship between the factors – but we need to know what the slices are before we can investigate the functional form. And most often, for real policy cases in real time, there is not much hope we will make much headway on the full functional form. That is why I have opted to focus on INUS conditions – at least when we have a reasonable understanding of these we will know what auxiliaries will be necessary if the policy variable is to have a hope of being effective. But it is at least worth having the ideal in mind since it is structurally just like the less ideal cases we must deal with in social policy.

Physics example. An object of charge q_1 at a distance r' from the earth's centre is accelerating at a distance r from a second object of charge q_2 . It is also of course subject to the earth's pull. Letting M represent the mass of the earth, its acceleration is given by¹²

$$\text{Acc} = \varepsilon q_1 q_2 / r^2 \oplus GM / r'^2.$$

The first term (the 'Coulomb acceleration') is a **sufficient** condition for acceleration – it is enough to cause acceleration. But it is **unnecessary**. Since there are many other possible causes of acceleration, the object can accelerate even without any Coulomb force. So too with the second term (the 'acceleration due to gravity'): it is sufficient but unnecessary.

Consider next q_1 . Without it there is no Coulomb force. So it is a **non-redundant**, or necessary, part of the first term. But it is **insufficient** since it cannot produce an acceleration on its own but only in consort with another charge (q_2) and some separation (r). The same is true of each of these other factors appearing in the first term as well as of the factors M and r' in the second term.

The factors q_1 , q_2 , r , M , and r' are all *causes* of the acceleration in anybody's books. And they are each, as Mackie claims, INUS conditions; each is an insufficient but necessary part of an unnecessary but sufficient condition for the acceleration.

II.5. Two central principles for a theory of use

We now have two assumptions that form the core of a theory of evidence for policy effectiveness:

Principle 1: A sure-fire way to evaluate whether a policy will be effective for a targeted outcome is to employ a 'causal model' comprising

¹² Assuming there are no other forces at work and ignoring the generally negligible gravitational attraction between the two objects themselves.

- A list of causes of the targeted outcome that will be at work when the policy is implemented
- A rule for calculating the resultant effect when these causes operate together.

Principle 2: Causes are INUS conditions

Part III: The neglected questions

With these two theoretical principles in place we can return to the three issues of quality, relevance and evaluation. If we are to evaluate policy counterfactuals via causal models, as I propose, this imposes criteria of relevance and, via that, also affects the construction of standards of quality. A causal model, even if rough and approximate, requires a great deal more information than we are in the habit of looking for.

Requisite information for evaluating policy effectiveness:
Information is needed about –

- The causal factors that will operate:
 - What factors causally relevant to the targeted outcome are in the situation? This breaks naturally into two questions:
 - What's there?
 - Is it causally relevant?
 - What factors that are introduced during implementation will be causally relevant? Again this breaks into two questions:
 - What will we do?
 - What factors among those we introduce will be causally relevant?
- How these combine in producing the effect. Here we want to pay particular attention to
 - What auxiliary factors are necessary along with the policy variable to produce the targeted effect?
 - How do different factors within a single complex (different segments of the same pie) combine?
 - How do different causal complexes (different pies) combine?

These are empirical questions and any answers that are proposed should have evidence to support them. This sets our criterion of relevance:

An empirical claim is *evidentially relevant* to a policy effectiveness estimate just in case it helps to establish

- i. What's there in the target situation
- ii. What will be introduced in implementing the policy
- iii. The causal relevance of any of the above factors for the targeted effect
- iv. The method of calculating joint effects.

I note that this formulation does not eliminate questions of relevance; it only pushes them back a level. One still needs to know what kinds of evidence are relevant for establishing what's there, what factors are causally relevant, and for claims of how they combine. The point at the moment is that relevance is a far broader church than the one we are used to practicing in. In principle we should have evidence for all the components that need to be used in supporting an effectiveness claim. In practice some facts will be fairly obvious and not need much evidencing; and we will necessarily take a good many shortcuts. But the task for this paper is not to jump into shortcuts but rather to lay a principled foundation for judging policy effectiveness, including evaluating shortcuts and deciding how much to bet on them.

The broad-church relevance criteria in turn affect issues of quality. Most current guides focus on the quality of *efficacy* claims. Depending on context and philosophical leanings, these can be read as claims that the policy can work, or that it does work under specific conditions, or about its average effect under special implementations across some range of conditions. Efficacy claims help support the causal relevance of the policy variable, which is part of category iii. The usual ranking schemes police the quality of efficacy claims. But how shall we police the quality of the other kinds of claims needed as evidence for the remainder?

This issue needs to be faced and dealt with, however fallibly, in designing a well-grounded comprehensive advice guide, convenient as it would be to ignore it. Recall my cautions about chains of argument. It is no use having one or two highly certain premises in arguing for or against policy effectiveness. The conclusion can be no more certain than the weakest premise. In adopting a policy, one is betting, willy-nilly, that all the requisite questions have the right kinds of answers. One can do that on a wing and a prayer. But that is not an evidence-based decision. So it is incumbent on us here to figure out reasonable and usable sets of advice about how to manage the need for evidence and not institutionalize ignoring the need.

Here is probably where I first get into trouble with those who maintain that RCT-backed policies are the only ones with a reasonable evidence

base. I am very happy to take RCTs as a gold standard. In my view, they are provably good at establishing efficacy conclusions, as are a number of other methods, such as deduction from sound theory and certain econometric methods.¹³ But that is from the point of view of the evidence producer.

Evidence users want to know if a policy will work for them. That, as we knew all along and as I have been stressing here, requires a lot more information than the information supplied by an RCT or a good econometric model that establishes the efficacy of the policy variable; and that information needs evidence, including evidence about what can sometimes be a really tough question – how the causes combine.

Things look very different when we survey the whole problem and from the user's point of view than they do when we look from the point of view of the scientist charged with producing sound results to offer up as evidence. Imagine we are offered two policies. One has very good RCT evidence in favor of its efficacy but we have very weak ideas and information about what the requisite helping factors and major inhibitors for it are. The second is a policy that comes with a theory that suggests what helping factors are needed – and these are ones that are either in place for us or cheap to put in place. Suppose the theory has some reasonable evidence in its favor and the associated policy has some evidence for efficacy, but not gold standard? Which has stronger evidential support in favor of its claim to be effective if we implement it?

This is a question that depends on the actual details, and in many cases there won't be any very good answer. But sometimes normal educated judgment will – and should – reasonably go for the second policy though the evidence for its efficacy is clearly less compelling. That's why I made such an issue at the start about chains of support, which are only as strong as their weakest link. Adding more rigor at one point can raise the overall probability but that can be easily offset by too much guessing later on. We all know this all too well. I keep stressing it because I think we do not have guides that provide enough of the right kind of advice considering all that is required.

It would be wrong of course to suggest that these other issues have not been tackled at all. A lot of hard work and serious thought has been put into what is already available. But much of it is piecemeal, directed at specific problems, starting from specific places in midflow. We need a foundation that considers the problem of evaluating effectiveness

¹³ See 'Causal Claims: Warranting Them and Using Them' in NC's *Hunting Causes and Using Them*.

counterfactuals as a whole. It is only on the basis of such a foundation that we will be in a position to judge how reasonable it is to leave out specific considerations, to take specific shortcuts, and to make specific heroic assumptions. The theoretical foundation proposed here is meant to do just that job. It is not the only one possible, but it is a foundation laid specifically with a view that practicable advice needs to be built up from it.

Part IV. Making life somewhat easier

Perhaps suggesting that we want to provide an advice guide based on the idea of constructing a causal model sounds like a tall order. Sometimes it is. Particularly when if there is a demand for very precise predictions or predictions that we can be very sure of. But we should not be too frightened of the project. For it is one we are well used to. We regularly build causal models in making decisions in our daily lives as we think through the possible effects of our actions and policies. Consequently, the schema should not be seen as too exotic or impractical. It, or something like it, is used all the time.

Yesterday, for example. My favorite red-and-white-striped tee shirt was soiled looking. Should I wash it in hot water? Well: hot water only works if the shirt has a reasonable amount of cotton in it; and it won't work against coffee or ink stains. Even with cotton it can be counterproductive if the hot water makes the stripes run. And I know that I have to be especially careful in loading a hot wash since the shirt will go grey if I inadvertently include some dark socks. All told, given my cotton shirt with garden dirt and the determination to be careful in loading the machine, I reckoned (correctly) that the shirt would come out clean in a hot water wash.

This is a homely example but it illustrates my claim that we build what I call 'causal models' all the time in making policy decisions. The problem for evidence-based policy is how to use evidence to build them better and to estimate the degree of confidence we should have in the results of our efforts.

Perhaps you do not find this familiar kind of example comforting, nor the other real-life story I told about evaluating the effectiveness in situ of the distant signature writing machine. The idea of insisting on causal models stills sounds too daunting. Nevertheless, Nature will use a causal model to decide what outcomes to produce when we implement our policies

whether we wish to follow her lead or not. The right answers to our questions of quality and relevance will depend on the models she chooses. So, daunting or not, I think advice on these questions should reflect that.

We can however sometimes make the job less daunting. Consider: We would in general like to be able to predict the actual value of the effect that would follow the implementation of a proposed policy. By *just how much* will household burglaries drop if a community-wide property marking program is adopted? But often that will be difficult because we do not know how to predict what else will be going on. What other causes of burglaries will be in place at the time? Often we cannot assume that the causes will be the same then as they are now. (This is the reason JS Mill said economics cannot be an inductive science.) So we can't estimate what other 'sufficient' causes will be at work, let alone what their combined effect will be. In these cases we may be satisfied with reasonable assurance that the policy will produce an improvement in the effect over what would be the case without it, whatever that is. If so, life is somewhat easier.

In this case establishing just a couple of facts will allow us to ignore the other complexes that make up alternative 'sufficient' causes (all the other 'pies') and concentrate on complexes that include the policy variable.¹⁴ What we need to know is that no alternative complex of causes will be so dominant that it swamps the policy variable, either positively or negatively, making its effects negligible. For instance, there is no point offering a low cholesterol diet to improve longevity to a man who will be executed in the morning. Nor in installing a fancy electronic lock on my old Rover since, my daughter assures me, there is no chance that it will be stolen.

So...If we are content to settle for the claim that the policy will make an improvement on what would otherwise have been the case were the policy not implemented, and we have good enough reason to think that nothing will swamp the effects of the policy, then we are justified in focusing just on the policy variable and the factors necessary for it to succeed in producing the targeted effect.

¹⁴ Complex relations between the sufficient causes are possible, however, so sometimes even for these kinds of cases it is not a good idea to ignore other causal complexes. Suppose, for example, that adjusting one component cause of a cluster (one slice of a pie) modifies another component cause of the same cluster – the example about bicycle helmets illustrated this – then, if the secondary modified component is also a component of another cluster, the effect of the second sufficient cluster will be modified.

A warning reminder is worth making, however. We all know that a successful policy – one that did indeed produce an improvement over what would have been – can easily be judged a failure if it does not produce an improvement over what used to be. Policy consumers are apt unimpressed by the claim: ‘Yes things have gotten worse. But they would have been far worse still if we hadn’t acted as we did’ even if it is true. In these cases one needs to have a good account of what other causes operated to counter the policy effects, and good evidence that that is really the correct story.

Part V. Mechanisms: A principle in aid of practical advice

The primary purpose of the ‘theory of evidence for use’ is to provide principled grounds for practical advice. To this end I propose to borrow one more tenet from my colleagues in philosophy to add to the basic principles of the theory, albeit one more informally put.

Principle 3: Mechanisms matter.

Methodologists like RCTs in part because RCTs provide evidence for causal relations without our having to know the mechanisms by which the cause produces its effect. Policy makers generally share this lack of interest in mechanisms. They are concerned only with whether the policy will produce the targeted results and do not care about the mechanisms that will drive the result. Still, when we want to try to put a cause to work, getting a better understanding of the mechanism can make a big difference. The importance of mechanisms for causal discovery, causal understanding, and causal prediction has been heavily stressed in recent philosophical literature. What though is a mechanism?

I told you that causation is all the rage in philosophy now; mechanisms are centre stage in the discussion. Not surprisingly then there are a wide variety of different characterizations on offer.¹⁵ Here I am not going to

¹⁵ I shall describe some of these approaches to stress by contrast that none of these are what I mean by ‘mechanism’ here. Here I mean an answer to a how question that can help in finding INUS auxiliaries. As to other senses of mechanism: Judea Pearl has shown you causal models that take the form of linear equations, one equation for each effect variable on the left-hand-side, laying out a complete set of causes for it on the right-hand-side. Many people call these equations ‘mechanisms’, as in a simple supply and demand model in economics where the equation for the quantity supplied is said to describe ‘the supply mechanism’; that for the quantity demanded, ‘the demand mechanism’. I talk about a mechanism (or a ‘nomological machine’) as a fixed (enough) arrangement of parts that has the capacity when set running to give rise to stable in-put/out-put relations. (NC, *The Dappled World*) For my UCSD colleague Bill Bechtel, “A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the

rely on any of these (including my own) since they are generally both too narrow and too abstract to be of help to those non-expert in the sciences. Rather I want to make use of an informal notion of mechanism common to many of the formal accounts. This is a notion that can provide a help for policy makers – a prod for the imagination – in identifying the auxiliary factors that are necessary along with the policy variable to produce the targeted effect. For these purposes I take a mechanism to be an answer to the question:

How would the policy variable bring about the desired effect?

Two different ways of answering can help in finding auxiliary factors:

1. Trace out the causal pathway from policy variable to effect. Seeing what should come next at each step helps focus on what would be required in addition to the policy variable to make the next step happen.
2. Many social results are achieved by calling into play general, often familiar, routine phenomena, such as loyalty, mother-love, fear of punishment, desire to conform, desire to be recognized. Different helping factors will be required, besides the policy variable, to set different general mechanisms into operation. So recognizing which general mechanisms will be called on can be a big help in identifying the necessary auxiliaries.

V.1. Tracing the causal pathway: an example from economics.

Robert Lucas famously argued that it is generally counterproductive for governments to intervene to regulate the economy on the basis of observed regularities.¹⁶ That's because people will figure out what is happening and act differently, in consequence undermining the very regularity the government depends on for predicting the effects of its policies. One of his striking examples is that of the Phillips curve, the empirically observed trade-off between inflation and unemployment that was used by policy makers in the 50s and 60s to control unemployment

mechanism is responsible for one or more phenomena." (Bechtel, W. and Abrahamsen, A. (2005). "Explanation: A Mechanistic Alternative." *Studies in History and Philosophy of the Biological and Biomedical Sciences*, 36, 421-441.) Alternatively Peter Machamer, Lindley Darden, and Carl Craver define mechanisms as: "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer, P., Darden, L., Craver, C. "Thinking About Mechanisms" *Philosophy of Science*, Vol. 67, No. 1. (2000), pp. 1-25.)

¹⁶ Lucas, Robert (1976). "Econometric Policy Evaluation: A Critique." Carnegie-Rochester Conference Series on Public Policy 1: 19–46.

via inflation. Lucas uses a ‘rational expectations’ model to show that the Phillips curve will break down if people know what the government is doing. His model reflects a story that answers the question, ‘How does rising inflation produce a lowered rate of unemployment?’ In so doing it unearths some crucial auxiliary factors that have to be in place besides inflation if inflation is to reduce unemployment.

We have seen a version of the Phillips curve already in section *II.3.b.*:

$$* \quad y_t = \theta\beta[p_t - p_{t-1}] - \theta\beta\pi + y_{pt}.$$

According to this equation an increase in p should make for an increase in output. We can suppose that an increase in output will in turn lead to an increase in employment. Hence the equation describes a trade-off between inflation and unemployment. But it is of no use for policy says Lucas. His story goes like this: How much output suppliers produce depends on the price they expect their good to sell for and on what they expect their expenses to be. In the Lucas model, the average price for goods in the economy serves as a proxy for expense. So in the model, the amount of a good supplied in a given period depends on the ratio of the price of the good to the expected economy-wide price for goods in that period. Lucas assumes that suppliers will be good guessers about the economy-wide price: The economy-wide price that they expect is the average economy-wide price that actually obtains. In this case overall output of a good will be proportional to the ratio of the price of the good to the mean of economy-wide prices. So the output of a good will be greater when the price of the good exceeds the mean of prices across the economy. That means that there will be a positive relationship between output and price increase. Another causal process that I won’t describe provides Okun’s law, under which increases in output lead to increases in employment. The two processes together thus imply that rising prices will reduce unemployment.

What happens if the government decides to intervene to increase inflation over what it would have been? Assuming that the Phillips curve (along with Okun’s law) still holds, unemployment should go down. Not so, Lucas argues, because suppliers are good guessers about the average price. If they know about the government’s actions, they will predict the average price rise that will in fact occur. The expression for output of a good has price for the good in the numerator and, assuming suppliers are good guessers, average price rise in the denominator, recall. So the rise in price suppliers see for their product, which appears in the numerator, will prompt an increase in output only if it is not offset by the increase in the

average prices in the denominator that inflation will entail. Indeed, if the denominator goes up proportionately faster than the numerator, the government policy to increase prices in the economy can even create a drop in output and thereby cause an increase in unemployment.

Where do we see this important factor – the average of economy-wide prices – in equation *? It is hidden in θ . But rehearsing the causal process step-by-step, as in the Lucas story, brings it out of hiding. The only way that inflation can increase output is if the average price rise this will involve does not result in an increase in the overall price rise expected by suppliers big enough to offset the rise in price the suppliers see for their own products. The trade-off between inflation and unemployment holds when it does just because suppliers do not expect the overall rise in prices. The requisite helping factor we learn about then is the failure of the suppliers to foretell the inflation. That suggests that if the government is going to succeed in the strategy of encouraging inflation in order to reduce unemployment it had better not let people know that that is what it is doing.

This case illustrates two points of interest here. Equations are nice because they express precise quantitative relationships. Still, true equations may leave a lot out, and especially a lot we need to know for policy success. Even equations that are 100% descriptively accurate can fail to lay out the factors necessary to enable the cause they picture to produce the expected effect. Second, thinking through the causal process step-by-step – answering a *how* question – can make these helping factors apparent.

V.2. Identifying the means of production: a criminology example.

I should like to quote an example from Nick Tilley¹⁷ at length to illustrate how thinking about the general mechanisms called into play by the policy variable in order to produce the effect can also help in identifying auxiliary factors:

Take property marking. What is it about it that is expected to ‘work’ as a crime prevention measure? Property marking might increase the risk to offenders by making it more likely that they will be caught with stolen property, successfully prosecuted and punished. This in turn may mean:

¹⁷ Nick Tilley, ‘What’s the ‘what’ in “what works?”’, ms, Jill Dando Institute of Crime Science, UCL

1. More offenders are incapacitated,
2. Some offenders are deterred from future crime,
3. And/or other prospective offenders are deterred as they come to appreciate what will happen to them if they try to commit the crime.

Alternatively (or in addition), the perceived increased risk of apprehension, regardless of the reality:

4. May lead (some) prospective offenders not to commit crime in the first place.

For property marking to 'work' in relation to any individual offender in the first way,

- a) Property that is liable to be stolen has to be marked,
- b) Offenders have to fail to remove or disguise the marks,
- c) Authorities have to check that property that might be stolen has property marks on it,
- d) Police have to link the marked property back to those from whom it has been taken,
- e) Those found with the stolen property have to be unable to cook up a plausible enough story about why they legitimately have it in their possession,
- f) The prosecutor has to be persuaded that the case is worth taking to court,
- g) The judge/jury have to be persuaded by the evidence,
- h) A custodial sentence has to be passed, and
- i) There have to be offences that the incarcerated person would otherwise be committing but for the fact that he or she is in prison.

For property marking to work in the second way, (a-i) have to be in place, and

- j) the penalty has to be sufficiently salient that the offender makes decisions that do not lead to further offences or which lead to fewer offences.

For property marking to work in the third way (a-j) have to be in place, and

- k) Prospective offenders need to know, appreciate and sufficiently fear the penalties applied that they will make decisions not to commit offences that would otherwise commit.

For property marking to work in the fourth way (a-k) need not be in place, but,

- l) Prospective offenders must know that property is (or may very likely) be marked
- m) Prospective offenders must be persuaded that the marking significantly increases their risks of being caught and penalised if they steal the marked goods, and
- n) The expected penalties must be sufficient to lead them to decide not to commit the offences they would otherwise commit.

If a net fall in crime is to be produced by property marking, further conditions are needed,

- o) The crimes prevented by any of the four means must not be substituted in terms of volume, value or severity, either by the same or substitute offenders, and/or
- p) Offender uncertainty about the range of offences, goods and places where property marking has taken place leads them to avoid offences even where or in relation to goods not property-marked.

Thus, what might work in property marking to bring about a crime drop through property marking depends on contextual contingencies.

Tilley's 'contextual contingencies' are just the auxiliary factors I have been talking about in discussing INUS conditions, factors that must be in place along with property marking in order for property marking to bring about a drop in crime. Focusing, as he recommends, on *how* property marking is supposed to achieve these results directs our attention to these essential factors.

Part VI. In Sum

The ultimate aim is to construct a relatively comprehensive advice guide for evaluating policy effectiveness claims, a guide that is practicable and at the same time rests on sound general principles. To this end I propose three principles. First, policy effectiveness claims are really causal counterfactuals and the proper evaluation of a causal counterfactual requires a causal model that i) lays out the causes that will operate and ii) tells what they produce in combination. Second, a cause for these

purposes will be an INUS condition, and it is important to review both the different causal complexes that will affect the result (the different pies) and the different components (slices) that are necessary to act together within each complex (or pie) if the targeted result is to be achieved. Third, a good answer to the question, ‘How will the policy variable produce the effect’, can elicit the set of auxiliary factors that must be in place along with the policy variable if it is to operate successfully.

A guide based on these principles will have to help users construct their own causal models and use evidence to judge how good they are. It should also provide shortcuts, what Gerd Gigerenzer¹⁸ has called ‘cheap heuristics’, that can achieve near enough the same conclusions with less input. Most of these will apply only in special conditions. Part of the job before offering them to users will be to show that these shortcuts are indeed good ones in the right circumstances, then to describe the circumstances for the users in a way that can be understood and applied.

All this is something of a tall order for users. That just makes our job hard. We need to do the best we can to help those who need to evaluate effectiveness do so as best possible, even if the process will inevitably be flawed. Recognizing that it will be flawed means making clear that policy effectiveness judgments will almost never be very secure; and so far as possible, one should hedge one’s bets on them. It does not mean giving up on the attempt to construct a causal model, or alternatively defending that a particular short cut will do almost as well. For, as I have stressed, when one bets on an effectiveness counterfactual, one is betting, willy-nilly, on the causal model that underwrites it. The whole point of evidence-based policy is that bets like this should be taken consciously and be as well informed by evidence as is practicable. It’s no good ducking the problem. We’d better just get on with figuring out how to make this all as simple and user friendly as possible.

¹⁸ *Simple Heuristics That Make Us Smart*. 2000. Gerd Gigerenzer, Peter M. Todd, ABC Research Group. Oxford University Press.

REFERENCES

Tilley, N (Forthcoming) 'What's the "what" in "what works?"? Health, policing and crime prevention.' In J. Knutsson and N. Tilley (eds.) *Evaluating Crime Reduction*. Crime Prevention Studies Volume 24. Monsey NY: Criminal Justice Press.