

**Nancy Cartwright**  
***Evidence for Evidenced-Based Policy***  
Talk at a Home Office Seminar on Criminology and Evidence-Based Policy  
10th June 2008

I start by announcing that I am firmly in favour of evidence-based policy, despite the disappointments with it. Though evidence is not the only consideration, it is clearly better to look at the evidence and think hard about it than not. I am not at all surprised however that it has had disappointing results in both the UK and the US since we are not giving good advice about how to go about it.

*Two lost questions*

For today I stick with the issue of evidence for effectiveness; that is, for the claim that a policy will produce the desired ends when implemented – how, when and where it will be implemented. I start with a truism: In evaluating the effectiveness of a proposed policy we need credible evidence that speaks for or against the policy and we need to know what to do with the evidence when we have it. This truism naturally generates three questions:

1. What counts as *credible* evidence – **what evidence claims are likely to be true?**
2. What evidence claims are *relevant* – **what claims speak for or against the proposed policy and how strongly?**
3. How should the evidence be *integrated* – **how should the probability of the policy being effective be established in light of all the evidence?**

Begin with *credibility*. We do not wish to enter claims into the record of evidence that are themselves not likely to be true. Compare: In deciding if a person is guilty or innocent we do not take into account the testimony of a witness without good reason to think the witness is telling the truth. So too in deciding if a policy will be effective or not.

Suppose however just for the moment that this is no issue. Suppose we have at our disposal the entire encyclopaedia of unified science, an encyclopaedia that contains within it all the true claims there are. But for deliberating about a particular policy we are not going to cart the entire encyclopaedia to the table. Rather we want a selection – we want on the table only true facts that are *relevant* to the effectiveness of the policy. And given a collection of relevant facts we want to know how to assess the probability that the policy will be effective in light of them. How are we supposed to make all these decisions?

Let us look at an American case that is particularly egregious. It does not have to do with crime but with education: the so-called ‘No Child Left Behind’ legislation. I use it because it illustrates my points clearly and sharply. Go to the US Dept. of Education website, as school masters are supposed to do, and this is the advice you find:

**From**  
**Identifying and**  
**Implementing Educational**  
**Practices Supported by**  
**Rigorous Evidence:**  
**A User Friendly Guide**  
December 2003

U.S. Department of Education  
Institute of Education Sciences  
National Center for Education  
Evaluation  
and Regional Assistance

**How to evaluate whether an educational intervention is supported by rigorous evidence: An overview**

---

**Step 1. Is the intervention backed by "strong" evidence of effectiveness?**

<p><b>Quality of studies needed to establish "strong" evidence:</b></p> <ul style="list-style-type: none"> <li>• Randomized controlled trials (defined on page 1) that are well-designed and implemented (see pages 5-9).</li> </ul>	+	<p><b>Quantity of evidence needed:</b></p> <p>Trials showing effectiveness in —</p> <ul style="list-style-type: none"> <li>• Two or more typical school settings,</li> <li>• Including a setting similar to that of your schools/ classrooms. (see page 10)</li> </ul>	=	<b>"Strong" Evidence</b>
--	---	--	---	--------------------------

---

**Step 2. If the intervention is not backed by "strong" evidence, is it backed by "possible" evidence of effectiveness?**

<p><b>Types of studies that can comprise "possible" evidence:</b></p> <ul style="list-style-type: none"> <li>• Randomized controlled trials whose quality/quantity are good but fall short of "strong" evidence (see page 11); and/or</li> <li>• Comparison-group studies (defined on page 3) in which the intervention and comparison groups are <i>very closely matched</i> in academic achievement, demographics, and other characteristics (see pages 11-12).</li> </ul>	+	<p><b>Types of studies that do <u>not</u> comprise "possible" evidence:</b></p> <ul style="list-style-type: none"> <li>• Pre-post studies (defined on page 2).</li> <li>• Comparison-group studies in which the intervention and comparison groups are not closely matched (see pages 12-13).</li> <li>• "Meta-analyses" that include the results of such lower-quality studies (see page 13).</li> </ul>
--	---	---

The left-hand side of the + sign – ‘quality’ of evidence – plus all of step 2 is in answer to the first question: when is a particular kind of evidence claim credible? The US Dept of Education’s is typical of a large number of evidence-ranking schemes currently available, including the Maryland rules. These schemes gauge the credibility of causal claims based on their associated research design. The advice they give is excellent, and detailed. As we know, there are very long checklists of demands a trial must meet to earn the label ‘well-designed and well-implemented’.

Questions 2 and 3 are lumped together under the heading ‘quantity of evidence’. The advice on relevance, Question 2, is short indeed – the RCT should be in ‘settings similar to that of your schools/classrooms’ and page 10, which this page refers to, adds only 4 lines describing one case – trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students. This is like the warning: Beware, results about police-initiated fear reduction programmes in large estates in South Birmingham and Southwark may not carry over to Belgrave Square.

The answer to question 3 is equally short: two positive relevant RCTs are strong evidence for effectiveness. So with two good RCTs we can assign a high probability that the policy will work in our school.

One obvious problem is that this violates the principle of total evidence, which is at the heart of practice in the natural sciences: look at all the evidence, strong and weak, of various kinds and from various sources. Other guidelines do better in this regard. For instance the NICE guidelines allow us to consider all the evidence, weak and strong, pro and con. But we are still without advice about how to go about considering it.

The well-known Maryland rules prominent in criminology are very much the same. Here is the entire section headed '*integration of evidence*'. This is where we should look for an answer to question 3 – how should we combine the evidence to evaluate the probability that the policy will be effective when we implement it? The integration offered is no integration at all. We are told to judge that a programme 'works' if it has positive results in two rigorous studies, with an acknowledged need for 'judgement' in the end:

**Integration of Evidence.** The end product of the analysis of empirical evidence contains a range of findings with respect to effectiveness. In the interests of clarity of presentation for policy analysis purposes, we organize the presentation of material in each chapter by the content of the findings, rather than the priority of the program. The content is defined both the strength of the scientific evidence and the strength (and direction) of the program effects. We ultimately report on four categories of effectiveness.

Program categories are sorted into these effectiveness categories using the following rule:

**Works (1):** At least two studies with methodological rigor greater than or equal to "3" reporting significance tests have found crime prevention effects for the program condition, and where effect sizes are available, the effect is at least one-tenth of one standard deviation (e.g., effect size = .1) better than the effects for the control condition, and the preponderance of the evidence supports the same conclusion.

**Doesn't Work (2):** At least two studies with methodological rigor greater than or equal to "3" reporting significance tests have found no effect favoring the program condition, and the preponderance of other evidence supports the same conclusion.

**Promising (3):** At least one study with methodological rigor greater than or equal to "3" reporting conventional significance levels has found crime prevention effects for the program condition, and where the effect size is available, the effects are at least one-tenth of one standard deviation better than the effects for the control condition OR the preponderance of evidence favors the program.

**Don't Know (4):** Categories with empirical evidence which do not fit one of the above are included in this residual category.

We must, of course, be extremely careful in labeling any program category as highly certain to be good or bad in its effects. Yet we must also be clear enough to make our conclusions useful, no matter how much we anticipate that science is always provisional and that our conclusions may be changed by next year. The large number of programs to be reviewed almost guarantees that some have strong evidence of extreme effects, both positive and negative. Yet most extreme results are from single, unreplicated studies. It is just as important not to conclude too much from a single negative result as from a

single positive one. A single evaluation, even with strong evidence, cannot be assumed to generalize to all or most other settings. The primary objective of identifying promising results should be to foster replications, and the more promising the results the greater the replication need. Where there is substantial consistency of evidence in one direction, the senior author of each chapter makes a judgement and defends it in the text.

Here again we have *no* advice about question 2 – when is an evidence claim relevant to assessing the probability of effectiveness for a proposed policy; and faulty advice about question 3 – how to settle on the probability of effectiveness: to wit, don't look at the total picture. (This despite the fact that it is contrary to what we do in physics.)

So, we have arrived at my first point. We put a huge amount of effort and expense into question 1, which ensures that there really is a *causal* connection between programme and outcome somewhere. We do so in aid of evidence-based policy. But question 1 is not a policy question; it is a question in pure science. This is like Galileo rolling balls down an inclined plane or dropping them from the leaning tower of Pisa to test the effects of gravity on them. Galileo's experiments yield a great deal more precise information than any RCT; yet Galileo's results are a long way from the policy question: predicting the trajectory of our cannonballs under the influence of gravity.

Question 1 is a question in pure science; the policy-relevant issues about evidence are encoded in questions 2 and 3. Yet all the rigor, and almost all 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 defence 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 huge waste.

So if we do want decision-makers to use evidence as the basis for judgements about whether a policy will be effective when implemented, we need to develop far better guidelines for questions 2 and 3. Whatever else is going on, I am not surprised that evidence-based policies are not proving as effective as hoped. For if policy makers were assiduously following the advice of the current dominant guides, they would be making wrong assessments a great deal of the time.

#### *How to use evidence: the need for a causal model*

My second point is that the right answers to these two crucial questions in any particular case depend on the right choice of a causal model for that case, and advice on the two questions should reflect that.

Consider a simple case using everyday physics. I choose this because it is simple, well understood and I am not likely to get involved in subject-specific debates in criminology. 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 at this stage *entirely on 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. This has easy analogues in crime reduction. Consider the nice example of Nick Tilley and Ray Pawson. If CCTV cameras in car parks reduce car theft by discouraging thieves, they need to be visible to be effective at all. But if they work by alerting the police to get there in time to arrest the thieves, they had better be hidden. We need to know the necessary auxiliary factors.

Then the acceleration caused by the magnet is 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 boot. Finally, we need to know how all these factors combine to produce a result. Often in criminological or other social contexts we assume simple 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 necessarily get 74.

The point is that whether the magnet will be effective at all in the target situation and to what extent depends on the causal structure of the situation. So the most direct way of predicting its effects is to construct a causal model of the situation and estimate them.

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 the uncertainties. Sometimes there are shortcuts, 'cheap heuristics'. For instance, one powerful cause can swamp everything else so you don't need to model the rest. If you are going to put a bullet through someone's heart you do not need to find out what his cholesterol levels are to calculate his longevity. Or, as with the magnet and the matchstick, the absence of some necessary auxiliary can show that a policy will not be effective without further thought. For instance an elaborate schedule of rewards and punishments is not going to work in cases where people's actions aren't responsive to their utilities.

Failing a nice heuristic for a case, the right advice is: do your best with the resources and time available to build a causal model. This is my second point. 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. Generally then it is better to think about it than not, and to do so in a systematic and deliberate way.

Finally, tying the two points together. If what we are aiming for are reasonable causal models for our policy decisions, this provides direction for constructing advice for how to answer the three central questions, including the first, which I have ignored – the question of credible evidence claims. Because now we see we need information far beyond the kinds of causal claims that are the subject of the standard evidence-ranking systems. Those causal claims bear on one piece of the causal story. The guides show us how to decide if a magnet can lift – because it has definitely lifted somewhere in some circumstances. That does not tell us at all that it will lift in our circumstances. For that we need to know what the rubbish is like as well, what situation it is in and how all these factors behave together. These judgements should have solid backing as well if predictions are to be relied on.

To repeat, our assessment of the probability of effectiveness is only as secure as the weakest link in our chain of reasoning to arrive at that probability. We may have to ignore some issues or make heroic assumptions about them. But that should dramatically weaken our degree of confidence in our final assessment. Rigor isn't contagious from link to link. If you want a relatively secure conclusion coming out, you'd better be careful that each premise is secure going in.

---

Where do RCTs fit? Curtis Meinert, prominent expert on clinical trial methodology and outspoken opponent of the US NIH act demanding studies of subgroups: 'There is no point in worrying whether a treatment works the same or differently in men and women until it has been shown to work in someone.' (p. 108 in S. Epstein's *Inclusion: The Politics of Difference in Medical Research*, 2007, Chicago: Chicago University Press) That's what the RCT shows – and that is pure science.