Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps

Nancy Cartwright*

There is a takeover movement fast gaining influence in development economics, a movement that demands that predictions about development outcomes be based on randomized controlled trials. The problem it takes up—of using evidence of efficacy from good studies to predict whether a policy will be effective if we implement it—is a general one, and affects us all. My discussion is the result of a long struggle to develop the right concepts to deal with the problem of warranting effectiveness predictions. Whether I have it right or not, these are questions of vast social importance that philosophers of science can, and should, help answer.

1. A Focus on Development Economics. The World Bank estimates that in developing countries 178 million children under age 5 are stunted in growth and 55 million are underweight for their height (World Bank 1995). Malnutrition leaves children vulnerable to severe illness and death and has long-term consequences for the health of survivors. The bank has funded a wide range of nutritional interventions in developing countries, in Latin America, the Caribbean, Africa, and East and South Asia. This included the Bangladesh Integrated Nutrition Project (BINP), modeled on its acclaimed predecessor, the Indian Tamil Nadu Integrated Project (TINP). What was integrated? Feeding, health measures, and, centrally, education of pregnant mothers about how better to nourish their children.

TINP covered the rural areas of districts with the worst nutritional status, about half the Tamil Nadu state, with a rural population of about 9 million. Malnutrition fell at a significant rate. The World Bank Independent Evaluation Group concluded that half to three-fourths of the decline in TINP areas was due to TINP and other nutrition programs in those areas.

*To contact the author, please write to: Department of Philosophy, Durham University, 50 Old Elvet, DH13HN, Durham, UK; e-mail: nancy.cartwright@durham.ac.uk.

Philosophy of Science, 79 (December 2012) pp. 973–989. 0031-8248/2012/7905-0033$10.00

The Bangladesh Project was modeled on TINP. But Bangladesh's project had little success. A Save the Children UK assessment concludes that program areas and nonprogram areas still had the same prevalence of malnutrition after 6 years, despite the fact that the targeted health educational lessons sank in to some extent: Caregivers in the BINP areas had on the whole greater knowledge about caring practices than those in non-BINP areas. Why then did the project fail in Bangladesh?

Before that we had better ask: Why should it have been expected to succeed? The extrapolation to Bangladesh from uncontroversial success in India was not warranted, I shall argue, because it was based on simple induction; and simple induction is no better a method in social science than in natural science and no better in policy science than in pure science. Moreover, we can do better, and often with knowledge already at hand.

My talk will concentrate on development economics and on a vigorous takeover movement fast gaining influence there, a new methodology to improve development outcomes: randomized controlled trials (RCTs). As a Public Radio International interview reports, "A team of economists at [Massachusetts Institute of Technology] says it's time for a new approach—one that makes prescriptions for poverty as scientifically-based as prescriptions for disease" (http://www.pri.org/theworld/?q=node/10887). MIT's Esther Duflo is one of the leaders of this movement. She tells us that "the last few years have seen a veritable explosion of randomized experiments in development economics" (Banerjee and Duflo 2009) and that "creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century" (Duflo, quoted by *Lancet* 2004, 731). Witness also the recent *Journal of Economic Perspectives* symposium on a paper commending RCTs by my London School of Economics colleague Steve Pischke and another MIT economist, Joshua Angrist. They cite one exemplar of good research design: "in a pioneering effort to improve child welfare, the Progresa program in Mexico offered cash transfers to randomly selected mothers, contingent on participation in prenatal care, nutritional monitoring of children, and the children's regular school attendance" (Angrist and Pischke 2010, 4). They add, quoting Paul Gertler, one of Progresa's original investigators, that "Progresa is why now thirty countries worldwide have conditional cash transfer programs" (4). That's serious extrapolation!

And, to see why I am concerned: Even since I wrote this in draft I have learned that the father of Progresa, Santiago Levy, says that many of the places that want conditional cash transfer programs are places where they will obviously fail. In some of these countries success would require people to go to clinics that do not exist (Deaton 2010, 449).

Here's another, from the Jamil Poverty Action Lab (J-PAL), which Duflo and other MIT economists work with: the Deworm the World Movement.

The J-PAL website reports that "Research by J-PAL associates . . . Kremer and . . . Miguel has shown that school-based deworming is one of the most cost-effective methods of improving school participation" (http://www.povertyactionlab.org/scale-ups/school-based-deworming). The Kremer and Miguel study looked at 75 primary schools in Busia, Kenya. Busia, the J-PAL website explains, "is a poor and densely-settled farming region in western Kenya adjacent to Lake Victoria. [It has] some of the country's highest [intestinal worms] infection rates, in part due to the area's proximity to Lake Victoria Kenya." The website goes on: "The evidence from [the Kremer and Miguel] study has helped inform the debate and has contributed to the scale-up of school-based deworming across 26 countries where over 7 million children have been dewormed since 2009." I focus on development and on RCTs. But the problem of using evidence of efficacy from good studies and pilots to predict whether a policy will be effective if implemented is a general one. And it is a megaproblem. It affects us all. This megaproblem, like a good many other problems involving the practice and use of science, is one philosophers of science can contribute to. We are in a position to step in and help, and we should. If we don't step forward to act to improve the decisions that influence all our lives, what is philosophy good for? So let's look at some philosophy that can help. I start with a familiar philosophical concern.

2. Let's Get Straight What We Are Talking About. RCTs, proponents argue, are the 'gold standard' for warranting causal claims. But there's startlingly little attention to what these claims claim. In particular, there's widespread conflation of three distinct kinds of causal claims. RCTs are especially good only for the first: (1) It works somewhere. (2) It works in general. (3) It will work for us. Here's a typical example from a paper by Duflo and Kremer (2005, 205). Already in line 5, in one single sentence, all three kinds of claims are mixed together without note: "The benefits of knowing which programs work . . . extend far beyond any program or agency, and credible impact evaluations . . . can offer reliable guidance to international organizations, governments, donors, and . . . NGO's beyond national borders." I take it from the language and use that they mean as follows:

- Which programs work = It works in general.
- Impact evaluation = It works somewhere.
- Reliable guidance = It will work for us.

I focus on these three kinds of causal claims because I endorse evidence-based policy and I want to improve policy outcomes by the use of evidence. The first—it works somewhere—is where we are encouraged by evidence-based policy guidelines to start. These are the kinds of claims that our best scientific study designs can clinch. The third is where we want to end up: the proposed

program will produce the desired outcome in the target situation as it will be implemented there. The middle—'general' causal claims—is the central route by which 'It works somewhere' can make for evidence that it will work for us. But the road from 'It works somewhere' to 'It will work for us' is often long and tortuous. There are four essential materials for building a passage across:

1. *Roman laws*. I call them this on account of Luke 2:1: "And it came to pass in those days, that there went out a decree from Caesar Augustus, that all the world should be taxed." The laws involved need not be really universal. But they must be wide enough to cover both the evidence and the prediction the evidence is evidence for.
2. *The right support team*. We need all those factors without which the policy variable cannot act.
3. *Straight, sturdy ladders*. So you can climb up and down across levels of abstraction without mishap.
4. *Unbroken bridges*. By which the influence of the cause can travel to the effect.

You must have all four; if any one is missing, you can't get there from here.

3. What's an RCT and What's It Good For? I would hope to stay away from formulas in an address like this, but we do need some technical results to get started. An ideal RCT for cause X and outcome Y randomly assigns individual participants in the study, $\{u_i\}$, into two groups, where $X = x$ universally in the treatment group and $X = x' \neq x$ universally in the control group. No relevant differences are to obtain in the two groups other than X and its downstream effects. The standard result measures the average 'treatment effect' across the units in the study: so T average is the average of Y in the treatment group minus its average in the control group. Of what interest is this strange statistic about randomized units in a study group?

Supposing that Y values for the units in the study are determined by a causal principle that governs the study population, the RCT can reveal something about the role of X in this principle. Without significant loss of generality we can assume that the principles governing Y look like this:[1]

$$L\colon\ Y(u)c = \alpha(u) + \beta(u)X(u) + W(u),$$

where W represents the net contribution of causes that act additively in addition to X and where X may not play a role in the equation at all if $\beta = 0$. So doing a little algebra (and letting $\langle\Phi\rangle$ represent the expectation of Φ),

1. The important lessons follow equally for more complicated functional forms.

$$\begin{aligned}\langle T\rangle &=_{df} \langle Y(u)/X(u)=x\rangle - \langle Y(u)/X(u)=x'\rangle\\ &= \langle \alpha(u)/X(u)=x\rangle - \langle \alpha(u)/X(u)=x'\rangle\\ &\quad + \langle \beta(u)/X(u)=x\rangle x - \langle \beta/X(u)=x'\rangle x'\\ &\quad + \langle W(u)/X(u)=x\rangle - \langle W(u)/X(u)=x'\rangle.\end{aligned}$$

Suppose, as is hoped, that the random assignment of u's to x and x' implies that for u's in the study, X is probabilistically independent of α, β, and W. Then

$$T = \langle \beta(u)\rangle (x - x').$$

Recall L: $Y(u)c = \alpha(u) + \beta(u)X(u) + W(u)$. So $T \neq 0 \rightarrow X$ is a contributing cause for Y in L.

You don't really need to follow the details here; just note the bottom line: If the standard assumptions for an ideal RCT are met, the average treatment effect is the difference in X between treatment and control times β average. So if the average treatment effect is positive, then β is too, in which case X genuinely appears as a cause for Y in law L. This, however, provides no evidence that X will produce a positive difference in the target unless the target and the study share L.[2] Law L must be general to at least that extent. But the stretch of L is in no way addressed in the RCT, and for the most part generality cannot be taken for granted. That's because the kinds of causal principles relevant for policy effectiveness are both *local* and *fragile*.

4. Roman Laws Are Not All That Easy to Come By. The causal laws we rely on for reliable predictions in real policy, real technology, and real experimental settings are *local*. They are local because they depend on the mechanism or the social organization, what I have called the 'socioeconomic machine' that gives rise to them (see Cartwright 1989). Economists know about this kind of locality. The Chicago School notoriously used it as an argument against government intervention: the causal principles that governments have to hand to predict the effects of their interventions are not universal. They arise from an underlying arrangement of individual preferences, habits, and technology and are tied to these arrangements. Worse, according to the Chicago School, these principles are *fragile*. When governments try to manipulate the causes in them to bring about the effects expected, they are likely to alter the underlying arrangements responsible for those principles in the first place, so the principles no longer obtain (see Lucas 1976).

2. Or at least share the important feature of L that X genuinely appears in it.

British econometrician Sir David Hendry urges the use of simple 'quick catch-up' models for forecasting rather than more realistic causal models because the world Hendry lives in is so fluid that yesterday's accurate causal model will not be true today (see Hendry and Mizon 2011). J. S. Mill does too. Economics cannot be an inductive science, he argued, because underlying arrangements are too shaky; there's little reason to expect that a principle observed to hold somewhere sometime will hold elsewhere or later because there's no guarantee the underlying arrangement of basic causes will be the same (see Mill 1836/1967; 1843/1850, bk. VI).

Because so many of the causal principles we employ are tied to causal structures that underpin them, you can't just take a causal principle that applies here, no matter how sure you are of it, and suppose it will apply there. After all, common causal structures are not all that typical, even in the limited and highly controlled world of structures we engineer. Consider for instance these three toasters I found on sale in Oxford: the Cuisinart Classic four-slice at £41.46, the Krups expert black and stainless steel at £44.99, and the Dualit three-slice stainless steel at £158.03. Even these three toasters—man-made and for the same job—do not have the same structure inside. (Or at least we hope not given the big price differential.)

Perhaps you think—as many other economists and medical RCT advocates seem to—that the different populations you study, here and there, are more likely to share causal structure than are toasters. That's fine. But to be licensed in that assumption in any given case, you better be able to produce good evidence for it.

Simple induction is no more warranted here than anywhere else. It requires stable principles, and stable principles require stable substructures to support them. Without at least enough theory to understand the conditions for stability, induction is entirely hit or miss. This I take it is a key point of Princeton economist Angus Deaton's British Academy Keynes lecture in economics. He says of RCTs that they are "unlikely to recover quantities that are useful for policy or understanding. Following Cartwright . . . I argue that evidence from randomized controlled trials has no special priority. . . . The analysis of projects needs to be refocused towards the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work. . . . Thirty years of project evaluation in sociology, education and criminology was largely unsuccessful because it focused on *whether* projects work instead of on *why* they work."[3] Moving on, let's suppose though that (1) there are causal principles that enable *X* to produce *Y* in the study, (2) these are shared in the target, and (3) contrary

3. Read at the academy, October 9, 2008, and published, in a revised form, as Deaton (2009).

to expectations from the Chicago School of economics, these principles will be unaffected if the proposed policy is implemented in the target. There are still three central problems for the prediction that the policy will work in the new setting. The next problem concerns the *support team* necessary if X is to produce a contribution to Y.

5. Support Teams. Return to the abstract form L for the causal law that, for purposes of argument, we are now taking to be shared between study and target situations:

$$L\colon\ Y(u)c = \alpha(u) + \beta(u)X(u) + W(u).$$

The RCT tells about β. It is tempting to think of β as a constant or as an undecomposable random variable. But it isn't. And this despite the fact that you can find it treated thus in sundry works in our field (maybe not from A to Z but at least from Cartwright to Woodward). The difference depends on the kinds of factors that the variables represent. When I write β as a constant or a random variable, I assume that X represents a full, not a partial, cause. But most policy variables represent only partial causes—INUS causes, extending J. L. Mackie's (1965) sense to multivalued variables:

> X is an INUS contributor to Y: X is an *i*nsufficient but *n*onredundant part of a complex of factors that are *u*nnecessary but together *s*ufficient to produce a contribution to Y.[4]

What matters here is that policy variables are rarely sufficient to produce a contribution; they need an appropriate support team if they are to act at all. The support factors are represented by β.[5] And the values of these factors can be expected to vary across the units just as the values of X and W vary.

This is well known in philosophy and in social science. Nevertheless the consequences are frequently ignored. Consider for example the usual advice in the evidence-based policy literature about how to grade policy proposals on the basis of evidence. The US Department of Education explains that what you need are successful RCTs in two or more typical school settings, including "school settings similar to yours" (2003, 10). And the Scottish In-

4. 'Contributions' are, at least as I make sense of them, defined relative to a metaphysics of capacities, other contributions, and laws of composition. In a law of form L, each separate additive term on the right-hand side represents a contribution. See Cartwright (2009).

5. In this case we are supposing that the size of the contribution of X to Y is fixed once the values of the 'helping factors' are set. But this contribution could still vary arbitrarily from unit to unit. It would be more usual, though, to suppose that a full set of helping factors would at least fix the probability for a contribution of a given size.

tercollegiate Guidelines Network, used to help set best practice for the UK National Health Service, provides an A grade to a policy if it is supported by "at least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population" (2011, 51). This advice is vague, surprisingly so given how specific the guidelines are in assessing RCTs, meta-analyses, and systematic reviews. Moreover, if properly spelled out, it is hard to follow. Worst, it is generally bad advice.

Start with hard to follow and consider a paper by a team of authors from Chicago, Harvard, and Brookings, "What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?" (Ludwig et al. 2008). The paper explicitly addresses the question of where outside the experimental population we are entitled to suppose the experimental results will obtain. The authors first report "MTO defined its eligible sample as . . . ." I won't read their long list because I am about to cite it in their conclusion: "Thus MTO data . . . are strictly informative only about this population subset—people residing in high-rise public housing . . . in the mid-1990s, who were at least somewhat interested in moving and sufficiently organized to take note of the opportunity and complete an application. The MTO results should only be extrapolated to other populations if the other families, their residential environments, and their motivations for moving are similar to those of the MTO population" (154–55). The list is a potpourri. It seems as if they have tossed in everything they can think of that might matter without any systematic grounds; why, for instance, did they leave out the geographical location of the cities in the experiment? And anyway, the list gets at what's necessary indirectly. Look again at β in principle L and in the treatment effect:

$$L:\ Y(u)c = \alpha(u) + \beta(u)X(u) + W(u),$$
$$\langle T\rangle = \langle\beta(u)\rangle(x - x').$$

The term β represents in one fell swoop all the different supporting factors necessary if X is to contribute to Y. Each separate combination of values of these factors corresponds to a different value of β. The average treatment effect depends on the average of these values across the study population. That means we suppose that each different arrangement of values of the supporting factors represented by a different value, b, of β appears in that population with a specific probability: $\text{Prob}_{SP}(\beta = b)$.

So, supposing L obtains in both the study and target populations, when can we expect $\langle\beta(u)\rangle$ to be the same? Exactly when $\text{Prob}_{SP}(\beta = b) = \text{Prob}_{TP}(\beta = b)$ for all b's, that is, when all the combinations of values of the supporting factors have the same probability in the study and target popu-

lations. Otherwise it is an accident of the numbers. I expect that the distributions in the study population are rarely duplicated in other populations.

Independent of that, the list in the MTO article does not seem to be a list of supporting factors. Perhaps the hope is that the list includes sufficient 'indicator' factors to ensure that populations that share these indicators will have the same probability distributions over β. Maybe sometimes this is the best we can do. But if we resort to it, we need some defense of why the indicators might be up to the job. And this will be hard to provide without explicit discussion of what the supporting factors might be.

Suppose, though, we solve the problems of identifying these factors. Still advice like that of the Department of Education is wasteful. The treatment effect averages over arrangements for the supporting factors. Some of these arrangements enable X to make a big contribution, others only a small contribution, and for others X may even be counterproductive. We shouldn't aim for the *same* mix of these arrangements as in the study population but rather for a *good* mix—a mix that concentrates on arrangements that allow X to do the most for us.

I am not alone in this view. In 1983 Edward Leamer wrote a classic paper, "Taking the Con out of Econometrics." The symposium discussing the Angrist and Pischke paper was called "Con out of Economics." Leamer's contribution to that symposium makes the same point about supporting factors I have long argued. Here are Leamer's words:

> With interactive confounders [my 'supporting factors'] explicitly included, the overall treatment effect [our $\langle\beta\rangle$] is not a number but a variable that depends on the confounding effects. . . . If little thought has gone into identifying these possible confounders, it seems probable that little thought will be given to the limited applicability of the results in other settings. (Leamer 2010, 35–36)

> [This] is a little like the lawyer who explained that when he was a young man he lost many cases he should have won but as he grew older he won many that he should have lost, so that on the average justice was done. (35)

For a final example of sensitivity to supporting factors, return to the integrated nutrition program. The need for getting the requisite supporting factors into place was not ignored in either Tamil Nadu or Bangladesh. One of the central ideas of the nutrition program was that better nutrition can be secured with meager resources, but to do so, mothers need to know what makes for good nutrition. However, nobody expects that education is enough by itself. You can't feed children better if you can't feed them at all. So the edu-

cational program for mothers was coupled with a supplemental feeding program. Nevertheless the results were disappointing. To see what is supposed to have gone wrong, despite the presence of a good support team, turn to my third problem: *ladders*.

6. Ladders. I am a pluralist and a particularist, inclined to suspect that everything is different. Economists are often more homogenizing (though not Hendry and Mill). They believe that they can base their economics on relatively Roman laws. We are, they argue, really much the same at base, governed by the same motivations and the same laws of human nature. Gary Becker is a notorious limiting case. Becker won the Nobel Prize for modeling great swathes of what we do in day-to-day life under the principles of market equilibrium and rational choice theory, from drug addiction to racial discrimination to crime and family relations. Basically, Becker supposes that the agents he models act so as to maximize their expected utility. The trick is to prescribe just what in the case under study utility consists in, which can include anything from financial gains to inconvenience to serious illness or the joys of watching your spouse consume. As you will see, I shall call this 'climbing down the ladder of abstraction'. Note that in Becker's cases this enterprise is relatively unconstrained, so the accounts are unfalsifiable, which many of us still take to be a damning charge. As economist Robert Pollak argues, "The devil is in the details" (2003, 120).

Angrist and Pischke seem to have an optimistic view about breadth: "anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future" (2010, 23). As I remarked, I am suspicious about principles of behavior that are supposed to apply almost across the board. But that is not the source of my worries about ladders. After all, even though the specific causal principles describing the functioning of the Cuisinart, the Dualit, and the Krups toasters are all different, still I agree that there are a set of even more basic principles that all three share. Even assuming shared principles and laying aside worries about falsifiability, trouble looms: There may be a set of laws that enable X to be a contributing cause to Y in the study and these laws may be shared with the target, but in the target they do not connect X and Y. That's because what counts as a realization of a given factor in the study often cannot do so in the target.

This problem arises because of the way properties at different levels of abstraction piggyback on one another. To use vocabulary familiar from another problem area, abstract features are generally multiply realizable at the concrete level, but the abstract does not supervene on the concrete. The causes in a causal principle can be more or less abstract; because of the piggybacking, principles involving factors at different levels can all obtain at once. On a sphere, 'The trajectories of bodies moving subject only to in-

ertia are great circles' is true; so too is 'The trajectories of bodies moving subject only to inertia are geodesics (i.e., the shortest distance between two points)'. They are equally true because on a sphere, being a great circle is to be a geodesic.[6] For spheres there's a 'ladder' down from the abstract 'geodesic' to the more concrete 'great circle', but there is no such ladder for Euclidean surfaces.

Generally the higher the level of abstraction of a causal principle, the more widely it is shared across populations. Bodies on Euclidean planes subject only to inertia follow geodesics but not great circles. And the lower the level, the more likely that the principle is only locally true. This can make serious problems when it comes to the stretch of the principles that RCTs can establish. The Bangladesh nutrition program provides a vivid example.

There was good evidence that the integrated nutrition program had worked in 20,000 Indian villages. But it failed on average in Bangladeshi sites. Looking at the standard account of what went wrong, we will see that issues about levels of abstraction were at the heart. Nothing in this account supposes that Bangladeshis and Indians are altogether different. On the contrary, it seems likely they share a common principle that allowed the program to improve children's nutrition in India. But this principle couldn't do the same job in Bangladesh because things in Bangladesh just aren't what they are in India.

I imagine those who adopted the program in Bangladesh expected Bangladesh and India to share a simple, commonsense principle:

> **Principle 1**. Better nutritional knowledge in mothers plus food supplied by the project for supplemental feeding improves the nutritional status of their children.

But they did not. The first reason for the lack of impact in Bangladesh, it seems, was 'leakage': The food supplied by the project was often not used as a supplement but as a substitute, with the usual food allocation for that child passing to another member of the family (Save the Children 2003). The principle 'Better nutritional knowledge in mothers plus food supplied by the project for supplemental feeding improves children's nutrition' was true in the original successful cases but not in Bangladesh. This suggests that a better shot at a shared principle would be:

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

> **Principle 2**. Better nutritional knowledge in mothers plus supplemental feeding of children improves children's nutrition.

This is a principle about features at a higher level of abstraction than those in the first principle. In the successful cases in India the more concrete feature 'food supplied by the project' constituted the more abstract feature 'supplemental feeding'. But not in Bangladesh. There the ladders are missing that connect the abstract features in the shared principles with the concrete features offered by the program.

A second major reason for the lack of positive impact is also a problem with connecting ladders between the abstract and the concrete. It's labeled 'the mother-in-law factor' by Howard White, who also points out what I call 'the man factor': "The program targeted the mothers of young children. But mothers are frequently not the decision makers . . . with respect to the health and nutrition of their children. For a start, women do not go to market in rural Bangladesh; it is men who do the shopping. And for women in joint households—meaning they live with their mother-in-law—as a sizeable minority do, then the mother-in-law heads the women's domain. Indeed, project participation rates are significantly lower for women living with their mother-in-law in more conservative parts of the country" (2009, 6). This suggests yet another proposal for a shared principle:

> **Principle 3**. Better nutritional knowledge results in better nutrition for a child in those who (*a*) provide the child with supplemental feeding, (*b*) control what food is procured, (*c*) control how food gets dispensed, and (*d*) hold the child's interests as central in performing *b* and *c*.

Just as the food supplied by the project did not count as supplemental feeding in the Bangladesh program, mothers in that program did not in general satisfy the more abstract descriptions in points *b* and *c*.

The all-too-common fact that things in one setting may not be what they are in another makes real trouble for the use of RCTs as evidence. The previous successes of the program in India are relevant to predictions about the Bangladesh program only relative to the vertical identification of mothers with the more abstract features in points *b*, *c*, and *d*. But not all of these identifications hold. So the previous successes are not evidentially relevant.

7. Roman Laws, Ladders, and Structural Parameters. The lesson of BINP is that the way abstract and concrete features relate implies that (1) in different contexts the same isn't always the same, and (2) this limits the usefulness of claims of 'it works somewhere' for predicting 'it will work for us'. But the very same facts about the relations between the abstract and the concrete equally imply: (1′) In different contexts very different things can be the

same. And because of this, (2′) claims of 'it works somewhere' can support policy predictions in contexts far away and very different from the study populations that warrant them. Angrist and Pischke employ this in their commendation of RCTs. "Small ball sometimes wins big games," they tell us (2010, 25). How so? Because sometimes from RCTS, they urge, you can learn 'structural econometric parameters', where following David Hendry, "Structure . . . is defined as the set of basic features of the economy which are invariant to [various specific] changes in that economy," including "an extension of the sample" (Hendry and Mizon 2010, 1–2). How wide an extension? That depends on the theory. For the moment let us assume, wide enough at least to cover the policy target.

Suppose that in the study a *structural* law of form *L* allows *X* to cause *Y*. Then β from that law is a structural parameter. Because β is a structural parameter, $\beta \neq 0$ in the study population shows that it's unequal to 0 in extensions of the population. This line of reasoning is familiar. Because the gravitational constant *G* is a structural parameter, Galileo can measure it on balls rolling down inclined planes and Euler a century later can put the same *G* into formulas calculating the 'true curve' of cannonballs that are subject to the buoyant and resistant forces of the air as well as to gravity.

The parameter discussed by Angrist and Pischke is the "intertemporal [labor supply] substitution elasticity" (2010, 4), that is, a parameter that represents how much transitory wage changes contribute to hours of work a worker supplies. This is a theoretical parameter in, for example, life cycle theory. Is it constant enough for Angrist and Pischke to play the Galileo-Euler game? Maybe, maybe not. As Angus Deaton remarked in a private conversation, "Structural parameters are in the eye of the beholder." Or according to Mervyn King, governor of the Bank of England (in a paper read at the Royal Society, March 22, 2010), "There are probably few genuinely 'deep' (and therefore stable) parameters or relationships in economics."

I don't know if the labor supply elasticity is a structural parameter or how far the structure stretches if it is. But Angrist and Pischke must take it that way. Here is the longer passage from which I quoted before: "Small ball sometimes wins big games. In our field, some of the best research designs used to estimate labor supply elasticities exploit natural and experimenter-induced variation in specific labor markets. Oettinger . . . analyzes stadium vendors' reaction to wage changes driven by changes in attendance, while Fehr and Goette . . . study bicycle messengers in Zurich who, in a controlled experiment, received higher commission rates for one month only" (2010, 25). Oettinger's (1999) analysis of stadium vendors at major league baseball games supposes that the vendors' expectations about the size of the crowd constitute their wage expectations, and in turn their wage expectations constitute 'laborers' wage expectations' in this case. Similarly, the number of vendors constitutes the labor supply in this case. So Angrist and Pischke

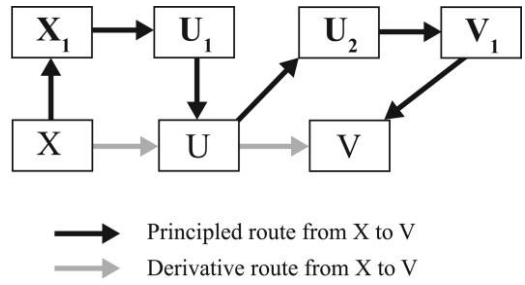

Figure 1. Two routes from a cause to an effect, at different levels of abstraction.

seem to assume that labor supply elasticity is a structural parameter and that the parameter connecting vendors' expectations of crowd size with the number of vendors showing up at the stadium is the labor supply elasticity in this situation.

What warrants these two assumptions? We confront here the twin problems of Roman laws and warranted ladders. For the first, it is usually theory that teaches that there is a structural parameter, but it had best be a credible well-supported theory. As to the second, we need help in both climbing up the ladder of abstraction in the study situation and then, in new settings, climbing down. How do we know that what Oettinger measured on his stadium vendors was an instantiation of the labor supply parameter? And when we turn to a new situation with this parameter in hand, how do we figure what concrete features count as labor supply elasticity there? Theory can help. But it will also take sound knowledge of the local context. The point is that studies like Galileo's and Oettinger's—and RCTs—can measure structural parameters, but they cannot tell us that there is a structural parameter to be measured. That information must come from elsewhere.

8. Unbroken Bridges. My final problem involves *causal chains*. Generally getting from cause to effect is not a one-step process. Rather the policy variable is at the head of a causal chain with the hoped-for outcome at the tail, with a number of links in between. Policy X causes outcome Y in the study situation because X causes U which causes V which causes W which causes . . . which causes Y. We can expect X to cause Y in a different situation only as long as the chain is unbroken.

Consider figure 1 and look at the first step. What enables X to cause U? I have been arguing that it is often not because of a general principle connecting X and U but rather because X and U are concretizations of features $\mathbf{X}_1$ and $\mathbf{U}_1$ at a higher level of abstraction, where $\mathbf{X}_1$ and $\mathbf{U}_1$ are joined by a reasonably general principle. Similarly, U may cause V not because of a principle connecting U and V but rather because of a general principle between more

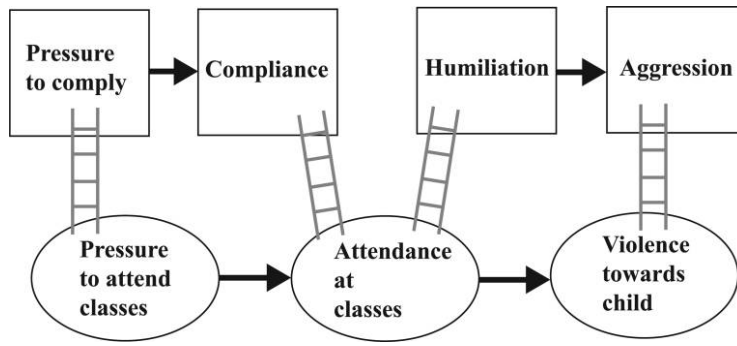

Figure 2. Example of a "broken bridge."

abstract features $\mathbf{U}_2$ and $\mathbf{V}_1$ that they instantiate. Note the new subscripts. There is no reason that the very same features under which *U* is the effect of *X* should be the feature in virtue of which *U* is the cause of *V*.

Let me illustrate with a possible example social workers have been worrying about from UK child welfare policy in which a child's caregivers are heavily encouraged, perhaps badgered, into attending parenting classes. It is illustrated in figure 2.

Consider making fathers attend parenting classes. Different cultures in the United Kingdom have widely different views about the roles fathers should play in parenting. Compelling fathers to attend parenting classes can instantiate the more abstract feature, 'ensuring caregivers are better informed about ways to help the child', in which case it can be expected to be positively effective for improving a child's welfare. But it may also instantiate the more abstract feature 'public humiliation', in which case it could act oppositely. Attending classes as a result of pressure can constitute a public humiliation and by virtue of being a public humiliation can lead to aggressive and violent behavior, which may be directed toward the child. There is then no unbroken bridge at the level of the more widely applicable principle, but there is a linked-up sequence at the more concrete level.

This of course has mixed policy implications. If we found that pressing fathers to attend parenting classes in this cultural group led to negative outcomes, that would not mean it should be expected to do so in other groups. The general principles that affect the different populations may be the same, but they don't make an unbroken bridge for the negative effects to move along. However, getting positive results in other groups in which the humiliation mechanism is not activated does not tell us what will be the overall outcome where it is activated. This is yet another case in which knowing that a policy works—or fails—somewhere is at best a starting point for figuring out if it will work for us.

9. Conclusion. We can do better at predicting policy effectiveness. And philosophy helps show how. RCTs can help too, as their advocates maintain. But, as I have argued, it is a long and tortuous road from learning that a policy works somewhere, which is the kind of claim an RCT can clinch, to correctly predicting that it will—or won't—work for you. And you can go wrong in both directions: accepting programs that won't work for you, as Levy claims has repeatedly happened with Progresa, and rejecting ones that would, like the J-PAL rejection of textbooks in favor of deworming, or in my hypothesized example, sending caregivers who won't feel humiliated to parenting classes.

I've rehearsed four essential materials it takes to secure a safe pathway: (1) shared laws, (2) supports, (3) ladders, and (4) laws that interlock. No matter how secure the starting point, if any one of these is missing, you just can't get there from here.

I don't need to remind you that a conclusion is only as secure as its weakest premise. RCTs may be the gold standard for underpinning the start point, but you can't pave the road in between with gold bricks. Evidence for these other factors is necessarily different and varied in form: theory, big and little, consilience of inductions, and a great deal of local information about study and target situations. Philosophy matters because once you know what you need, you can hunt for it. And often you can find it. Here is Howard White again: "In the Bangladesh case, identification of the 'mother-in-law' effect came from reading anthropological literature" (2009, 15). But to find it you must be encouraged to look. And where it doesn't exist, the sciences must be encouraged to uncover it. It's no good just putting all your money into gold bricks.

We philosophers of science are faced then with a hard job. Here as elsewhere in the natural and social sciences, in policy, and in technology, we can help. But to do so we need to figure out how better to engage with scientific practice and not just with each other.

REFERENCES

Angrist, Joshua, and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Banerjee, Abhijit, and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1:151–78.

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

———. 2009. "Causal Laws, Policy Predictions and the Need for Genuine Powers." In *Dispositions and Causes*, ed. Toby Handfield, 127–58. Oxford: Oxford University Press.

Deaton, Angus. 2009. "Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development." *Proceedings of the British Academy* 162: 123–60.

———. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48:424–55.

Duflo, Esther, and Michael Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, ed. George Pitman, Osvaldo Feinstein, and Gregory Ingram, 205–32. New Brunswick, NJ: Transaction.

Hendry, David, and Grayham Mizon. 2010. "Econometric Modelling of Changing Time Series." Discussion paper series, Oxford University.

———. 2011. "What Needs Rethinking in Macroeconomics?" *Global Policy* 2:176–83.

Lancet. 2004. "The World Bank Is Finally Embracing Science." *Lancet* 364:731–32.

Leamer, Edward. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24 (2): 31–46.

Lucas, Robert. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, ed. Karl Brunner and Allan Meltzer. Amsterdam: North-Holland.

Ludwig, Jens, Jeffrey B. Liebman, Jeffrey R. Kling, Greg J. Duncan, Lawrence F. Katz, Ronald C. Kessler, and Lisa Sanbonmatus. 2008. "What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?" *American Journal of Sociology* 114:144–88.

Mackie, John Leslie. 1965. "Causes and Conditions." *American Philosophical Quarterly* 2:245–64.

Mill, John Stuart. 1836/1967. "On the Definition of Political Economy and on the Method of Philosophical Investigation in That Science." In *Collected Works of John Stuart Mill*, vol. 4. Toronto: University of Toronto Press.

———. 1843/1850. *A System of Logic*. Repr. New York: Harper.

Oettinger, Gerald. 1999. "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors." *Journal of Political Economy* 107:360–92.

Pollak, Robert. 2003. "Gary Becker's Contributions to Family and Household Economics." *Review of Economics of the Household* 1:111–41.

Save the Children. 2003. *Thin in the Ground: Questioning the Evidence behind World Bank–Funded Community Nutrition Projects in Bangladesh, Ethiopia and Uganda*. London: Save the Children UK.

Scottish Intercollegiate Guidelines Network. 2011. *SIGN 50: A Guideline Developer's Handbook*. Edinburgh: Scottish Intercollegiate Guidelines Network.

US Department of Education. 2003. *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington, DC: Coalition for Evidence-Based Policy.

White, Howard. 2009. "Theory-Based Impact Evaluation: Principles and Practice." Working Paper 3, International Initiative for Impact Evaluation, New Delhi.

World Bank. 1995. *Tamil Nadu and Child Nutrition: A New Assessment*. Washington, DC: World Bank.