



ORDER: GOD'S, MAN'S AND NATURE'S

Will this Policy work for You?

Predicting Effectiveness Better: How Philosophy Helps

Nancy Cartwright

Presidential Address PSA 2010

Support of the Templeton Foundation is gratefully acknowledged.

It is of course a great honour to be giving this address this evening and I thank you.

The World Bank estimates that in developing countries 178 million children under five are stunted in growth and 55 million are underweight for their height.¹ 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), modelled 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 programmes in those areas.

¹ Independent Evaluation Group World Bank. 2001. "Tamil Nadu and Child Nutrition: A New Assessment". The World Bank Group.

The Bangladesh Project was modelled on TINP. But Bangladesh's project had little success. A Save the Children UK assessment concludes that programme areas and non-programme areas still had the same prevalence of malnutrition after 6 years and this despite the fact that the targeted health educational lessons sank in to some extent: Carers in the BINP areas had on the whole greater knowledge about caring practices than those in non-BINP areas.

'Why 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 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 take-over movement fast gaining influence there, a new methodology to improve development outcomes: randomized controlled trials. As a Public Radio International interview reports: "A team of economists at MIT says it's time for a new approach -- one that makes prescriptions for poverty as scientifically-based as prescriptions for disease."² MIT's Esther Duflo is one of the leaders of this movement. She tells us:

- ❑ "The past few years have seen a veritable explosion of randomized experiments in development economics."
- ❑ "Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century..."

Witness also the recent *Journal of Economic Perspectives* symposium on a paper commending RCTs by my LSE colleague Steve Pischke and another MIT economist, Joshua Angrist.³ As one exemplar of good research design they cite:

"...in a pioneering effort to improve child welfare, the Progresá programme in Mexico offered cash transfers to randomly selected mothers, contingent on participation in prenatal care, nutritional monitoring of children, and the

² <http://www.pri.org/theworld/?q=node/10887>

³ [Spring 2010, vol 24, no. 2]

children's regular school attendance.... Progresas is why now thirty countries worldwide have conditional cash transfer programmes." [p 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 Progresas, Santiago Levy says that many of the places that want them are places where they will obviously fail: In some of these countries success would require people to go to clinics that don't exist.⁴

Here's another, from the Jamil Poverty Action Lab (J-PAL), which Duflo and other MIT economists work with: the Deworm the World Movement. J-PAL website reports: '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.' 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 mega problem. It affects us all. This mega problem, 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.

⁴ the JEL version of Deaton's paper

I start with a familiar philosophical concern:

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 3 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: 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."⁵

■ 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 programme will produce the desired outcome a. *in the target situation* b. *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 4 essential materials for building a passage across:

⁵ [Use of Randomization in the Evaluation of Development Effectiveness p 93]

1. Roman laws.

The laws involved need not be 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.

I would hope to stay away from formulae 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 tell us 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:

$$L: Y(u) = \alpha(u) + \beta(u)X(u) + W(u)$$

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 – you don't need to follow the details, 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 beta 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 . 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**.

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. 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.

Or, 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. JS Mill 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 is the same.

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. Even these three toasters – man-made and for the same job – do not have the same structure inside. 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 lecture. 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.”

Moving on. Let's suppose though that

- ☐ there are causal principles that enable X to produce Y in the study
- ☐ these are shared in the target and
- ☐ contrary to expectations from Lucas critique, principles will be unaffected if the proposed policy is implemented in the target.

There are still 3 central problems for the prediction that the policy will work in the new setting.

Return to the abstract form L for a causal law we take to be shared between study and target situations.

$$L: Y(u) = \alpha(u) + \beta(u)X(u) + W(u)$$

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' represent a *full*, not a *partial* cause. But most policy variables represent only partial causes – INUS causes, extending JL Mackie's sense to multivalued variables:

X is an INUS contributor to Y: X is an *insufficient* but *nonredundant* part of a complex of factors that are *unnecessary* but together sufficient to produce a contribution to Y.

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 .⁶ 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 a setting *similar to that of your schools/classrooms*'. And SIGNs, 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...*'

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?' The paper explicitly addresses the question of where outside the experimental population we are entitled to suppose the

⁶ 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.

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-1990's, 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.'⁷

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:

$$L: Y(u) = \alpha(u) + \beta(u)X(u) + W(u)$$

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

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; i.e., when all the combinations of values of the supporting factors have the same probability in the study and target populations. Otherwise it is an accident of the numbers. I expect that the distributions in the study population are rarely duplicated in other populations.

⁷ AJS 114, 144-88. Ludwig et al

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 defence 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. Leamer:

...With interactive confounders [my 'supporting factors'] explicitly included, the overall treatment effect [our β] 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.

... which 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.⁸

For a final example of sensitivity to supporting factors, return to the integrated nutrition programme. The need for getting the requisite supporting factors into place was not ignored in either Tamil Nadu or in Bangladesh. One of the central ideas of the nutrition programme was that better nutrition can be secured with meagre

⁸ P 34,35

resources, but to do so, mothers need to know what makes for good nutrition. On the other hand nobody expects that education is enough by itself. You can't feed children better if you can't feed them at all. So the educational programme for mothers was coupled with a supplemental feeding programme. 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 3rd problem.

I am a pluralist and a particularist, inclined to suspect that everything is different. Economists are often more homogenizing. We are 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 modelling 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. Note that 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." (p 9)

Pischke and Angrist 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. [p23]

As I remarked, I am suspicious about principles of behaviour 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 cause Y in the study and these laws may be shared with the target but in the target they do connect X and Y. That's because what counts as a realization of a given factor in the study often can not 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 to philosophers, 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 inertia 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.⁹ For spheres there’s a ladder 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 programme provides a vivid example

There was good evidence that the integrated nutrition programme had worked in 20,000 Indian villages. But it failed on average in Bangladesh 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 programme 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 programme in Bangladesh expected Bangladesh and India to share a simple, common-sense principle:

⁹ I shall here be relatively cavalier about the metaphysics of properties. I treat abstract features and concrete ones both as real and I treat them as different features even if having one of these (the more concrete feature) is what constitutes having the more abstract one on any occasion. I take it that claims like this can be rendered appropriately, though probably differently, in various different metaphysical accounts of properties.

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

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 programme.

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 labelled '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. (White 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 programme, mothers in that programme did not in general satisfy the more abstract descriptions in 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 programme in India are relevant to predictions about the Bangladesh programme only *relative to* the vertical identification of mothers with the more abstract features in b., c., and d. But not all of these identifications hold. So the previous successes are not evidentially relevant.

Roman laws, ladders and structural parameters

The lesson of BINP is that the way abstract and concrete features relate implies

- 1. In different contexts the same isn't always the same.

And

- 2. This limits the usefulness of it-works-somewhere claims 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.' It-works-somewhere claims can support policy predictions in contexts far away and very different from the study populations that warrant them.

Pishke and Angrist employ this in their commendation of RCTs. 'Small ball sometimes wins big games,' they tell us. 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’.¹⁰

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 formulae 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 Pishke is the ‘...intertemporal labour supply substitution elasticity’; 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 Pishke to play the Galileo-Euler game?

Maybe, maybe not. As Angus Deaton remarks, ‘Structural parameters are in the eye of the beholder.’ Or see Mervyn King, Governor of the Bank of England: ‘There are probably few genuinely “deep” (and therefore stable) parameters or relationships in economics’¹¹

I don’t know if the labour supply elasticity is a structural parameter nor how far the structure stretches if it is. But Pischke and Angrist 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

¹⁰ Hendry and Grayham Mizon, “Econometric Modelling of Changing Time Series”, vol in honour of Svend Hylleberg

¹¹ Royal Society March 2010

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.

Oettinger's analysis of stadium vendors at major-league baseball games¹² supposes that the vendors' expectations about the size of the crowd constitute their wage expectations and in turn their wage expectations constitute 'labourers' wage expectations' in this case. Similarly the number of vendors constitutes the labour supply in this case. So Angrist and Pishke seem to assume that labour supply elasticity is a structural parameter and that the parameter connecting vendor's expectations of crowd size with number of vendors showing up at the stadium IS the labour 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 credible well-supported theory. As to the second, we need help both in climbing up the ladder of abstraction in the study situation; then in new settings, in climbing down. How do we know that what Oettinger measured on his stadium vendors was an instantiation of the labour supply parameter? And when we turn to a new situation with this parameter in hand, how do we figure what concrete features count as labour 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 can not tell us that that there is a parameter to be measured. That information must come from elsewhere.

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... ...which causes Y. We can expect X to cause Y in a different situation only so long as the chain is unbroken.

¹² Journal Of Political Economy 1999, 107(2), 360-392

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 X_1 and U_1 at a higher level of abstraction where X_1 and U_1 are joined by 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 abstract features U_2 and V_2 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 where a child's care-givers are heavily encouraged, perhaps badgered, into attending parenting classes. Consider making fathers attend parenting classes. Different cultures in the UK 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 care-givers 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 behaviour, which may be directed towards the child. There is then no unbroken bridge at the level of 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. On the other hand, getting positive results in other groups where 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 where knowing that a policy works – or fails – somewhere is at best a starting point for figuring out if it will work for us.

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 programmes that won't work for you, as Levy claims has repeatedly happened with Progresá, and rejecting ones that would, like the J-PAL rejection of textbooks in favour 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.
4. Laws that interlock.

No matter how secure the starting point, if any one of these is missing then 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 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....' 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 technology, we can help. But to do so we need somehow to figure out how better to engage with scientific practice and not just with each other.

Thank you.

