Durham University

Durham Research Online

Deposited in DRO:

15 April 2016

Version of attached file:

Accepted Version

Peer-review status of attached file:

Peer-reviewed

Citation for published item:

Cartwright, Nancy D. (2009) 'What is this thing called 'efficacy'?', in Philosophy of the social sciences : philosophical theory and scientific practice. Cambridge : Cambridge University Press, pp. 185-206.

Further information on publisher's website:

http://www.cambridge.org/gb/knowledge/isbn/item2427693/Philosophy%20of%20the%20Social%20Sciences/?site$_locale = en_GB$

Publisher's copyright statement:

This material has been published in Philosophy of the social sciences : philosophical theory and scientific practice edited by C. Mantzavinos. This version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. © Cambridge University Press 2009

Additional information:

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full DRO policy for further details.

Durham University Library, Stockton Road, Durham DH1 3LY, United Kingdom
Tel : +44 (0)191 334 3042 | Fax : +44 (0)191 334 2971
https://dro.dur.ac.uk

What is This Thing Called 'Efficacy'?[1]

1. The Topic

This paper is about efficacy, effectiveness, the need for theory to join the two, and the tragedies of exporting the Cochrane medical-inspired ideology to social policy. Loosely, *efficacy* is what is established about causes in RCTs – randomized controlled trials. *Effectiveness* is what a cause does 'in the field'. The theory, like that describing forces in mechanics, underwrites the assumption that the cause contributes the same effect in the field as in the experiment. The tragedies are multiple and snowball on one another. On conventional Cochrane Collaboration doctrine, following the model of testing pharmaceuticals, the RCT is the gold standard for evidence of effectiveness in evidence-based policy. The first tragedy is that on dominant characterizations of 'efficacy', including, especially, many that try hard to be scientific, it does not make sense to suppose that efficacies make any difference outside experiments. The second tragedy is that once 'efficacy' is characterized so that it does make sense, the RCT can hardly be a gold standard since it goes no way towards establishing the theory, or more loosely the story or account, that it takes to get out of the experiment and into the field. The third tragedy is that much of the teaching about evidence-based policy pays little attention to the need for such theories or accounts. Indeed there is often the suggestion that RCTs should replace such accounts since the accounts are almost always controversial. The worst tragedy is that we offer advice that lets policy down, wasting the powerful knowledge that could be provided by RCTs. We pay heavily to measure efficacies in RCTs but efficacy is not evidence at all for effectiveness without the right kind of account or theory to make it so. Yet we provide hardly any guidance about how to manage when accounts are dicey.

2. Why it Matters

Evidence-based or evidence-supported policy is all the rage now, mandated throughout the U.S. and the U.K. and increasingly in Europe, at the international, national and local levels. But mandates need policing and policing here requires a theory of evidence. To know whether a genuine attempt has been made to support

[1] I would like to thank Chris Thompson for his help with the research and production of this paper and the Spencer Foundation for support of the research.

decisions about policy by evidence that the proposed policy will be effective, some notions are required of what effectiveness is and what counts as evidence for it.

Much serious work has been done to supply these, a great deal of which has been inspired by the Cochrane Collaboration, a ‘volunteer’ organization that has been very instrumental in the evidence-based medicine movement. The relatively new cousin of the Cochrane Collaboration, the Campbell Collaboration, takes as one of its chief tasks to provide an account of evidence for evidence-based social science and policy. Much of the current debate in the Campbell Collaboration is about the extent to which the Cochrane methodology can be taken over and especially about the prevalent view that RCTs should be the gold standard, providing the very best evidence in favour of policy and, sometimes it is argued, the only reliable evidence. Consider for example the U.S. Department of Education advice: Good evidence for a new educational policy is two successful RCTs for that policy in ‘typical’ schools that are ‘like yours’.

Part of these efforts are driven by the laudable desire to produce advice that on the one hand is relatively sound and on the other can realistically be expected to be put to use by policy-makers who may wish to do the best they can but are untrained in both natural and social science and in the handling and evaluation of evidence and also are pressed for time and resources. It is understandable in these circumstances if the advice does not meet either goal to a very high standard. A lot of compromise should be expected and satisficing rather than optimizing should be the standard for success.

This paper looks at one standard way of thinking about the use of RCTs as evidence for policy: RCTs establish the efficacy of a cause and that provides one central piece of evidence about how effective the cause will be in live settings outside the experiment. The most standard assumption in practice seems to be that a cause that is efficacious in the experimental setting will be efficacious outside unless its effects are swamped by opposing causes. Consider the widely discussed California class-size reduction programme of 1996 which is generally taken not to have produced the good results hoped for. The general verdict is that well-conducted trials in Tennessee provided good evidence for the efficacy of class-size reduction in improving

academic achievement[2]. What went wrong in California was not that this evidence is faulty. Rather, the programme was rolled out in California rapidly, with little time for schools to prepare. As a result there was a big increase in demand for teachers and for classrooms, a demand that well exceeded supply. Implementation lagged in schools serving minority and low income students, in part because they lacked adequate classroom space. As a consequence most of the unqualified teachers ended up in the schools with the most disadvantaged students.[3]

This argument seems to suppose that class-size reduction is an efficacious cause whose tendency to produce good effects was overwhelmed in California by the actions of causes that operate in the opposite direction. There are other ways of thinking about efficacy, or indeed of thinking about what exactly it is that is established in RCTs and what use can be made of experimental results, some of which I outline in section five. But this is one very standard and widespread way of treating RCTs and their lessons and it is the one I shall concentrate on in this paper. My conclusions are unfortunate. If it is to be used in the way just described, efficacy takes far more evidence than RCTs to establish, and evidence that is very different in kind. What the RCT can do is to measure *how much* efficacy a cause has *given* that there is efficacy there to be measured.

Besides explicitly addressing issues of efficacy, evidence, RCTs and policy, this paper has a subsidiary purpose. Scientists, including social scientists, are often dismissive of philosophy. Philosophy, it is said, is too abstract, too fussy and too taken up with its own problems to matter to real practice. With the issues discussed here I think just the opposite is the case. Bad practice, I maintain, is being recommended without intention and without sufficient notice in part because prissy issues that philosophy fusses about are being ignored, issues like what counts as a proper definition and whether an argument has been laid out with all the necessary premises.

---

[2] Nye, B., Hedges, L.V., Konstantopolous, S., 2000, "The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment", *American Educational Research Journal*, Vol. 37, No.1, pp.123-151

[3] Bohrnstedt, G.W., Stecher, B.M. (eds.), 2002, "What We Have Learned About Class Size Reduction in California", California Department of Education

3. Efficacy and Effectiveness

In 1978 the U.S. Office of Technology Assessment (OTA) made a distinction between efficacy and effectiveness that is widely cited. The definition and the summary statement given in the OTA report are themselves a little at odds with each other but here is the kind of lesson that practitioners take away from it: *Efficacy* is 'the ability of a treatment to produce benefit if applied ideally' and *effectiveness*, 'the benefit that actually occurs when a treatment is used in practice'. (p 317 in G Andrews 1999) How shall we understand 'ideally' here? This author also maintains that 'a good treatment' should 'be shown to be clinically and statistically better than that due to an ineffective or placebo treatment in randomised controlled trials'. (p 316) The OTA itself explained in its 1978 report, 'For efficacy the conditions of use are considered to be ideal, or, as a substitute, experimental research findings'. (p. 16) It is laudable to try to be more precise and to make distinctions that help prevent mistaken inferences, such as the easy slide from experimental results to predictions that just the same results will obtain in the field. Nevertheless it is this characterization of efficacy in terms of experimental results that makes the first set of troubles I want to highlight. The more recent report of 1994, 'Identifying Health Techniques that Work', concurs with the identification of efficacy with experimental results: 'Traditionally, RCTs have been the tool associated with narrowly defined efficacy studies.' (Summary, p 17)

The same kind of characterization of efficacy in terms of experimental conditions is common among careful statisticians, and for good reason. Efficacies are average effects and averages are relative to distributions. But distributions and averages make no sense without reference to the population of units over which the distribution is defined and the chance set-up that is supposed to generate it, and for an average it must be clear whether it is conditional or unconditional and what it is conditional on.

The canonical source here is probably DR Cox's 1958 *The Planning of Experiments*, which defines the notion of 'true treatment effects': These effects are defined as effects in an experiment. One of the most standard ways of defining efficacy nowadays is that championed by Paul Holland and Donald Rubin. Although there have been a variety of sophistications and improvements, the fundamental point I

want to make still holds. Since my point is most easily illustrated with the original characterization, I will stick with that.

In an early careful account Rubin defines the formal cousin of efficacy: 'the average causal effect in P of one treatment relative to another' (where P is a 'population of N experimental units' [p. 36]). This is the difference in *conditional* means for the outcome it would experience given exposure to the first treatment and the outcome a unit would experience given exposure to the second. [p 43]

AP Dawid objects to the fact that the means are taken for subjunctive conditional variables: The values of these two variables cannot possibly both be observed but at most only one or the other.[4] I am sympathetic with his worries but they are not mine here. Rather I want to point out what these means are conditioned on: observed values of treatment outcomes, of covariates, a missing-indicator variable and, what matters for my discussion here, a 'treatment assignment variable' W. The treatment assignment variable picks up two kinds of information. One is 'the treatment assignment mechanism which determines the sample experimental units to receive each treatment' (p. 37) – for instance the random assignment mechanism of the RCT. The other is 'the sampling mechanism which determines the experimental units to be studied' i.e., to be exposed to a treatment. Together these fix whether or not the means concern units in an experimental setting and exactly how the assignment to treatment occurs for units in the experiment. Dawid does a similar thing in his own decision-theory account, which takes means only over actual, not counterfactual, outcomes, for Dawid conditions his means on a treatment assignment variable, F, that can take three values: receives treatment 1 by assignment, receives treatment 2 by assignment, is not assigned but takes values for either treatment naturally; that is, Dawid conditions on a variable that determines whether one is in an experiment or not.

In one sense this is exactly as it should be. We should not be defining means without specifying the intended population, chance set-up and conditioning factors. But in

[4] Dawid, A.P., 2000, 'Causal Inference Without Counterfactuals', *Journal of the American Statistical Association*, Vol. 95, No. 450 , pp. 407-424; and Dawid, A.P., 2007, 'Counterfactuals, Hypotheticals and Potential Responses: a Philosophical Examination of Statistical Causality'. In *Causality and Probability in the Sciences*, Russo, F., Williamson, J. London: College Publications, Texts In Philosophy Series Vol. 5, pp. 503–32.

another, it makes for a very odd sense of 'efficacy'. For the causal effect is now always *defined only relative to a particular experimental design*. That means that it if we take this as our formal definition of the efficacy of a cause, it *does not make sense* to talk about the cause being efficacious in any other setting. The concept only applies to those situations in which the experimental design obtains, and in no others[5].

It is important to note that this is not the familiar problem of external validity. In the problem of external validity we consider a quantity that *could* hold across a variety of situations; we observe the value of that quantity in one situation, like an RCT; and we ask if the very same value of that quantity holds everywhere or in some other specified situation. But with the definitions discussed here, the quantity in question could not possibly take a value anywhere except in the experiment with respect to which it has been defined. By way of analogy, the International Adult Literacy Survey (IALS) is used to measure and compare the functional literacy of adults across the OECD. It is reasonable to ask whether the test of literacy used by the IALS is applicable to populations outside the wealthy countries that make up the OECD. But it is inappropriate to ask whether the IALS can tell us anything about the numeracy rates in different countries.

We can take this problem to be the result of over operationalizing, and that is the approach I shall discuss here. That is, we mistakenly define a quantity that can hold across a variety of different situations in terms of one way of measuring the quantity in one particular setting. The problems that arise from operationalizing are exacerbated by the drive to provide definitions in terms of probabilities so that the techniques of statistics can be brought to bear. The 1978 OTA report for instance asserts (without defence or comment), 'Efficacy is best expressed in probabilities'. (p. 4) But probabilities are always defined relative to populations, chance set-ups and conditioning variables. So definitions expressed in terms of probabilities will always characterize concepts that ipso facto have a very limited range of applicability.

[5] Alternative to choosing a specific population and assignment mechanism for defining the efficacy of a cause C is to define a concept with an open variable "for any population P and assignment mechanism W, the P.W efficacy of C is…" One could then suppose that finding the RCT- efficacy of C in any population P would give one information about the P.W efficacy of C for any target P and W. This alternative raises exactly the same demands as those adumbrated in section 8, for the capacity alternative I focus on.

4. Capacities

So then suppose that we take the approach that RCTs provide evidence for the effectiveness of causes in real-life settings because the RCT provides information about some fact about what a cause can do that can obtain outside the RCT, in the real-life setting. I have always called these facts about what causes do across a broad range of settings, facts about the *capacity* associated with the cause, where capacity is modelled on JS Mill's notion of a *tendency*. Indeed, in the book where I discussed capacities at length (*Nature's Capacities and their Measurement*) one of the prime motivations for positing capacities was to make sense of the highly lauded scientific technique of using the effects that causes produce in very special circumstances – RCTs, ideal Galilean experiments or when all other causes are held fixed – to teach something about what will happen when those causes are present elsewhere. The idea is that

1. Causes have, or have associated with them, relatively enduring capacities
2. We can learn about the natural effects of these capacities in various nice situations (like Galilean experiments, RCTs, or when all other causes are held fixed at one set of values) and
3. What we learn in these very special situations has some systematic relation to the effect that is produced when the cause obtains elsewhere.

This seems to mesh neatly with the talk about efficacy that we frequently see in practice. Recall the California class-size reduction programme. The Tennessee trials provide evidence about how efficacious small class sizes can be on academic achievement; the effect that actually occurred in California was taken to be the resultant of that effect damped down and modified in an intelligible way by other causes, themselves conceived of as being efficacious in opposite or at least different directions.

This leads to the following scenario. The efficacy of a cause T for an outcome O (or the efficacy of T for O relative to alterative T') is a relatively enduring feature associated with T. The magnitude of T's efficacy for O can be measured more or less accurately in a variety of ways and that magnitude contributes in some systematic or

intelligible way to what happens whenever T is present. The ideal RCT is then taken to be a good way to measure aspects of[6] the efficacy of T.

This is a good scenario but philosophers of science and metaphysicians generally do not like it, and these are people who spend their lives thinking about what kinds of categories it makes sense to use to describe reality. I can hardly refer to them as an objection to the hypothesis that efficacy is a capacity though since I have spent considerable effort arguing that their point of view is mistaken. It is only fair to note for practising social scientists, however, that my claims that the notion of capacity makes sense and that it plays an important role throughout the natural, social and biomedical sciences, and especially in their link with technology and policy, is by far a minority view in philosophy. Here I want to raise my own problems with deploying it for evidence-based policy. To do so it will help to rehearse some conventional lessons from philosophy of science, especially in the venerable discussion about the realism and acceptability of scientific concepts, which differs from discussions of construct validity that social scientists may be more familiar with. Before doing so, however, I shall give a very brief review of some other methods on offer for avoiding the problems raised by the kinds of definitions of efficacy described in section 3. This should help us to understand better what thinking in terms of capacities involves and what its advantages are.

5. Alternatives to Capacities

Let me recap the problem. We want to know what facts about a cause can be established in an RCT that could allow positive results in an RCT to be evidence – albeit possibly only partial and defeasible evidence – for the effectiveness of the cause elsewhere. One answer, the one I look at in detail here, is that the RCT can provide information about the size and direction in which a (relatively enduring) capacity operates. Other strategies propose

[6] I say 'aspects of' because one would seldom expect the actual mean differences to be the very effect the cause *contributes* in this capacity sense. At the very least, we know that this difference is an average over the contribution it makes in various subpopulations represented in the population enrolled in the experiment. It is a complicated – and context-dependent – question of exactly what information an RCT yields about the contribution of a cause on the capacity interpretation. I ignore the issue here in order to avoid excess complication.

a. Causal claims and inductions from them. Given some natural assumptions about causality, an increase in probability of an effect, E, in the 'treatment' arm of an ideal RCT over that in the 'placebo' arm shows that for at least some individuals in the population enrolled in the experiment, the treatment, T, *caused* E. The defence of this as a valid conclusion to draw from a successful RCT depends on how one characterizes an ideal RCT and on what assumptions one makes about singular causal claims. This is a long story that I shall not review. What matters for the issues here is how establishing this kind of causal conclusion about T for the population enrolled in the experiment can provide evidence for the effectiveness of T outside the experiment. How does establishing that T causes E in some individuals in the experiment give reason to think that T causes E outside?

The answer is that it does not, without a great deal more ado. Here we do encounter the problem of external validity, and writ large. In this case it makes sense to ascribe the same feature inside and outside the experimental set-up, but there is no reason yet to do so. We can attempt an induction: Some singular causal events that happened once somewhere will happen in the target population. I do not mind inductions, but wild-eyed ones, on a wish and a prayer, are to be avoided in evidence-based policy. We are only entitled to an induction on some given feature when there is reason to think that the new case is like the original with respect to that feature. We properly induce the colour of the camellia buds that have not yet opened on the plant at the bottom of my garden from the colour of those that have, but we do not induce the colour of the flowers my geranium will produce from the flowers on my camellia. That is why the U.S. Department of Education remarks that evidence for efficacy of a programme in your school should be positive results in RCTs in two schools *like* yours. But it is notorious that the kinds of similarities that make for a good basis for induction are seldom transparent.

So, what features could a cause have that make induction reasonable? Once one begins to lay out the kinds of assumptions presupposed in making inductions from experimental outcomes to real-life situations, the enterprise begins to look a lot like the postulation of a capacity. Why should we think that the cause T will do for some targeted individuals what it did for some individuals in the experiment? If we have

reason to ascribe an enduring capacity to T, we have a reason: T caused E in the observed individuals because it had the capacity (power, tendency) to do so, and having that capacity, it will do so for other individuals unless too strongly interfered with. So it may not be that this alternative differs significantly from the capacity idea after all. At the very least the postulation of capacities and the use of induction from singular causal conclusions established in RCTs have in common the concern I am worried about in this paper: The RCT itself goes no way to underwriting the assumption crucial for taking it as evidence for effectiveness outside the setting, whether this be seen as evidence that the cause T is associated with a relatively enduring capacity or some kind of evidence that there is a proper base for an induction. In particular analogues for the three requirements I outline in section 8 must be satisfied just as much for inductions as for the use of the logic of capacities.

b. A casual law and its analogues. Ideal RCTs can also establish what I call *causal laws*. Again, the proof depends on what one counts as an ideal RCT, what one takes a causal law to be and what features one takes systems of causal laws to have. Causal laws as I characterize them – and this is the characterization that is employed in the proof that ideal RCTs can establish causal laws – are always relative to a 'test population'. These are populations that are homogeneous with respect to 'all other' causes for the same effect. So, what the positive RCT can establish is that there is some test population, δ, which is a subset of the population enrolled in the experiment, for which 'T causes E in δ' is true.

Laying aside problems with policies that shift the underlying causal structure, this means that the causal law that T causes E holds for any population that satisfies δ. Two immediate problems face employing this nice result as evidence for effectiveness. First, we usually do not know what δ is. The ideal RCT tells us that there is such a subpopulation but not how to pick it out. Second, we could seldom expect that our target population is a δ population, or has δ as a subpopulation within it. So the information would be useless – unless we suppose that the causal law that holds in one population can be relied on to hold in others, which we often do assume. We do not expect exactly the same effect to appear in other kinds of populations since what actually happens in a given population depends on the combined efforts of all the causes at work in it. But sometimes we expect that the contribution guaranteed by

one causal law will be the same for different kinds of populations, and often with very good reason. But again this returns the focus to capacities and their justification: The logic that supports this assumption is the logic of capacities, or something very akin to it.

c. Probabilities and inductions from them. A third alternative is to abjure causal and efficacy talk altogether; stick with just the probabilities. Take the conclusion simply to be the difference in means between the two experimental groups. Even better, take each mean separately since, without the causal connotations, the difference clearly gives less information than the two means separately, each of which can be used independently in decision-theoretic analyses, as Dawid suggests. Again the question arises, how do the means in the experiment provide evidence about anything outside? Again the answer is that without a lot more ado, they cannot. The problem in this case is usually worse than with external validity of causal conclusions. In general the mean in two populations is the same only if the probability distribution over confounding factors is the same in the two. There may be some cases where the population enrolled in the experiment can be seen as a really good sample from the target population so that the distributions should be the same. But we know this is impractically rare. Indeed this is why JS Mill thought that political economy could never be an inductive science. The probabilities with which a given kind of event occurs depend on the complex mix of causal factors that bring it about and there are generally a very large number of these, each itself brought about by yet an earlier large mix of factors, and so on. So the probabilities for an outcome should be expected to be in constant flux; they are the very thing we would want not to do inductions on. The contribution of a cause (in the technical 'capacity' sense of 'contribution') at least is the kind of thing that we have seen to be relatively stable for a variety of different causes across a variety of different fields – like the gravitational action of one mass on another. Mill for one took the fundamental social causes to be among these. If he is right that social causes are often associated with relatively stable capacities, then there is some hope that RCTs can be of use for social policy. For pure probabilities, by contrast, we almost always have good reason to suppose change rather than stability.

d. Econometrics. This alternative does not rely on RCTs as evidence at all. There are a number of different formal treatments of efficacy on offer in econometrics. These

treatments define efficacy in a population-relative way, not however relative to an experimental population but rather to the target population. These then provide schema for modelling the causal structure of the population and investigate econometric techniques for estimating efficacy from data. As with the treatments in section 3, it is still the case that by definition a cause can only be efficacious in the population relative to which it is defined but now it is defined where the policy-maker needs it. These techniques can provide powerful evidence directly relevant for policy. But they have a big drawback, which is just the flipside of this virtue. They deliver results about the target population directly, but then they require both data and careful modelling on that population, and this can frequently be too expensive and too time consuming to help in policy considerations. Better to have results concerning capacities, 'off-the-shelf' results that do not have to be established anew for each population – if only we can get them.

Let us turn then to considerations about what it takes to establish capacity claims and what role RCTs can play in doing so.

6. What Makes a Concept Legitimate?

Contemporary science studies features that can be ascribed to real systems in the world: the charge of an electron, the structure of a DNA string, the fitness of a population, or – in the case of interest here – the efficacy of a treatment for a given effect, like the efficacy of class-size reduction to improve academic achievement or of phonic awareness to reading development. Very often, as with charge and efficacy, the features are *quantities* – they have magnitudes that measure the amount of the quantity and allow for comparisons of amounts of the quantity that obtain in different systems. With respect to quantities three different notions need to be distinguished.

1. The quantity itself, e.g., *negative charge.*
2. The scientific representation of it, e.g., negative charge is represented as a discrete-valued function that maps physical systems into $-n \times 1.6022 \times 10^{-19}$ *coulombs* where n is any integer including 0, and -1.6022×10^{-19} coulombs is the charge of the electron, the smallest unit of negative charge.

3. Reliable operations for measuring the quantity – e.g. measuring the change of a particle by observing the deflection in its trajectory produced by a known electromagnetic field.

I rehearse these three because they are often conflated, sometimes harmlessly, sometimes not. For instance, "Fuel poverty is defined as circumstances where a household has to spend more than 10 per cent of its income on fuel to maintain an adequate standard of warmth."[7] conflates 1 and 3. Or consider debates about measures of economic freedom – e.g. how adequate is a sheer cardinality measure, which measures the economic freedom of individuals by the size of their set of options? These are debates about the best scientific representations of economic freedom, but they can often look like debates about 1. – What is economic freedom? – or 3. – What operations will reliably measure the amount of economic freedom? Physics too is rife with conflations. 'The quantum state is evolved deterministically by the quantum Hamiltonian' is typical. The quantum state is supposed to be a real feature of systems in the world. The quantum Hamiltonian is definitely a piece of mathematics, which cannot evolve anything in the real world. Either some real feature of systems in the world must be pointed to that is represented by the Hamiltonian or the claim must be read as a purely mathematical claim about the relation of the two representations, with 'deterministically' read as some appropriate mathematical characteristic. From the point of view of capacities, the efficacy definitions from section 3 conflate 1. and 3. (Some of the econometric approaches discussed in section 5 make efficacy model-relative, which conflates 1. and 2.)

Which of these three matters for establishing that a scientific concept stands for something real, or at least for licensing the concept as legitimate? All three have their advocates. But I would wish to argue that in the ideal for a good scientific concept we need all three and the three need to mesh tightly. The account of what the quantity is must dovetail with how it is represented, where the canonical way to show they dovetail is via representation theorems, like those involving probabilities, preferences and utilities (where expressed preferences are considered a function of probabilities and preferences) or those for 'measures' of economic freedom. These theorems show

[7] DWP, 2006, "Opportunity for All, Eighth Annual Report", p.70

that the features picked out to characterize the quantity are appropriately captured by the mathematics used to represent it. (For instance, lengths do not get represented by negative numbers and temperatures do not get represented in a way that insists that twice the numerical value means twice as hot.) Similarly, both the characterizations of the quantity and the representation of it must fit with how we actually measure it in the world: The measurement procedures must be justified as good ways of finding out about the very thing we have characterized and represented. The justification may well be elaborate and indirect, relying on a large number of background assumptions, as in the justification for using the deflection of the trajectory of a particle in an electromagnetic field to measure its charge. But the justification needs to be there if we are to place trust in our measurement procedures.

7. Return to Efficacies as Capacities

From this ideal point of view what is wrong with taking efficacies as enduring capacities measured by RCTs? Generally far too much is missing. We could conceive of the real-life experiment as a way of measuring, but measuring what, represented how? We could additionally think of definitions in terms of mean effects in ideal RCTs relative to well specified populations, like the ones described in section 3, as laying out the requisite representations for some quantity, with real experiments as the real-life measurement procedures. Then we would at least have a transparent connection between the representation and the measurement procedures. But we would still lack an account of what the quantity is that is being represented and measured. The efficacy itself, conceived as a quantity that can be measured in one setting and relied on for prediction in another, is a mere shadow. We have a procedure for measuring, measuring something, but we do not have an account of what it is. We can of course provide such an account by reading off a definition of the quantity from the representation or the measurement procedure. But then we are back where we started in section 3. We would have a good, scientifically well-formulated definition of efficacy. But causes could never be efficacious outside of the experimental settings which enter into our definition.

The problem I note here is in no way peculiar to ‘efficacy’ but is widespread throughout the social sciences. To legitimate a quantity concept we need to characterize the quantity itself, its representation and its measurement procedures. Then we need to tie these three together – in a way that can be justified. And we need to be explicit, precise and rigorous. We naturally face problems here, but which problems depends on which end we start from in tying the bundle together. What generally happens is that the account of representation slides close to one side or the other, making the distance on the far side difficult if not impossible to traverse. We can build a representation very close to our measurement procedures, in which case the bond between those two is transparent and easy to justify. But then, how to reach over to an account of the quantity itself becomes a problem. As we saw with the representations for ‘efficacy’ described in section 3, we sometimes tie representation so closely to measurement that the gulf to the other side where the quantity dwells is literally unbridgeable.

On the other hand the representation can flow naturally from the account of what the quantity is. Consider the representation of charge referred to above, as a discrete-valued quantity. This is readily justified by the principles of electromagnetic theory that help explain what negative charge is, especially the principles that explain that electrons carry the smallest unit of negative charge, that this charge is indivisible and that the charge of an electron is -1.6022×10^{-19} coulombs. In this case we have the converse problem to that with efficacy: It is very difficult to justify our measurement procedures – Why look at the deflection of a particle’s trajectory in an electromagnetic field in order to measure its charge? – and doing so will demand a large number of auxiliary empirical assumptions. It is no surprise then that those who distrust almost all claims to knowledge in the human sciences err in the other direction and tie representation too closely to measurement.

If the situation is so bad, you may wonder why I have defended efficacies as capacities for so long. The reason depends on noting the difference between the abstract description of a category and a concrete instance of something that fits the category. I have defended capacities as an appropriate abstract category of concepts for use in science and I have defended specific efficacies as concrete instances of that category and I have done the first largely on account of the second. That’s because I

see specific efficacy concepts at work in science, concepts that are well defined, appropriately represented and properly measurable.

The paradigm is Mill's own – forces in physics. First of all, 'force' is defined implicitly via the laws in which it participates. Importantly for counting 'force' as a capacity term, among these laws is the law of composition of forces, which fixes what it means to say that a given cause, like gravity or charge, 'contributes' its canonical effect even when other forces are at work as well. This then gives sense to the idea, central to the concept of efficacy as capacity, that the cause should contribute in some 'systematic or intelligible way to what happens whenever it is present'. Second, force has a well-understood mathematical representation, roughly as a vector in 3-space whose components are non-negative real numbers. Third, it is measured in a vast variety of ways that can be defended as appropriate given the force laws in conjunction with acceptable auxiliary assumptions.

To make vivid the gaps that beset the concept of efficacy, consider the force of gravity. The effect of one mass on another is not defined as the force or acceleration that is measured in a very carefully controlled experiment, like those Galileo aimed for. It is defined instead by the role laid out for it in the laws of motion, laws that maintain that the mass contributes the same effect universally, not just in Galileo's experiment, laws whose implications in this regard have been widely confirmed. These well-confirmed laws also show why Galileo's experiments are good for measuring the effect of one mass on another – these are experiments in which the pull due to the mass operates almost on its own, with no other forces to interfere. They tell us in addition exactly what effect to take away from Galileo's experiment and, as I stressed above, they tell us just what it means to say that this effect is contributed in other situations outside the ideal measurement setting.

8. Speaking Realistically

The take-home lesson from the considerations in section 7 is that, for making sure our concepts are good ones, theory matters, and so do auxiliary empirical assumptions, good well-confirmed theory and sound, reliable empirical assumptions. But it is well

known that policy cannot wait for the advance of theory. So we had best consider what is minimally required if we are to take a well-conducted RCT as evidence for effectiveness, even if only very partial evidence. Perhaps there are aspects of the ideal we can get along without. For instance, we may not need a thick theory to tell us just what the capacity involved is and under just what laws it operates. But the very logic involved in using the idea of enduring capacities as a rationale for taking RCTs as evidence bearing on the effectiveness of the cause outside the experiment makes three clear demands. Meeting these three demands provides the three ingredients necessary to turn the efficacy measured in an RCT into evidence for effectiveness in a new setting. It's like making pancakes. RCTs are the baking powder. But the baking powder is useless without the flour, milk and eggs.

1. We need good reason to think that the effect produced in the experiment is an enduring one, reason to think that when we see differential effects with the cause present versus absent in the RCT setting, we are seeing the effects of a genuine capacity, one that can reliably be expected to operate in various new settings and in new populations.
2. We should have good reason to think that the proper effect has been identified, that the effect we focus on in the experimental setting is not piggybacking in a misleading way on the true, generalizable effect. "For example, let us say that an experiment [RCT] is conducted to increase security and reduce theft in two schools through the introduction of closed circuit television (CCTV). The effect is a reduction in theft in the experimental school. Exactly what is the cause here? It may be that potential offenders are deterred from theft, or it might be that offenders are caught more frequently, or it might be that the presence of the CCTV renders teachers and students more vigilant…"[8]
3. We need some sense of what it means for the observed effect to be *contributed* in new cases. In formal theories this is supplied by the *rules for composition.* But there are different methods in different theories. The vector addition of forces is different from simple addition. Sometimes we introduce new, intermediate words to describe how the cause contributes even though we don't know how to predict the ultimate outcome. Consider: The magnet

[8] Morrison, K., 2001, "Randomised Controlled Trials for Evidence-based Education: Some Problems in Judging 'What Works'", *Evaluation and Research in Education*, Vol.15, No.2, pp.69-83, p.72

always *pulls on* the pin even when we cannot predict whether the pin will be set free. Or the commitment about contribution can be couched in a cautious, modalized form: Eating Wheaties *can* (or 'may') improve your heart health. And of course threshold effects are a notorious problem.

To illustrate the importance of these three basic ingredients for using efficacies to help predict effectiveness let us look in more detail at the third and consider – in caricature – the difference between two well-known examples, simultaneous equation models in economics and forces in physics. These different notions of how causes contribute give radically different predictions about what happens when the cause is present in new circumstances.

Economics first, where supply and demand jointly determine quantity exchanged, via two equations:

Supply: $q_s = \alpha p + \mu$
Demand: $q_d = -\beta p + \nu$.

The theory supposes that in equilibrium $q_s = q_d$ and that *both equations are satisfied at once.* This last is a rule of composition. The supply equation describes the *contribution* to the quantity exchanged from the combined supply-side causes (here price, p, and other unnamed causes, μ) but it does not select what the actual value will be because price and quantity are taken to be fixed simultaneously by both equations of the model. Similarly, the demand equation describes the contribution to the quantity exchanged from the combined demand-side causes (price, p, and other unnamed causes, ν) and again it does not select any actual value. The rule of composition describes how these two separately-contributed effects combine to fix the value that the quantity exchanged actually takes: Both equations must be satisfied at once.

Now forces. A particle of negative charge q_1 hovers in midair, pulled down by the mass M of the earth and upwards by the pull of a positive charge q_2. The equations that govern its acceleration are from gravitational and electromagnetic theory respectively:

$acc_g = gM/\mathbf{R}^2$

$acc_{em} = \varepsilon q_1 q_2/\mathbf{R'}^2$.

Here **R** is the vector representing the particle's distance from the earth; **R'**, its vector distance from the second charge; g, the gravitational constant and ε, the electromagnetic constant. As is well known, the acceleration the particle actually experiences – which in this example is zero – is given by vector addition.

Compare. There is a sense in which the separate capacity claims in the economics model are more informative than those in the physics case. That's because the effect described is bound to happen no matter what particular form the contribution from the other side takes since both equations must be satisfied. Even though the effect is not fixed precisely, there is information about the quantity exchanged that is bound to hold no matter – the information that it lies on the line described. This information may turn out to be of great use or it may be of little use in evaluating the efficacy of a policy that proposes to tinker with demand or supply-side causes. That depends on the policy setting. But (as the theory has it) knowing either the demand equation puts the quantity on the demand line and that can be relied on even if we know nothing about the supply equation, and vice versa.

The economics rule for calculating the net effect of different contributions is in sharp contrast with the physics case. With vector addition, knowledge of either contribution separately does not provide any information about the actual value of the resulting acceleration. The acceleration is not constrained in any way by the presence of gravitational or of electromagnetic causes alone whereas the quantity exchanged is constrained to lie along given lines by both the supply-side and the demand-side causes separately. For physics the best we get are counterfactual constraints: If the other causes at work (whatever they are) were to stay fixed, adding the pull of gravity would change the acceleration by the vector addition of $acc_g = gM/\mathbf{R}^2$; similarly, adding the pull of another charge would change the acceleration by the vector addition of $acc_{em} = \varepsilon q_1 q_2/\mathbf{R'}^2$. These strongly *ceteris paribus* counterfactuals are of limited use in situations where other changes may get introduced by the policy as well.

For most cases of social policy the causes to be worried about are not described by nice theories like these with nicely articulated rules of combination, nor do they have nicely articulated accounts supporting the claim that the causes are associated with relatively enduring capacities that guarantee effects that contribute in some systematic way across new situations. What do we do then?

In the case of combination, we often assume simple linear addition: If the cause produced an improvement of size x in the RCT, it will produce that size improvement elsewhere. But of course we know better and normal intelligence dictates caution. Sometimes the effects taper off at the margins, sometimes there are clearly drawn thresholds, sometimes, without the appropriate helping factors the effect cannot be produced at all. Even if, unlike what happened in California, other factors are held fixed, smaller classes may make very little difference where reading scores are already high and they will make none at all in classes of children who do not have the capacity to read.

What legitimates the assumption of one kind or another of additivity and where do cautions like these about the assumption come from? They come from some story, account, or theory of what the capacities of small class size are and how they work. And this is where they must come from. We need – we always need – an account that supplies the three basic ingredients if we are to turn efficacies into evidence for effectiveness. We may well not have very good accounts to supply these basic needs but that does not mean we can pretend we do not need them.

If we are not to be led astray, sometimes far astray, by introducing RCT results into policy considerations, it is not enough to be told that the RCT was well conducted. These three requirements must be met in some reasonable way as well. You can't make pancakes without flour, milk and eggs no matter how much baking powder you pour into the bowl. By itself the RCT is not evidence for efficacy in a new setting. It is evidence only *conditional* on these other three ingredients. If we have little idea about how to supply these basic ingredients, we need to figure out how to cope with that problem, and ignoring it is not coping. We can make bets in situations of ignorance but for policy we should make intelligent bets. The probability that the RCT provides evidence and exactly what it provides evidence for depends on what

the other three ingredients might be like that will turn it into evidence. These issues need to be thought about, not swept under the rug; and in the end, as with most cases where bets are dicey, perhaps we need to hedge them heavily.

These remarks are not meant to be of serious practical help. They are rather a warning and a call for refocusing attention in evidence-based policy. Clearly the possibility of establishing efficacies in well-conducted RCTs is a great boon, no matter which definition of efficacy is settled on. But they are only a help when the other ingredients are there to make them so. So, far more effort needs to be focussed on how to secure the other ingredients and how to cope when our knowledge of them is insecure, as it so often is.

9. A Brief but Important Warning

I say that we need a theory – a story, or an account – strong enough to supply the three basic ingredients of section 8 if efficacy is to contribute evidence for effectiveness. It is a major problem for policy that often the theory is missing. What is important to keep in mind is that in many cases our problem may not be due to the fact that we have not yet found the appropriate theory but rather that there is no theory to be found. Much of science works by postulating capacities, by the use of the analytic method: We study the components separately then make predictions by 'adding' their effects in the appropriate manner. This works for forces and for biological mechanisms (such as chemical transmission between neurons)[9] and Mill was concerned it would work for political economy. But we should be wary. There is no necessity that what a cause produces in one setting will have any systematic relation with what it does in another. Much of social phenomena may be too holistic to yield to the analytic method. Indeed both Anna Alexandrova and Julian Reiss argue that a lot of the recent work in experimental economics suggests that the analytic method is not working so well in political economy as Mill – and a great many more recent economists – had hoped. In many cases evidence is mounting up against Mill's hope that the important economic causes are associated with stable tendencies (or, in the vocabulary used here 'capacities'). And if this is true for

[9] Machamer, P., Darden, L., Craver, C.F., 2000, 'Thinking about Mechanisms', *Philosophy of Science*, Vol.67, No.1

economic causes, it seems all the more likely to be true for the social causes outside the domain of economics.

What then of RCTs? When the analytic method fails they may be suggestive of what to look for in new cases, but they can hardly count as evidence at all. Inferring from efficacy to effectiveness is induction on a wing and a prayer. I doubt, however, that serious social scientists conduct many RCTs without good reason to think the causes tested are associated with relatively stable capacities. The reason though is generally bound up with theory, which is despised by many of those who advocate taking over Cochrane doctrines for policing social evidence. It is despised because the relevant theories are controversial or ill-formed or poorly supported or all three. This returns us to the finishing point of section 8: we need theoretical assumptions, and assumptions of just the right kind, else all our very careful efforts at experimentation go to waste. So we had best not despise dicey theories but learn how to manage their uncertainties.

10. In Sum

The purpose of this paper is not to attack RCTs or any other way to inform ourselves about efficacy, nor to promote theory above or to the exclusion of everything else in deciding policy. Its aim is merely to argue that in our attempts to get clear and rigorous about what we mean by evidence, we have concentrated too much on getting the methodology of RCTs straight and too little on getting clearer what theories, or accounts, or stories we need to make an RCT evidence at all. For without some theory about the capacities of the causes we are studying, the evidentiary value of an RCT is not just weakened, but, I argue, made empty, zero.

The contemporary theory and practice of the RCT is a brilliant intellectual achievement. But a good RCT –and there are many of them – is only one component of a set of ingredients necessary to create a piece of evidence that policy will be effective. I say ‘a piece of evidence’ deliberately. Even where we have a very good, very full theory, one measurement by one team in one environment is not very strong evidence. (It certainly never suffices in physics, where the theory is often exceedingly strong – though a result from an experimental team that is known to be

exceedingly good can carry considerable weight.) This may in the end be all we have – and lucky to have it at that. But that does not make it strong evidence. Again it is important to recognise this and to factor in the large uncertainty that is left, not rather than avert our gaze and pretend it does not stalk on policy. The efficacy that we claim to have established by an RCT to test a particular policy is an empty notion with zero evidentiary power if we have no theory to supply the missing ingredients that turn the efficacy into evidence for effectiveness. This is a more fundamental concern about RCTs than the worry that in very many cases there may be difficulties in eliminating bias, controlling for confounding factors, getting the right sample, and so on. It says that even if all these technical problems are absent, there is no read across from efficacy to effectiveness without some theory of some sort, and efficacy as a notion has no power beyond that defined for it by the statistical procedures themselves. Its connection with effectiveness is no more than sharing its first three letters.

I use words like 'theory' or 'account, or 'story' and 'some sort of' advisedly. We all agree that in the social sciences, as compared with in the natural sciences, our theories are typically quite poor. We wish it otherwise, and work to make it so. But we are where we are. And the consequence of recognising this poverty is not that we should abandon theory until it performs better. This is because without some sort of a theory, some non statistical idea of what may be going on, we have no evidence from RCTs at all. We just have words and numbers written on a page. The theory is not some fancy add-on which can be bypassed by going straight to the evidence. It is what gives the RCT life.

It is easy to see how RCTs have earned so much attention. Consider the position of policymakers - not methodologists - wanting conscientiously and intelligently to decide what to do. Their raw material will include some theory, about what, say, makes juveniles offend; some factual evidence, e.g. if they are lucky an RCT; and a good deal of anecdote and folk wisdom. (Forget the politics.) When they talk to serious academics and similar, they learn to neglect the anecdotes, as not scientific enough to be evidence. They also learn that there are a lot of theories about juvenile offending, from the sociological to the medical to the psychological, but none of them looks conclusive, and they conflict.

So they say: I will just look at the evidence, at what works. And advocates of RCTs say - and this paper agrees with them –that RCTs represent an extraordinarily powerful engine for testing efficacy, and hence – and this paper does not agree – for identifying effectiveness. At the very worst, the policy makers may then think: We are at least looking at that part of the raw material that is telling us something precise and relatively certain and uncontroversial. It would be nice to have a good theory of juvenile criminal behaviour, as the physicists have of gravity. But we don't. Theory plus evidence may add up to the gold standard for policy determination. But we have to use what we actually have – the evidence of what works. And maybe that is all we need. Why do we need theory if you have incontrovertible evidence of efficacy?

My argument is that if you don't have both you don't just have half or whatever of what you should have, but that you have nothing. We all recognise that theory without evidence to support it leads to no conclusions. The reverse is true as well.