Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1356-1294

The limitations of randomized controlled trials in predicting effectiveness

Nancy Cartwright PhD FBA[1] and Eileen Munro PhD[2]

[1]Professor of Philosophy, London School of Economics and Political Science, London, UK and Professor of Philosophy, Department of Philosophy, University of California, San Diego, La Jolla, California, USA
[2]Professor of Social Policy, London School of Economics and Political Science, London, UK

Keywords
capacities, external validity, multisystemic therapy, randomized controlled trials

Correspondence
Nancy Cartwright
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
UK
E-mail: n.l.cartwright@lse.ac.uk

Accepted for publication: 19 December 2009

doi:10.1111/j.1365-2753.2010.01382.x

Abstract

What kinds of evidence reliably support predictions of effectiveness for health and social care interventions? There is increasing reliance, not only for health care policy and practice but also for more general social and economic policy deliberation, on evidence that comes from studies whose basic logic is that of JS Mill's method of difference. These include randomized controlled trials, case–control studies, cohort studies, and some uses of causal Bayes nets and counterfactual-licensing models like ones commonly developed in econometrics. The topic of this paper is the 'external validity' of causal conclusions from these kinds of studies. We shall argue two claims. Claim, negative: external validity is the wrong idea; claim, positive: 'capacities' are almost always the right idea, if there is a right idea to be had. If we are right about these claims, it makes big problems for policy decisions. Many advice guides for grading policy predictions give top grades to a proposed policy if it has two good Mill's-method-of difference studies that support it. But if capacities are to serve as the conduit for support from a method-of-difference study to an effectiveness prediction, much more evidence, and much different in kind, is required. We will illustrate the complexities involved with the case of multisystemic therapy, an internationally adopted intervention to try to diminish antisocial behaviour in young people.

1. Introduction

What kinds of evidence reliably support predictions of effectiveness for health and social-care interventions? There is increasing reliance on evidence that comes from studies whose basic logic is that of JS Mill's method of difference. These include randomized controlled trials (RCTs), case–control studies, cohort studies, and some uses of causal Bayes nets and counterfactual-licensing models like ones commonly developed in econometrics [1,2]. References therein discuss these methods further. The topic of this paper is the venerable issue of the 'external validity' of causal conclusions from these kinds of studies. We shall argue two claims. Claim, negative: external validity is the wrong idea; claim, positive: 'capacities' are almost always the right idea, if any idea is right.

If correct, these claims imply big problems for policy decisions. Often in making these decisions one hopes to rely on some solid Mill's-method-of-difference studies. Indeed many advice guides for grading policy predictions give top grades to a proposed policy if two good Mill's-method-of difference studies support it. But if capacities are to support an effectiveness prediction from a method-of-difference study, much more evidence, and much different in kind, is required, and at two stages: First, to support the capacity claim; second to supplement this claim to ensure the capacity will operate in the target situation, and operate in the expected way. So a lot more work is required than hoped. To illustrate the complexities involved we consider multisystemic therapy (MST), an intervention internationally adopted to diminish antisocial behaviour in young people.

Section 2 below explains Mill's method of difference, what it can establish, and why; section 3, some sufficient conditions for claims established by method-of-difference studies to be externally valid; section 4 introduces three kinds of causal claims; section 5 explains capacities and their logic; section 6 explains why reasoning with capacities requires different evidence than method-of-difference studies provide; and section 7 describes the discrepant findings of RCTs on MST and the current difficulties in deciding how to interpret these results.

2. Mill's method-of-difference studies

Mill's method-of-difference studies locate differences in the probability of a selected outcome (O) with and without the treatment/intervention (T) across two groups that have identical distributions

for all factors causally relevant to the outcome except those causally downstream from the treatment. The intent is to draw a causal conclusion. But for that, some assumptions must be made to connect causes with probabilities. The most standard one, which we shall suppose here,[1] is

Causal fixing (CF): The probability of an effect is fixed by the values taken by a full set of its causes.

RCTs, case–control studies, cohort studies, and some uses of causal Bayes nets and counterfactual-licensing models like ones commonly developed in econometrics all follow the method-of-difference logic. This logic is deductive. That is its special strength. *Given* CF, *if* a positive probabilistic difference obtains and *if* the two groups have identical distributions of other causal factors, it follows T causes O in some subpopulation, φ, of the population, X, of the individuals in the study that is causally homogeneous with respect to O.[2] The methods differ by how they try to match the distribution of other causes in the two groups and sometimes by the techniques used to infer probabilities from observed frequencies.

3. External validity

We must be careful though. A positive difference between treatment and non-treatment groups shows that T causes O in some causally homogeneous subpopulation φ of X. This is a very narrow conclusion, of little use as it stands. Hence enters the venerable problem of *external validity*: When will the conclusion established for study population X hold for target population θ?

It is easy to provide sufficient conditions. T causes O in any new population θ in which both (i) the same causal laws for O hold as in X (so that the same factors will be causes of O as in X);[3] and (ii) some causally homogeneous subpopulation, φ, of X in which the probability of O is greater in the treatment than the non-treatment groups is a subpopulation of θ.

Not only are these conditions sufficient for T to cause O in θ. They are also necessary to justify exporting study results to θ. After all if θ has different causal laws or different subpopulations than those causally positive in X it may be true that T causes O there but whether it does so in X is irrelevant.

What if we back off though? Do not draw a causal conclusion; merely look at some probabilistic fact, for instance the *effect size*: the mean difference in O between the treatment and non-treatment groups in X. When will effect size be the same between X and θ? An answer requires assumptions about how the probability of O gets fixed. The most reasonable one we know is CF.[4] Given CF, the following conditions are sufficient. The effect size is the same in θ as in X when (i) X and θ are the same with respect to the causal laws for O; and (ii) X and θ are the same with respect to the probability of all causally homogeneous subclasses.

Otherwise it is an accident of the numbers.

So, very restrictive conditions must be met for effect size to travel from study to target populations. These conditions are restrictive not only in the sense that they may not hold widely, but also with respect to the epistemic demands they make. RCTs are now taken as a gold standard in causal inference throughout health care and social policy worlds because they are supposed to best control for bias from unknown confounders. So it is widely acknowledged that we generally don't know all the important causes for a factor, let alone knowing the distribution of subpopulations homogeneous with respect to these in the study and the target populations as (ii) requires.

Nor is (i) easier. Most causal and probabilistic relations relied on in health care and social practice are not fundamental: They do not just hold, they hold on account of some underlying structure that gives rise to them. When the structures are different, so too are the causal and probabilistic relations they create. For instance, stepping on the right-hand lever on the car floor – that is, the throttle – causes the object the lever is attached to accelerate. Stepping on the lever attached to the end of a toaster produces something entirely different. In this example the underlying structures are mechanical. In cases of interest for health and social policy they will be a mix of institutional, psychological and physical. The basic lesson is the same. Different underlying structures yield different causal and probabilistic relations. The problem is we often do not understand these underlying structures nor how they work to give rise to the causal relations an intervention might use. So we don't know when (i) is satisfied. For some causal relations it may be good to assume, as one economist recently claimed, that people are much the same wheresoever they are; for others that assumption can be disastrous. So the demands for exporting effect size from study to target population are generally far too great.

A weaker conclusion concerns the direction of the effect size: When will a positive effect size in X be sufficient for a positive effect size in θ? A number of separately sufficient conditions are immediately apparent. Effect size direction will be the same

- If T has same effect on every individual.[5]

Or

- If X and θ
  - **1** Have the same causal laws, and
  - **2** *Unanimity:* T acts in the same direction with respect to O in all causally homogeneous subpopulations.[6]

Or

- If θ has 'the right' subpopulations.

Again, these are strong conditions that may often fail to obtain. Is there then no other kind of useful conclusion to be exported more widely? We believe there often is. To see what kind of conclusion that is we offer some simple distinctions among types of causal claims.

[1] There are other ways to draw this connection. Probably the other most dominant is that of Holland and Rubin [3]. This links method-of-difference studies to effectiveness predictions via singular counterfactuals. These two methods are closely connected, however, which is apparent from the three-decade-old literature connecting probabilistic causality with the probability of counterfactual conditionals.

[2] Note that this is consistent with T doing exactly the opposite in other subpopulations. All that we can be sure of from a positive difference is just that T is causally positive in *some* subpopulation.

[3] If this condition doesn't hold then it makes no difference if φ is a subpopulation of the new population because now φ is defined as a population in which some given set of factors all have the same values. But this is irrelevant if those factors are not the causal factors for O in the new population.

[4] So it seems causality must enter at some stage of reasoning.

[5] This is similar to a requirement made by Holland and Rubin [3].

[6] Note for section 5 that this is sufficient but not necessary for the claim that T has a stable capacity to promote O.

4. Three kinds of causal claims

In order to understand the route from method-of-difference studies to effectiveness predictions it helps to distinguish three kinds of causal claims:

1 *It-works-somewhere claims:* T causes O somewhere under some conditions (e.g. in study population X, administered by method M).

2 *Capacity claims:* T has a (relatively) stable capacity to promote O.

3 *It-will-work-for-us claims:* T would cause O in population θ administered as it would be administered given policy P (i.e. effectiveness claims).

Given CF, method-of-difference studies can establish it-works-somewhere claims and medical and social sciences work hard to do so.[7] But what makes these evidence for effectiveness claims: T would cause O in θ administered as it would be administered given policy P (T will work for us)? The standard answer is external validity. An alternative is capacities.

5. Stable capacities

'T has a stable capacity to promote O.' What does this mean? Cartwright provides a detailed answer in a number of places [5,6] and there is currently a great deal of work by other authors on the related (possibly identical) notion of a causal power [7–10]. Rather than pursuing these details here some examples may suffice: Masses have a stable capacity to move other masses towards themselves; aspirins have a relatively stable capacity to relieve headaches.

A factor with a (relatively) stable capacity to promote O always (or across a range of situations under consideration) makes the same *fixed contribution* towards O. But this can – indeed, generally does – differ from what outcomes occur when the factor is present. The mass of the earth always pulls objects towards itself even if a magnet or the table-top prevents them falling. Similarly, aspirins generally have a positive effect on headaches even if my headache grows worse because of the stress of my job. At least the headache isn't as bad as it would be without the aspirins. One might say that aspirins 'try' to relieve the headache or that the mass of the earth 'tries' to make the body fall, although of course no conscious effort is involved.

In reasoning about effectiveness we often assume that factors have stable capacities. Consider for example the canonical explanation for the failure of the California class-size-reduction programme [11]. A well-conducted RCT in Tennessee established that small class sizes there improved reading scores: The study supported an it-works-somewhere claim. California reduced class sizes but reading scores did not improve. The usual explanation is not that the Tennessee study was flawed; nor that it was irrelevant due to different structural features in California; nor that it was an entirely local interactive effect from which no further lessons could be drawn.

The canonical explanation points instead to the fact that California rolled out its programme over a short time. Suddenly class sizes were cut in half. Twice as many teachers were required and twice as much classroom space. Neither was available. So teaching quality and learning experiences went down along with class size. But not because smaller classes do not contribute positively to reading scores. Rather, so the explanation goes, their good effect was offset by the bad effect of poor teaching and poor educational surroundings. The California scores were a result of all the contributions, positive and negative, 'added' together, just as when a magnet and gravity act together on a pin that doesn't fall.

So, some factors have relatively stable capacities and we regularly rely on that in our reasoning. When causes do have stable capacities, what we learn about their contributions in method-of-difference studies[8] can be exported to more situations than those where 'external validity' holds. External validity in all the forms we have discussed supposes that the same facts true in the study are true in the target, whether these be probabilistic facts or facts about what the factor causes (as opposed to what it contributes).

There are, however, two major problems with capacities. First, Mill's-method-of-difference studies can't establish them. A cause can make a difference in a specific situation yet there may be no stable contribution that can be relied on elsewhere. When there is a stable contribution we are entitled to what Mill called 'the analytic method' [12]. That is, we can establish what each separate cause contributes, then rely on some 'rule of combination' – like vector addition with forces or simple scalar addition[9] – to calculate what happens when a number of causes occur together. But not all fields are open to the analytic method. Mill argued that it can be used in mechanics and in political economy but not in chemistry. Chemistry, he thought, is more holistic: How a cause behaves depends on the other factors it interacts with and on its environment; what it does in one environment has little bearing on what it will do in another.

So it is an empirical question where capacities are likely to occur, whether a given cause has a stable capacity and across what range it is stable. A method-of-difference study can reveal something about the contribution of a capacity but it cannot establish that there is a capacity there to begin with. How to establish that is a complicated matter – just look at what it has taken to establish which causes carry forces in mechanics (i.e. what causes make stable contributions to motions) and how these combine. Unfortunately the methodology for establishing claims that a factor has a capacity is not laid out with anything like the degree of completeness and rigor available for establishing other hypotheses, such as it-works-somewhere claims.

Nevertheless a wealth of evidence of different kinds can clearly make specific capacity claims probable and when they are probable they are a powerful tool for predicting whether an intervention will work for us. What is observed in method-of-difference studies will contribute to this evidence base but the history of mechanics should remind us that a lot more is necessary as well, even if we

[7] See Meinert's claim, in Steven Epstein's book on diversity ([4], p. 98).

[8] What do we learn in method-of-difference studies about the contributions of causes, supposing there is a stable contribution there to begin with? What we see is an average of what the cause contributes across the causally homogeneous subpopulations in the study population. How we reason back from that to a representation of its contribution that can then be slotted into a rule of combination to predict what happen when the cause acts in consort with other causes depends heavily on what we otherwise know.

[9] For other rules of combination see [5,13].

cannot lay out a recipe for what the required ingredients are. Nor should we be discouraged. There is a wealth of scientific successes where we have acquired just the kind of detailed knowledge necessary to bring together factors with different capacities to produce relatively predictable outcomes, from global positional systems to heart transplanting and prosthetic knees.

6. Capacities and contexts

Establishing capacity claims then is difficult, far more difficult than establishing it-works-somewhere claims. This is in part because claims that a factor has a capacity to make a given contribution neither make sense nor are testable in isolation.[10] That's because there are a number of substantial implications that must be met if a capacity is to be ascribed to a factor. In particular, 'T has a stable capacity to promote O' implies that there are facts of the matter about

- *Mode of operation*: how T operates to promote O;
- *Necessary auxiliaries*: what must be in place for T to operate to promote O;
- *Destroyers*: what can destroy or overwhelm T's operation;
- What *other capacities* promote and retard O; and
- *Rule of combination*: what happens when many capacities are at work simultaneously?

So capacity claims make sense only relative to a far larger body of knowledge. That is, ultimately we need what we would call a *theory*. This of course runs contrary to much of the founding hope for evidence-based health and social policy. After all, RCT advocates like them because it seems no theory is required to do what they do – but recall, what they do is to establish 'it-works-somewhere' claims.

Consider a case using everyday physics. We choose this because it is simple, well understood and does not involve subject-specific commitments in health and social care. Magnets have the capacity to lift objects. Claims about their attractive powers have passed far more than two good RCTs; they have centuries of study behind them. Imagine: you have access to a desk magnet and a large industrial magnet and you know the exact strengths of these with a high degree of certainty. Should you use one of them to lift an object in your driveway? *That depends on features of the object and its surroundings*.

First, magnets need helping factors to be effective. A desk magnet is useless for lifting a matchstick; it is only the combination of a magnet and a metal object that produces a magnetic force. To predict whether the magnet will work for you, you need to know what the necessary auxiliary factors are.

Then the acceleration caused by the magnet is only one part of the story, often a small part. To know what happens when you use the magnet you need to know the other forces as well, especially gravity. The desk magnet may lift a pin but it is hopeless for your car, where you need the industrial magnet. You also need to watch for other forces you introduce while getting the magnet in place. Perhaps the industrial magnet would have lifted the car if you hadn't thrown its heavy packing case into the boot. Finally, you need to know how all these factors combine to produce a result. Often in health care and social contexts simple additivity is assumed: Add a good thing and the results can only get better. But that doesn't work in even this familiar physics case. We get so used to vector addition that we forget that it isn't simple scalar addition. Add a magnetic acceleration of 42 feet/second/second to that of gravity's 32 feet/second/second and you won't necessarily get 74 feet/second/second.

Whether and to what extent the magnet will be effective in the target situation depends on the causal structure there. It will be hard to make even roughly accurate predictions without investigating that situation and making a reasonable assessment of what the overall outcome will be when the relevant factors operate together.

This can seem daunting. But consider: You know industrial magnets would pass any number of method-of-difference studies, of any degree of stringency. But that's not anywhere near enough to know. None of us would rent an industrial magnet to remove a load of rubbish before examining the rubbish. Knowledge that magnets like this *can* lift is only a small part of what we consider when evaluating if the magnet will be effective in removing our rubbish. If this is so in everyday calculations and in applied science and engineering (like how to build a laser or an artificial limb), why expect predicting the effects of social and health care interventions to be substantially different – and substantially easier?

Of course this kind of complicated causal reasoning is hard, even if we are prepared to be rough in our approximations and figure out ways to tolerate uncertainties. Happily sometimes there are shortcuts, what psychologist Gerd Gigerenzer calls 'cheap heuristics' [14]. For instance, one powerful cause can swamp everything else so we needn't model the rest. If you are shooting a bullet through someone's heart you do not need to measure his cholesterol to calculate his longevity. Or, as with the magnet and the matchstick, the absence of some necessary auxiliary can show immediately that a policy will not be effective.

Failing a nice heuristic for a case, we advise: Do your best with the resources and time available. But reason in a sensible way. Do not optimistically expect external validity without reason to think that sufficient conditions for it are satisfied. In the same vein do not suppose that causal factors have stable capacities without good reason. Embrace capacities where you have reason to believe they hold because they are a powerful tool. But then remember that more work is needed to make reasonable bets about what the outcome will be when a cause with a known capacity is introduced. And recognizing that knowledge is missing at every stage, be prepared to manage uncertainty.

The next section gives a child welfare example to illustrate what options may be overlooked when we rely on a restricted set of evidence, mostly of an it-works-somewhere kind, and ignore causal capacities, both the power they might provide for our reasoning and the problems they entail.

7. Multisystemic therapy

A brief description of MST is

[10] Just consider precise claims about electromagnetic attraction and repulsion. This is never present on its own; gravitational attraction always acts as well. So these claims can only be properly tested by experiments on the motions of charged particles if we already know how to 'subtract away' the contributions of gravity.

> MST posits that youth antisocial behaviour is multi-determined and linked with characteristics of the individual youth and his or her family, peer group, school, and community contexts. . . . MST interventions typically aim to improve caregiver discipline practices, enhance family affective relations, decrease youth association with deviant peers, increase youth association with prosocial peers, improve youth school or vocational performance, engage youth in prosocial recreational outlets, and develop an indigenous support network of extended family, neighbors, and friends to help caregivers achieve and maintain such changes. Specific treatment techniques used to facilitate gains are integrated from those therapies that have the most empirical support, including cognitive behavioral, behavioral, and the pragmatic family therapies. [15]

MST has been widely adopted in North America and Europe and subject to many RCTs. A systematic review for the Cochrane Collaboration identified 13 studies as meeting the inclusion criteria for the review; it reported that the results of these studies vary [16]. Several studies found some positive outcomes for the young people treated but there is no consistency in which outcome variables show improvement; some show no improvement compared with standard intervention. A large study in Ontario, Canada where MST was offered to juvenile delinquents, found no significant difference in reconviction rates at 3-year follow-up. Similarly, an RCT in Sweden involving four sites and 156 youths who met the diagnostic criteria for conduct disorder reports: 'There were no significant differences in treatment effects between the 2 groups. The lack of treatment effect did not appear to be caused by site differences or variations in program maturity' [17].

Can policy makers use these mixed findings? The positive studies appear to show that MST 'works somewhere' but how can policy makers decide whether it will work for them? Current debates about MST show it is difficult to resolve these problems using the concept of external validity.

Treatment fidelity

One explanation for the inconsistent findings is that workers were not implementing the intervention correctly. The premise 'T causes O' is not falsified because these were not instances of T. This is offered in explanation of the poor results from the Canadian study: '[A]lthough the quality and quantity of adherence data are largely unknown, the site with apparently the worst adherence had the worst outcomes' ([18], p. 454).

The MST group, however, emphasize the importance of treatment fidelity and offer a package of services to secure it [19]. The Swedish study made measuring treatment fidelity one of the key aims in order to test whether it explained outcomes [20]. But their fidelity scores did not differ significantly among the six MST teams. On two outcome measures, higher fidelity scores were significantly correlated with higher outcome scores but there was no significant difference on the remaining factors. Therapists with high fidelity scores were compared with those with low scores. The results were mixed. On eight measures higher scores indicated more favourable outcomes, while for 10 measures the effect sizes indicated a negative outcome for the group with the highest fidelity measure ([17], p. 557).

A related claim is that as the programme becomes embedded, workers become more expert at applying the intervention: Programme maturity improves outcomes. Again, the evidence on this is mixed ([17], p. 557 and [18, p. 453]).

Weighing the evidence

Because treatment fidelity and programme maturity fail to provide convincing explanations of the inconsistent results, the field seems confused about how to rate MST. Using the advice guides for grading policy predictions, top grades could be awarded to it because more than two good Mill's-method-of-difference studies support it. Unfortunately, more than two good studies don't. In the health and welfare field, most of the organizations that rate interventions include MST as having demonstrated evidence of effectiveness. Indeed the MST web site provides a list of prestigious organizations that offer this endorsement [21]. But for the Canadians and Swedes who spent millions of pounds on the MST license and on evaluating their services this endorsement has been misleading. How should others proceed given this mixed message?

Section 3 stated that the effect size is the same in target population θ as in study population X when (i) X and θ are the same with respect to the causal laws for O; and (ii) X and θ are the same with respect to the probability of all causally homogeneous subclasses.

The inconsistent findings indicate that these two premises are not true for Canadian and Swedish populations. Either there are causal factors in Canadian and Swedish populations that are significantly different from those in the positive studies in some US states or their distributions are different enough to account for the differences in effect size. The trouble is, how do we work out what these are? Traditional advice on RCTs seeks to control for unknown confounders through randomization but this clearly is inadequate in this case because all studies were randomized. In the positive RCTs the variation in the factors showing significant improvements suggests that, even in this group, there are different causal factors in operation. How are we to find out which factors matter?

Stable capacities

Instead of some overall judgment that it will, or will not, work in new sites, can we identify stable capacities in MST instead? Capacities that because of site structure differences do not produce consistently positive results? There are some ready candidates for such a label. MST was developed using empirical research on key risk and protective factors for youth antisocial behaviour and incorporates empirically based treatments insofar as they exist, for example, cognitive behavioural approaches, behavioural parent training. It has nine core treatment principles, all of which have empirical support or are generally viewed as good practice, for example, being positive and strength-focused, present-focused, action-oriented and well-defined. The focus of theoretical exploration could thus be on what factors destroy or overwhelm T's operation, what other capacities promote or retard O, and what happens when many factors are at work simultaneously.

However, considerable work needs to be done both in theoretical development and testing to identify the contribution of other factors. Efforts to explain the discrepant results to date do not

generally consider capacities but the Swedish results produce speculations that fit better with the concept of capacities than with 'it-works' claims.

One hypothesis, for example, is that MST had positive results compared with treatment as usual (TAU). TAU varies between states and countries so negative results may arise because TAU in those sites is so similar to MST, that is, MST *does* work but so does TAU, so policy makers need to consider their relative cost-effectiveness [17]. This is plausible given MST's roots in empirically supported assumptions. However, to test this hypothesis work needs to be done analysing TAU to discover how and how much it resembles MST. Such research might strengthen claims to have identified stable capacities and help us understand how they operate *in situ*.

One explanation of the Swedish findings points to the differences in the social context. In the USA, juvenile offenders are treated by the justice system; in Sweden they fall almost always within the child welfare system and this, Sundell and colleagues suggest, promotes rehabilitation whatever the method [17]. However, this also holds true for Norway and MST achieved positive outcomes there so any link is not straightforward. Sundell and colleagues also identify poorer and higher crime and substance abuse neighbourhoods in the USA as more difficult contexts where the greater power of MST is required to reduce offending while the less powerful TAU is sufficient in Sweden. These speculations are examples of efforts to identify necessary auxiliaries (what must be in place for T to operate to promote O) and other factors that promote or retard O. However, the debate is underdeveloped and there is little show of sustained attempts to test these speculations more rigorously.

On the whole, the debate on how to interpret the results still centres around trying to determine whether MST does or does not work, sometimes deteriorating to personal attacks on one's opponents which suggests the disputants are unsure how to progress the debate when the usual pathway of labelling an intervention as effective is blocked.

8. Conclusion

Evidence-based policy and evidence-based practice are highly valued in health and social care. The dominant view at present of what evidence is reliable gives greatest weight to evidence from RCTs. This, it has been argued, is insufficient to meet the needs of policy or practice decision makers. A properly conducted RCT provides evidence that the intervention works somewhere (i.e. in the trial). The decision maker, however, needs to estimate 'will it work for us?' In health and social care the underlying social and physical structures in which an intervention is devised cannot automatically be assumed to be comparable to target localities in causally relevant aspects (assuming we knew what these were). Differences in institutional, psychological and physical factors yield different causal and probabilistic relations. Sweden and the USA, for example, have radically different ways of conceptualizing and responding to antisocial behaviour among young people. The examples cited of California class-size reduction and MST illustrate that we need much more information to jump from 'it works somewhere' to 'it will work for us'.

The concept of external validity is inadequate for the task because it assumes that the same facts observed in the study will occur in the target and this is rarely plausible in health and social care contexts. Stable capacities are an alternative. A factor T that has a relatively stable tendency to promote O makes a fixed contribution towards O in varied situations but this can, indeed generally does, differ from what occurs when the factor is present. Other factors may neutralize or enhance the positive effects of T so results can vary. While this looks a potentially more constructive way to evaluate adopting interventions with some RCT support, it is not simple. The methodology for establishing tendency claims is not laid out with anything like the completeness or rigor we have for establishing it-works-somewhere claims, using, for instance, method-of-difference studies. Claims that a factor T has a stable tendency cannot be tested in isolation. Research has to identify *how* T operates to promote O; what must be in place for T to operate to promote O; what can destroy or overwhelm T's operation; what other factors promote or retard O; and what happens when many factors are at work simultaneously. Ultimately, we need theory to judge which factors have stable capacities and to hypothesize when they are worth implementing.

The MST dispute illustrates the limitations of standard approaches to weighing evidence. MST has been subjected to several rigorous RCTs but these produced varied results. For the policy maker considering whether to implement it, the current situation is bewildering. There certainly is evidence that 'it works somewhere' but should the policy maker risk adopting it? Experiences in Canada and Sweden tell against this, whereas that of Norway favours it. The dispute cannot be settled by a power struggle or by conducting more RCTs to see if the balance of success to failure shifts. Decision makers need specific information to help judge whether it will be successful in their particular social context. This requires much more theoretical understanding of how MST operates and what factors in the environment help or hinder it. More RCTs and other methods able to establish it-works-somewhere claims alone will not settle these questions. A different kind of research, testing hypotheses about the role of other factors, is required to build a detailed picture of the circumstances under which MST is a valuable intervention.

Acknowledgements

We would especially like to thank Sophia Efstathiou for her help in editing and the AHRC project, 'Choices of Evidence: Tacit philosophical assumptions in the debates within the Campbell Collaboration', for support for this research.

References

1. Cartwright, N. D. (2007) Hunting Causes and Using Them: Approaches in Philosophy and Economics. New York: Cambridge University Press.
2. Reiss, J. (2007) Error in Economics: The Methodology of Evidence-Based Economics. London: Routledge.
3. Holland, P. W. & Rubin, D. B. (1988) Causal inference in retrospective studies. *Evaluation Review*, 12, 203–231.
4. Epstein, S. (2007) Inclusion: The Politics of Difference in Medical Research. Chicago, IL: Chicago University Press.
5. Cartwright, N. D. (1989) Nature's Capacities and Their Measurement. New York: Oxford University Press.
6. Cartwright, N. D. (2007) Causal powers: What are they? Why do we need them? What can and cannot be done with them? In Dissent in

Science Project Discussion Paper Series (ed. D. Fennell). London: Centre for Philosophy of Natural and Social Science, LSE.
7. Harré, R. & Madden, E. H. (1975). Causal Powers. Totowa, NJ: Rowman and Littlefield.
8. Mumford, S. (1998) Dispositions. Oxford: Oxford University Press.
9. Chakravarty, A. (2007) A Metaphysics for Scientific Realism: Knowing the Unobservable. Cambridge: Cambridge University Press.
10. Bird, A. (2007) Nature's Metaphysics: Laws and Properties. Oxford: Oxford University Press.
11. Bohrnstedt, G.W. & Stecher, B.M. (eds) (2002). What We have Learned about Class Size Reduction in California. Sacramento, CA: California Department of Education.
12. Mill, J. S. (1836 [1967]) On the Definition of Political Economy and on the Method of Philosophical Investigation in that Science, reprinted in Collected Works of John Stuart Mill, Vol. IV. Toronto: University of Toronto Press.
13. Cartwright, N. D. (1999) The Dappled World: A Study of the Boundaries of Science. Cambridge: Cambridge University Press.
14. Gigerenzer, G., Todd, P. M. & the ABC Group (1999) Simple Heuristics that Make Us Smart. New York: Oxford University Press.
15. MST Website, MST Services (2009) *Treatment model*. Available at: http://www.mstservices.com/mst_treatment_model.php (last accessed 2 September 2009).
16. Littell, J., Popa, M. & Forsythe, B. (2007) Multisystemic therapy for social, emotional, and behavioural problems in youth aged 10–17. *The Cochrane Collaboration*, Issue 4.
17. Sundell, K., Hansson, K., Lofholm, C., Olsson, T., Gustle, L.-H. & Kadesjo, C. (2008) The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology*, 22 (3), 550–560.
18. Hengeller, S. (2003) *Multisystemic therapy: an overview: dissemination, data and direction*. NASMHPD Research Institute Conference, February 2003.
19. MST Services (2009) *Organizational biography*. Available at: http://www.mstservices.com/organizational_biography.php (last accessed 2 September 2009).
20. Schoenwald, S., Sheidow, A., Letourneau, E. & Liao J. (2003) Transportability of multisystemic therapy: evidence for multi-level influences'. *Mental Health Services Research*, 4, 223–239.
21. MST Services (2009) *Index*. Available at: http://www.mstservices.com/index.php (last accessed 2 September, 2009).