Philos Stud (2010) 147:59–70
DOI 10.1007/s11098-009-9450-2

What are randomised controlled trials good for?

Nancy Cartwright

Published online: 1 October 2009

Abstract Randomized controlled trials (RCTs) are widely taken as the gold standard for establishing causal conclusions. Ideally conducted they ensure that the treatment 'causes' the outcome—in the experiment. But where else? This is the venerable question of external validity. I point out that the question comes in two importantly different forms: Is the specific causal conclusion warranted by the experiment true in a target situation? What will be the result of implementing the treatment there? This paper explains how the probabilistic theory of causality implies that RCTs can establish causal conclusions and thereby provides an account of what exactly that causal conclusion is. Clarifying the exact form of the conclusion shows just what is necessary for it to hold in a new setting and also how much more is needed to see what the actual outcome would be there were the treatment implemented.

Keywords Randomized controlled trials (RCTs) · External validity · Probabilistic theory of causality · Causal inference · Capacities · Contributions

1 Introduction

Randomized controlled trials (RCTs) are now the gold standard for causal inference in medicine and are fast becoming the gold standard in social science as well, especially in social policy. But what exactly does an RCT establish? To answer this question I turn to work from long ago by Suppes (1970), Skyrms (1980), myself

N. Cartwright (✉)
Centre for the Philosophy of Natural and Social Sciences, London School of Economics, Houghton Street, WC2A 2AE London, UK
e-mail: n.l.cartwright@lse.ac.uk

N. Cartwright
Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0119, USA

(Cartwright 1983, 1989) and others on the probabilistic theory of causality.[1] Given this theory plus a suitable definition of an ideal RCT, it is possible to prove trivially that from positive results in an RCT a causal conclusion can be deduced.

In the social sciences it is usual to talk about experiments or studies in terms of their *internal* versus *external validity*. A study that is internally valid is one that confers a high probability of truth on the result of the study. For instance RCTs are designed to establish causal conclusions and in the ideal, the design itself ensures that a positive result in the experiment confers a high probability on the causal conclusion. External validity has to do with whether the result that is established in the study will be true elsewhere.

I believe that the language of external validity obscures some important distinctions, distinctions that matter significantly when RCT results are offered in evidence that the cause established in the experiment will produce the desired effect in some target situation. As I see it there are two distinct issues conflated into one: 1. Do the RCT results travel *as RCT results* to the target situation? 2. What relevance has an RCT result in the target situation for predicting what will happen there when the cause is implemented, as it will be, without the experimental constraints and in an environment where many other causes may be at work as well? Seeing the roots of RCTs in the probabilistic theory of causality will help make these distinctions clear and reveal the strong assumptions that must be defended if external validity is to be warranted.

2 The probabilistic theory of causality

There is considerable contention about exactly how to formulate the probabilistic theory of causality. The fundamental idea is that probabilistic dependencies must have causal explanations. Take proper account of all the reasons deriving from an underlying causal structure that C and E might be probabilistically dependent. Then C and E are related as cause and effect just in case they are probabilistically dependent. The probabilistic theory takes account of the other possible causal reasons for C and E to be dependent by conditioning on some set of specially selected factors, which in my case was supposed to be a full set of causes of E (simultaneous with C or earlier than C) other than C itself. In the case of dichotomous variables, which is all I shall consider here for simplicity, this leads to the following formula: For an event-type C temporally earlier than event-type E

$$\text{C causes E iff } P(E/C\&K_i) > P(E/\neg C\&K_i).$$

Here K_i is a state description[2] over the specially selected factors. You will notice that K_i is dangling on the right-hand-side of this formula, making it ill formed. I shall return to this below.

[1] There are other routes to RCTs from other accounts of causality. The counterfactual account is noteworthy here (see especially Cartwright (1989) and Rubin (1974) and discussions thereof), but the link there requires more heroic assumptions than the link from the probabilistic theory.

[2] A state description over factors $A_1, \ldots, A_n$ is a conjunction on n conjuncts, one for each A_i, with each conjunct either A_i or $\neg A_i$.

The idea behind the use of the partial conditional probability is that any dependencies between C and E not due to a direct causal link between them must instead be due to a correlation between both C and E and some further factor, often called a *confounding factor*. Conditioning on these confounding factors will break the correlation between them and anything else; any remaining dependencies between C and E must then be due to a direct causal link between them. This is a standard procedure in the social sciences in testing for causality from observational data—by 'stratifying' before looking for dependencies. Formally this depends on what is called *Simpson's paradox*: A probabilistic dependency (independency) between two factors in a population may turn into a probabilistic independency (dependency) within each subpopulation partitioned along the values of a factor that is probabilistically dependent on the two original.[3]

Skyrms's proposal is the most directly responsive to this idea. He argued that the set of selected factors to condition on should include all and only factors with a temporal index prior to or simultaneous with C that are probabilistically dependent on E (Skyrms 1980). I maintained that Skyrms's proposal would not catch enough factors (Cartwright 1983). Wesley Salmon had argued that a cause can decrease the probability of its effect using an example in which a strong cause and a weak cause were anticorrelated: Whenever the weak cause was present the strong cause was absent so that the probability of the effect went down whenever the weak cause was present (Salmon et al. 1971). One need only adjust the numbers to construct a case in which the probability of the effect is the same with the weak cause as with the strong cause. In cases like this the effect will be probabilistically dependent on neither the weak cause nor the strong cause. So neither will appear in the list of selected factors to condition on before looking for a dependency between the other and the effect and thus neither will get counted a cause under Skyrms's proposal.[4]

The only solution I have ever been able to see to this problem is to require that the selected factors for conditioning on before looking for dependencies between C and E be a full set of causal factors for E other than C, where what constitutes 'a full set' is ticklish to define.[5]

My proposal of course is far less satisfactory than Skyrms's. First, it uses the notion of causality on the right-hand-side in the characterisation and hence the characterisation cannot provide a reductive definition for causation. Second, a direct application of the formula seems to require a huge amount of antecedent causal knowledge before probabilistic information about dependencies between C and E can be used to determine if there is a causal link between them. The RCT is designed specifically to finesse our lack of information about what other causes can

[3] See discussion in Cartwright (1983).

[4] The assumption that causes and effects are always probabilistically dependent is sometimes called 'faithfulness' in the causal Bayes-nets literature. Some authors argue that violations are rare. Others—like Kevin Hoover and me—argue the contrary: many systems are designed or evolved to ensure causes cancel. I shall not enter this debate here, however. For further discussion and references, see Cartwright (2007).

[5] See the definitions in Cartwright (1989, p. 112) and in Cartwright (2007, 1976, p. 64, footnote 8). The task is relatively easy if one can take the notion of a causal path as primitive. In that case a full set of causes of E contains exactly one factor from each path into E.

affect E. Before turning to that, however, we need some further consideration of this formula.

What it is to be done about the dangling K_i? There are two obvious alternatives. The first is to put a universal quantifier in front: for all i. This means that we will not say that C causes E unless C raises the probability of E in every arrangement of confounding factors. This makes sense just in case the cause exhibits what John Dupre called *contextual unanimity*: The cause either raises, lowers or leaves the same the probability of the effect in every arrangement of confounding factors (Dupré 1984). Where contextual unanimity fails, it is more reasonable to adopt the second alternative: relativize the left-hand-side causal claim to K_i:

Probabilistic causality: C causes E in K_i iff $P(E/C\&K_i) > P(E/\neg C\&K_i)$ and for any population A, C causes E in A iff C causes E in some K_i that is a subset of A.[6]

This allows us to make more specific causal judgements. It also allows us to say that C may both cause E and prevent E (say, cause ¬E) in one and the same population, as one might wish to say about certain anti-depressants that can both heighten and diminish depression in teenagers. It is especially important when it comes to RCTs where the outcomes average over different arrangements of confounding factors so that the cause may increase the probability of the effect in some of these arrangements and decrease it in others and still produce an increase in the average.

Over the years I, along with others, have noticed a number of other problems with this formula:

- When a confounding factor D can be produced by C in the process of C's producing E but can also occur for independent reasons, D should be conditioned on just in the cases where D is not part of the causal process by which C produces E (Cartwright 1989).
- When a probabilistic cause produces two effects in tandem, the effects will be dependent on each other even once the joint cause has been conditioned on. In this case the conditioning factors for deciding if C causes E need to include a dummy variable that takes value 1 just in case C has operated to produce the paired effect and the value 0 otherwise (Cartwright 1989, 2007).
- If a common effect of two separate causes is 'over represented' in the population the two causes for that the effect will typically be probabilistically dependent. This means that the selected factors for conditioning on must not include common effects like this—so we must not condition on too much.
- Sometimes quantities are probabilistically dependent with no causal explanation. The one widely recognized case of this is when two quantities both change monotonically in time. Say they both increase. Then high values of one will be probabilistically dependent on high values of the other. Vice versa if they both decrease. And if one increases and the other decreases, high values of one will be dependent on low values of the other.
- A standard solution to this problem in practice is to detrend the data. This involves defining two quantities whose values at any time are essentially the

[6] It need not be a proper subset.

values of the original quantities minus the change due to trend. This does not rescue the formula for probabilistic causality, however, unless we want further elaboration: *If* there is a dependence between C and E due to trend, then C causes E iff $P(E'/C'\&K_i) > P(E'/C'\&\neg K_i)$, where E', C' are new quantities defined by detrending C and E. The trick of course is to know when to detrend and when not, since a correlation in time between two monotonically changing quantities can always be due to one causing the other.

- One and the same factor may both cause and prevent a given effect by two different paths. If the effect is equally strong along both paths, the effect will not be probabilistically dependent on the cause. A standard solution in practice in this case is to condition on some factor in each of the other paths in testing for a remaining path. Again, a direct application of this strategy requires a great deal of background knowledge.

Given these kinds of problems, how should the formula be amended? I think the only way is by recognizing that at this very general level of discussion we need to revert to a very general formulation. We may still formulate the probabilistic theory in the same way, but now we must *let K_i designate a population in which all other reasons that account for dependencies or independencies between C and E have been properly taken into account.*

Nor should we be dispirited that this seems hopelessly vague. It is not vague but general. Once a specific kind of causal structure has been specified, it is possible to be more specific about exactly what features of that causal structure can produce dependencies and independencies.[7]

3 RCTs and the probabilistic theory of causality

RCTs have two wings—a treatment group of which every member is given the cause under test and a control group, where any occurrences of the cause arise 'naturally' and which may receive a placebo. In the design of real RCTs three features loom large:

a. *Blindings* of all sorts. The subjects should not know if they are receiving the cause or not; the attendant physicians should not know; those identifying whether the effect occurs or not in an individual should not know; nor should anyone involved in recording or analyzing the data. This helps ensure that no differences slip in between treatment and control wings due to differences in attitudes, expectations or hopes of anyone involved in the process.
b. *Random assignment* of subjects to the treatment or control wings. This is in aid of ensuring that other possible reasons for dependencies and independencies between cause and effect under test will be distributed identically in the treatment and control wings; this helps deal not only with 'other' causal factors

[7] For examples, see Pearl's linear causal structures (Pearl 2000) or my representations for structures in which causes act irreducibly probabilistically (Cartwright 1989).

of E but also with the other specific problems I mentioned for formulating the probabilistic theory at the end of Sect. 2 except for the last.

c. Careful choice of a *placebo* to be given to the control, where a placebo is an item indiscernible for those associated with the experiment from the cause except for being causally inert with respect to the targeted effect. This is supposed to ensure that any 'psychological' effects produced by the recognition that a subject is receiving the treatment will be the same in both wings.

These are all in aid of bringing the real RCT as close as possible to an *ideal RCT*. Roughly, an RCT is ideal iff all factors that can produce or eliminate a probabilistic dependence between C and E are the same in both wings except for C, which each subject in the treatment group is given and no-one in the control wing is given, and except for factors that C produces in the course of producing E, whose distribution differs between the two groups only due to the action of C in the treatment wing. An outcome in an RCT is *positive* if P(E) in the treatment wing >P(E) in the control wing.

As before, designate state descriptions over factors in the experimental population that produce or eliminate dependencies between C and E by K_i. In an ideal RCT each K_i will appear in both wings with the same probability, w_i. Then P(E) in treatment wing $= \sum w_i P(E/K_i)$ in treatment wing and P(E) in control wing $= \sum w_i P(E/K_i)$ in control wing. So a positive outcome occurs only if $P(E/K_i)$ in treatment wing $>P(E/K_i)$ in control wing for some K_i. This is turn can only happen if $P(E/C\&K_i) > P(E/\neg C\&K_i)$ for some i. So *a positive outcome in an ideal RCT occurs only if C causes E in some K_i by the probabilistic theory of causality*.[8]

The RCT is neat because it allows us to learn causal conclusions without knowing what the possible confounding factors actually are. By definition of an ideal RCT, these are distributed equally in both the treatment and control wing, so that when a difference in probability of the effect between treatment and control wings appears, we can infer that there is an arrangement of confounding factors in which C and E are probabilistically dependent and hence in that arrangement C causes E because no alternative explanation is left. It is of course not clear how closely any real RCT approximates the ideal. I will not go into these issues here, however, despite their importance.[9]

Notice that a positive outcome does not preclude that C causes E in some subpopulation of the experimental population and also prevents E in some other. Again, certain anti-depressants are a good example here. They have positive RCT results and yet are believed to be helpful for some teenagers and harmful for others.[10]

[8] As above, this argument sketch can be filled in more precisely once a particular kind of underlying causal structure is supposed. (The argument also supposes that the causal structure is the same in both treatment and control wings.).

[9] But see for instance Altman (1996) and Worrall (2002).

[10] See for instance the U.S. Food and Drug Administration medication guide at www.fda.gov/cder/drug/antidepressants/SSRIMedicationGuide.htm.

4 Causal principles: from experimental to target populations

There are two immediate models for exporting the causal conclusion of an RCT to a new situation involving a new population. Reconstructing from the suppositions made and the surrounding discussion across a number of cases in different fields, I would say both fit common practice, which more often than not seems a mish-mash of the two. One is the physics model that I have developed in my work on capacities (Cartwright 1989): Suppose the cause has some (relatively) invariant capacity; that is, the cause always makes some fixed contribution that affects the final outcome in a systematic way. For example, the gravitational attraction on a mass m associated with a second mass M always contributes an acceleration to m in the direction of M of size $GM/\mathbf{r}^2$, and this always adds vectorially with the contribution of other sources of acceleration acting on m.

I have been resurrecting my older work on capacities recently because something like this model often seems to be assumed, albeit implicitly, in the use of RCTs as evidence for predicting policy outcomes. RCTs are treated much like what I call 'Galilean experiments'. These strip—or calculate—away all 'interfering' factors. The idea is that what the cause produces on its own, without interference is what it will *contribute* elsewhere. The experiment measures the contribution assuming there *is* a contribution to measure; that is, assuming there is a stable result that contributes in the same systematic way across broad ranges of circumstances. To know that takes a huge amount of further experience, experiment, and theorizing. RCTs can measure the (average) contribution a treatment makes—if there is a stable contribution to be measured. But they are often treated as if the results can be exported in the way that Galileo's results could, without the centuries of surrounding work to ensure there is any stable contribution to be measured. Since I have written in detail about RCTs and capacities elsewhere (Cartwright, forthcoming a, b), I will not pursue the topic here. I bring it up only because the logic of capacities may be implicated in the second model as well, as I shall explain in Sect. 5.

The second model exports the causal principle established in the experimental population A directly to the new population A' in the new situation. Under what conditions is this inference warranted? We shall need a major modification later but right now it seems reasonable to propose:

Preliminary rule for exporting causal conclusions from RCTs. If one of the K_i that is a subset of A such that C causes E in K_i is a subset of A', then C causes E in A implies C causes E in A' under the probabilistic theory of causality.

But how do we know when the antecedent of this rule obtains? Recall, the beauty of the RCT is that it finesses our lack of knowledge of what exactly the confounding factors are that go into the descriptions $K_1, \ldots, K_m$. So in general we do not even know how to characterise the various K_i let alone know how to identify which K_i are the ones where C is causally positive, let alone know how to figure out whether that description fits some subpopulation of the target population A'.

In some situations matters are not so bad:

- If the experimental population is a genuinely representative sample of the target, then all the weights w_i will be the same in the target and the experimental

population. Any uncertainty about whether it is truly representative will transfer to the antecedent of our rule unless there is otherwise good reason to think the specific K_i's in which the causal principle holds are present in the target.

- If the cause is contextually unanimous, the antecedent clearly applies. But since many causes are not contextually unanimous, this cannot be assumed without argument.

There is another worry at a more basic level that I have so far been suppressing. Two populations may both satisfy a specific description K_i but not be governed by the same causal principles. When we export a causal principle from one population satisfying a description K_i to another that satisfies exactly the same description, we need to be sure that the two populations share the same causal structure. What is a causal structure? I have argued that there are a variety of different kinds of causal structures at work in the world around us; different causal structures have different formal characterizations (Cartwright 1999, 2007). Given the probabilistic theory of causality clearly what matters is this:

Two populations share the '*same causal structure with respect to the causal principle "C causes E"*' from the point of view of the probabilistic theory of causality iff the two populations share the same reasons for dependencies to appear or disappear between C and E (i.e. the same choice of factors from which to form the state descriptions K_i) and the same conditional probabilities of E given C in each.[11]

Sometimes this seems easy. I have heard one famous advocate of RCTs insist that for the most part Frenchmen are like Englishmen—we do not need to duplicate RCTs in France once we have done them in England. He of course realizes that this may not be true across all treatments and all effects. And we know that superficial similarities can be extremely misleading. I carry my pound coins from Britain to the US and put them into vending machines that look identical to those in Britain, but they never produce a packet of crisps for me in the US.

Clearly the rule of export needs amendment:

Rule for exporting the causal conclusion C causes E from an RCT. If populations A and A' have the same causal structure relative to "Causes E" and if one of the K_i that is a subset of A such that C causes E in K_i is a subset of A', then C causes E in A implies C causes E in A' under the probabilistic theory of causality.

The lesson to be learned is that although (ideal) RCTs are excellent at securing causal principles, there is a very great deal more that must be assumed—and defended—if the causal principles are to be exported from the experimental population to some target population. Advice on this front tends to be very poor indeed however. For instance the US Department of Education website teaches that two successful well-conducted RCTs in 'typical' schools or classrooms 'like yours' are 'strong' evidence that a programme will work in your school/classroom (U.S. Department of Education 2003). The great advantage of a formal treatment is that it can give content to this uselessly vague advice. From the point of view of the probabilistic theory of causality, 'like yours' must mean

[11] Or at least the same facts about whether the probability of E is greater conditioned on C as opposed to ¬C.

- Has the same causal structure
- Shares at least one K_i subpopulation in which the programme is successful.

Unfortunately, these two conditions are so abstract that they do not give much purchase on how to decide whether they obtain or not. Nevertheless, any bet that a causal principle does export to your population is a bet on just these two assumptions.

5 From causal principles to policy predictions

Consider a target population A for which we are reasonably confident that the causal principle 'C causes E in A' obtains. How do we assess the probability that E would result if C were introduced? Let's take a nice case first. Suppose we have tested for 'C causes E' in a very good RCT where the experimental population was collected just so as to make it likely that it was representative of the target. Then we can assume that P(E) if C were introduced will equal $\sum w_i P(E/K_i)$ in treatment wing $=P(E)$ in the treatment wing. Or can we? Not in general. We can if all three of the following assumptions are met:

- C and C alone (plus anything C causes in the process of causing E) is changed under the policy.[12]
- C is introduced as in the experiment—the C's introduced by policy are not correlated with any other reasons for probabilistic dependencies and independencies between C and E to appear or disappear in the target.
- The introduction of C's leaves the causal structure unchanged.

These are heavy demands.

If the RCT population is not a representative sample of the target matters are more difficult. Besides the three assumptions above we need to worry about whether the target contains the subpopulations in which C is causally positive before it will be true at all that C causes E there. That however does not ensure that introducing C, even as described in our three assumptions, will increase the probability of E since the target may also contain subpopulations where C is causally negative, and these may outweigh the positive ones. If C is contextually unanimous with respect to E this concern disappears. But, to repeat the earlier warning, contextual unanimity is not universal and a lot of evidence and argument are required to support it.

Finally, it is usual in policy settings to violate all three of the above conditions. Implementations usually change more than just the designated causal factor [e.g. in California, when class sizes were reduced, teacher quality also went down because there were not enough qualified teachers for all the new classes (Blatchford 2003)]; the changes themselves are often correlated with other factors [e.g. people who take up job training programmes may tend to be those already more prone to benefit from them (Heckman 1991)]; introducing the cause may well undermine the very causal

[12] This is a special case of the next requirement; I write it separately to highlight it.

principle that predicts E will result (e.g. the Chicago School of economics maintained that this is typical in economic policy[13]).

At least with respect to the first two of these worries I notice that there is a tendency to begin to think in terms of my physics model of capacities. Even if more than C is changed at least we can rely on the influence of C itself to be positive. In this case we need to pay close attention only to changes that might introduce strongly negative factors (as in the California class-size reduction programme) or undermine the causal structure itself (like banging on the vending machine with a sledge hammer).

The assumption here is a strong version of contextual unanimity: Whatever context C is in, it always contributes positively. It is stronger even than the cases of capacities I have studied, where the assumption is that whatever context C is in, it always makes the *same* contribution that affects the results in *the same systematic way*. Consider the vector addition of accelerations contributed by various different causes, described in Sect. 3. We are so familiar with vector addition that we sometimes forget that it is not the same as the simple linear addition supposed in saying that if C causes E in some K_i then C will always contribute positively. After all, a magnet pulling up will decrease the acceleration of a falling body not increase it. In any case, the point is that the assumption that at least C itself contributes positively is a strong one, and like all the rest, needs good arguments to back it up.

6 In sum

RCTs establish causal claims. They are very good at this. Indeed, given the probabilistic theory of causality it follows formally that positive results in an ideal RCT with treatment C and outcome E deductively implies 'C causes E in the experimental population'. Though the move from the RCT to a policy prediction that C will cause E when implemented in a new population often goes under the single label, the *external validity* of the RCT result, this label hides a host of assumptions that we can begin to be far clearer and more explicit about.[14] To do so, I think it is useful to break the move conceptually into two steps.

First is the inference that 'C causes E in the target population' or in some subpopulation of the target. The probabilistic theory of causality makes clear, albeit at a highly abstract level, just what assumptions are required to support this move. Unfortunately, they are very strong assumptions so one must make this move with caution.

Given the interpretation of the causal claim supplied by the probabilistic theory, this first step is essentially a move to predict what would happen in a new RCT in the target population, or a subpopulation of it, from what happens in an RCT in a different population. If we follow standard usage and describe RCT results as *efficacy* results, the inference here is roughly from efficacy in the experimental population to efficacy in the target. That is a far cry from what is often described as

[13] Cf. the famous Lucas critique of policy reasoning (Lucas 1976).

[14] For further discussion see also Cartwright (1976) and Cartwright and Efstathiou (2007).

an *effectiveness* result for the target: a claim that C will actually result in E when implemented there.

This comprises the second step: the move from 'C causes E in the target population' to 'C will result in E if implemented in this or that way'. Julian Reiss and I have argued, both jointly and separately (Reiss 2007; Cartwright 1999), that the best way to evaluate effectiveness claims is by the construction of a causal model, where information about the behaviour of C in an RCT is only one small part of the information needed to construct the model. We need as well a great deal of information about the target, especially about the other causally relevant factors at work there, how they interact with each other and with C, and how the existing causal structure might be shifted during policy implementation. I have not rehearsed this argument here but rather, in keeping with my starting question of what RCTs can do for us vis-à-vis policy predictions, I have laid out some assumptions that would allow a more direct inference from 'C causes E in the experimental population' to 'C will result in E if implemented in this or that way'. Again, these are very strong and should be accepted only with caution.

Throughout I have used the probabilistic theory of causality and at that, only a formulation of the theory for dichotomous variables.[15] This is not the only theory of causality making the rounds by a long shot. But it is a theory with enough of the right kind of content to show just why RCTs secure internal validity and to make clear various assumptions that would support external validity. Something similar can be reconstructed for the counterfactual theory and for Judea Pearl's account that models causal laws as linear functional laws, with direction, adding on Bayes-nets axioms (Pearl 2000). A good project would be to lay out the assumptions for various ways of inferring policy predictions from RCTs on all three accounts, side-by-side, so that for any given case one could study the assumptions to see which, if any, the case at hand might satisfy—remembering always that if causal conclusions are to be drawn, it is important to stick with the interpretation of the conclusion supplied by the account of causality that is underwriting that conclusion!

My overall point, whether one uses the probabilistic theory or some other, is that securing the internal validity of the RCT is not enough. That goes only a very short way indeed towards predicting what the cause studied in the RCT will do when implemented in a different population. Of course all advocates of RCTs recognize that internal validity is not external validity. But the gap is far bigger than most let on.

Acknowledgements I would like to thank Chris Thompson for his help, the editors and referees for useful suggestions, and the Spencer Foundation and the UK Arts and Humanities Research Council for support.

[15] I personally am not happy with any extant *general* formulations for multivalued or continuous variables. That is connected with my view that a proper formulation must be relativised to a specific kind of system of causal laws; for different systems, different formulations will be appropriate.

References

Altman, D. G. (1996). Editorials: Better reporting of randomised controlled trials: The CONSORT statement. *British Medical Journal, 313*, 570–571.

Blatchford, P. (2003). *The class size debate: Is smaller better?* Philadelphia: Open University Press.

Cartwright, N. (1983). *How the laws of physics lie*. New York: Oxford University Press.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. New York: Cambridge University Press.

Cartwright, N. (forthcoming 'a'). 'What is this Thing Called 'Efficacy'?' In C. Mantzavinos (Ed.), *Philosophy of the social sciences. Philosophical theory and scientific practice*, Cambridge University Press.

Cartwright, N. (forthcoming 'b'). Evidence-based policy: What's to be done about relevance?' *A talk given at Oberlin College Colloquium April 2008.*

Cartwright, N. & Efstathiou, S. (2007, August). *Hunting causes and using them: Is there no bridge from here to there*? Paper presented at the First Biennial conference of the Philosophy of Science in Practice, Twente University.

Dupré, J. (1984). Probabilistic causality emancipated. *Midwest Studies in Philosophy, 9*, 169–175.

Heckman, J. J. (1991) Randomization and social policy evaluation, *NBER Working Paper* No. T0107.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie Rochester Conference Series on Public Policy, 1*, 19–46.

Pearl, J. (2000). *Causality: Models reasoning and inference*. Cambridge: Cambridge University Press.

Reiss, J. (2007). *Error in economics: The methodology of evidence-based economics*. London: Routledge.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Salmon, W., Jeffrey, R., & Greeno, J. (Eds.). (1971). *Statistical explanation and statistical relevance*. Pittsburgh: Pittsburgh University Press.

Skyrms, B. (1980). *Causal necessity*. New Haven, USA: Yale University Press.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.

U.S. Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide* http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf. Accessed 29 August 2008.

Worrall, J. (2002). What evidence in evidence-based medicine. *Philosophy of Science*, Vol. 69, No.3, In Supplement: *Proceedings of the 2000 biennial meeting of the philosophy of science association*. Part II: Symposia papers, pp. S316–S330.