Reflections on RCTs

Angus Deaton and Nancy Cartwright

April 19, 2018

Even in our very long paper, we could not discuss all important issues about the conduct and use of RCTs in contemporary social science and medicine, and we are grateful for the many comments that add to and expand on what we have written. We take up a few of the omitted issues in these comments. We also want to correct a few serious misrepresentations of what we wrote, in case acquiescence is mistaken for consent. Beyond that, we have looked for critical objections that suggest mistakes in our arguments or ways to improve on our claims. We are grateful to the commentators for inspiring us and the editors for allowing us to respond.

One important comment, repeated—but not unanimously—can perhaps be summarized as 'All that said and done, RCTs are still generally the best that can be done in estimating average treatment effects and in warranting causal conclusions.' It is this claim that is the monster that seemingly can never be killed, no matter how many stakes are driven through its heart. We strongly endorse Robert Sampson's statement "That experiments have no special place in the hierarchy of scientific evidence seems to me to be clear." Experiments are sometimes the best that can be done, but they are often not. Hierarchies that privilege RCTs over ***any*** other evidence irrespective of context or quality are indefensible and can lead to harmful policies. Different methods have different relative advantages and disadvantages.

First, and while it is true that some of the issues we raise are also faced by other members of what Issa Dahabreh labels "the broader class of experimental comparative studies," advocates of these other studies are often more explicit in assessing whether the assumptions required for them to deliver correct results are likely to be satisfied. Second, many methods for causal inference about groups or individuals are not comparative at all, such as causal Bayes nets methods, causal structural modelling, examination and testing of empirical propositions, derivation from theory, case studies, or process tracing, though of course each method comes with its own weaknesses and strengths. We can often learn much about causality and about mechanisms without employing Mill's (1882) "Method of Difference." Third, when it comes to putting study results to use, whether for prediction about specific individuals or specific populations or for providing evidence for more general claims, other kinds of study designs may make far bigger contributions. We take up these issues in our first several points below, then move on to other comments.

1. Michael Oakes concludes his comments by saying that a well-done RCT is better than a p-hacked observational study, and of course it is. But we do not agree with Oakes' implicit message nor with what Etsuji Suzuki and Tyler VanderWeele explicitly say when they write "other methods have all of those problems, plus problems of their own," so that an RCT is generally better than ***any*** other study. Each case must be judged on its merits. A well-done RCT is better than a badly done RCT, or an observational study that thinks of itself as a wannabe RCT but without sufficient warrant for the causal assumptions needed to make it valid. There seems to be an idea that ***all*** observational, or non-experimental, studies are wannabe RCTs, or failed RCTs, so that, by definition, the real RCT is trivially better than the fake RCT. Of course, we agree. But that is not what we are talking about. There are many kinds of observational studies with different methods. An RCT is of necessity done on a special population, which is nearly always selected by the requirement that subjects must be available for the trial and agree to participate; it is an irony that we are repeatedly reminded that only an RCT can deal with selection, yet RCTs are ***almost always*** run on selected, special, populations. As Judea Pearl says, RCTs are done in artificial environments that often bear little resemblance to the target intervention. Those

who participate in a training program are different—had previous wage declines—than those who do not. Their ATE is likely to be different. Is an RCT on selected people relevant for another population? Is an RCT on training for women likely to work for men? A health insurance experiment is done on a test population, but if scaled up, it applies to the whole population.

There are two formal reasons for denying that general evidence rankings like these make sense, however cautiously put. First, all methods require assumptions to be met; methods for drawing ***causal*** conclusions require ***causal*** assumptions (no causes in, no causes out). For many methods, RCTs among them, it is ***provable*** that, if their accompanying assumptions are met, then a causal conclusion follows. For RCTs, the central assumptions are that the causal possibilities in the population studied can be represented in a potential outcomes equation, like our equation (1); that the study treatment is orthogonal to all causes of the outcome other than its own downstream effects; and assumptions about the shape of the underlying distribution of individual treatment effects in the study population. Other methods need other assumptions, some of which we recall in 13. in discussing Andrew Jones and Daniel Steel's contribution. So, to assume that RCTs are generally more reliable is to assume that their assumptions are generally more often met (or meetable) than those of other methods. What justifies that? Especially given that the easiest assumption to feel secure about for RCTs – that the assignment is done 'randomly' – is far from enough to support orthogonality (which we discuss in 2.), which is itself only one among the assumptions that need to be supported. Even if it turned out that the RCT assumptions were often easier to meet than assumptions for other methods, surely the sensible thing is to look at the matter case-by-case.

Second, there is a concern about the form of the causal conclusion. When we want to draw a conclusion about what holds anywhere except in the study population, RCTs have no privilege at all. They show that the treatment worked somewhere, usually a very special somewhere (often made more special by the stringent requirements of doing a good RCT) and it is a long and often difficult evidential road from that to 'It will work here'. And for this endpoint, the RCT is not a particularly natural starting point. Indeed, this is one of the misunderstandings that we are most concerned about, that a well-done RCT can be automatically transported, simply by virtue of being an RCT.

2. Many of the commentators sing the praises of randomization. But what is needed for unbiasedness is not randomization, but ***orthogonality***, which, as we stressed, is not guaranteed by randomization in and of itself. In all cases an ***argument*** is required that orthogonality is likely to have been achieved. We disagree with Guido Imbens notion that "a design involving randomization" simplifies estimation and inference (though many studies seem to believe so), however well or badly the randomization was done, whether it is real randomization or pseudo-randomization, or however long ago it was done. Randomization does not give orthogonality which requires, first, an assurance that the assignment mechanism really was random, as John Ioannidis puts it, "properly generated", and not by, for example, alphabetization or some other convenient 'natural' scheme. Second, problems are even more likely to occur due to what Ioannidis refers to as "post-randomization experiences", where, through lack of blinding, differential attrition, different times, places, lengths of treatment for the two groups, and a host of other problems, correlations with other factors affecting the outcome cannot be ruled out. The fact that an experiment is blinded (which is often not possible and is rare in economics) is a good start, but beyond this there should also be reasons to think the blinding was effective and that no other relevant systematic differences occurred between treatment and report of results, which cannot usually be taken as the default assumption. That orthogonality has been achieved 'well

enough' has to be defended case-by-case. Yet no-one talks much about this, and good arguments about post-randomization policing and exactly what can be inferred when it is lacking are rare in social and in biomedical science. An unblinded or unpoliced experiment might tell us something useful, but it might not, even if it is infinitely large and on the right population. Whether we are warranted in supposing it does so or not depends on what good arguments can be made in the case at hand.

3. Stephen Raudenbush defends randomization on the grounds that it creates groups whose characteristics are "statistically equated by random assignment." We do not know exactly what "statistically equated" means, but if the characteristics were ***actually*** equated—which of course they are not—the estimated ATE would be exactly true, so perhaps being "statistically equated" means the ATE is "exactly statistically true." It is this sort of loose language that creates the protective haze in which estimates from RCTs can be attributed magical properties that they do not possess. Unbiasedness—which is not even guaranteed by randomization, see 2.—is consistent with an estimate of the ATE that is arbitrarily far from the truth. An archer who misses two feet to the right half of the time and two feet to the left half of the time is shooting unbiased arrows but ***never*** hits the target. "Statistically equated" is an excellent example of an apparently precise statement leading to one of the misunderstandings we are trying to dispel. He goes on to says that "non-randomized studies are credible only to the extent that the selection of persons into comparison groups approximates random assignment." This is another classic misunderstanding. As Thomas Cook says, there are "unhappy randomizations" where all the randomly selected plots are on one side of the field, or two thirds of the sample are rich. Why would non-randomization seek to replicate such misfortunes? As we argue in the paper, people ***think*** that

randomization implies balance on observables and unobservables, they confuse randomness with representativeness, but these are not the same thing.

4. We recognize it is impossible to change the use of the terms 'internal validity' and 'external validity', and we do not dispute the familiar definitions by Shadish et al and by Rosebaum that Imbens quotes (as far as they go: the use of these terms goes well beyond the special cases where the study outcome is a cause-effect relation and well beyond studies that involve statistics). But we steadfastly resist the idea that these definitions of 'external validity' presuppose, which is that external validity is a property of a study design or of an estimate, in the sense that an externally valid estimate or causality applies 'as is' elsewhere. [1] 'External validity' is usually taken to mean that 'the same' claim (or estimate) warranted by some study for the set of systems studied holds for other targeted systems, either some specific set of others or even for 'all' others. Except in the rare circumstance where there is good warrant that the study design selects systems to study that are representative of those targeted with respect to the claim in question, this is not a characteristic of study design at all. It depends rather on material facts about the world, which need warrant from outside the study. The ATE or knowledge from an RCT will often be ***useful*** in other settings, but, as always, using it as evidence for claims not about the study population requires knowledge from outside the RCT. Asking that an RCT (or any other

[1] We also dispute Imben's suggestion that Cartwright says something odd in claiming 'economic models have all the advantages when it comes to internal validity but RCTs when it comes to external validity.'. This is a claim about theoretical 'pen-and-paper' models in economics (and similar models in physics) that introduce fictional systems characterized entirely by the premises of the model and that draw conclusions about those systems by deduction. Surely Imbens does not want to deny that deduction is bound to get it right since that's true by the definition of *deduction.* And it hardly seems surprising that facts we have found to be true of real people should have a better chance of being true of some other real people somewhere than do facts established for pen-and-paper fictions in pen-and-paper settngs.

study) be externally valid risks devaluing a useful study because it cannot do something that it was not designed to do and is capable of doing.

We also find it hard to believe Ioannidis' claim that RCTs in medicine do not disagree across settings provided relative treatment effects are used, not least from reading the examples cited by John Concato, but it is certain that no such claim is true in economics or other social sciences. It also illustrates the sort of claim we challenge in the previous paragraph, that RCTs, when done well, automatically apply elsewhere.

The emphasis in the RCT literature and within some of the commentaries is on taking a result well attested for a study population ('R holds for P', as in 'A good estimate of the ATE is X in this trial population'), and keeping the same linguistic form ('R', as in 'A good estimate of the ATE is X') but now ascribing that to a different population ('R holds for Q', as in 'A good estimate of the ATE is X for Q'). Or perhaps, more adventuresomely, making some of the related kinds of inferences that Pearl and Elias Bareinboim's methods allow, including but going beyond some kind of reweighting from one population to another. This, as we said, is to undersell what RCTs can do. Good RCT results, like any 'well established' empirical results, can play a role in evidencing claims that look very unlike the result itself. For instance, Millikan famously measured the ***strength of an electromagnetic field*** pulling up on charged oil drops as evidence about the ***charge of the electron***. In our paper, we give an example where an RCT on fertilizer use among a sample of farmers can be used to help construct an estimate of what fertilizer will do to all farmers together, even though the sign of the effect is reversed in the aggregate.

Of course, that one can do such things depends on a web of other scientific activities, not least of which is the development of useful concepts. But so too does making an 'extrapolation' from one population to another. Ioannidis comments that "Probably the large majority of inferences made from non-randomized data are substantially flawed." We have argued that one should in general not be drawing inferences from any ***single*** set of data in the first place, even about the subjects described in those data. This is important to underline. (For good arguments that one should (almost) always have mechanistic and comparative data for a casual conclusion in medicine (and elsewhere) about even a single population, see Parkkinen et al. (2018) cited in our paper. For a good study of the web of practice, theory, and evidence that goes into establishing credible claims in natural science see Chang (2004); for the human sciences, Wylie (2011).)

We are thus also in disagreement with Basu's claim: "For external validity, we have to use the kind of intuition discussed above. As Banerjee, Chassang and Snowberg (2016) point out, external validity is inherently subjective." This should be resisted, unless we want to judge that natural science results that we use all the time are "inherently subjective". It entirely overlooks the interconnected web of warrant of very different kinds, including experiment, observation, concept development, and theory testing and theory coherence, that makes these, like any scientific claims, credible. In passing, we note for the record that Basu's summary of the first part of our paper, that "when it comes to average treatment effects . . . .RCTs cannot be bettered," is not only wrong but close to being the ***exact opposite*** of what we say. Any doubts on this will be quickly resolved by reading what we wrote.

5. We are not persuaded that well-informed people can judge how to use results in other settings, as some commentators suggest. We think there is a serious problem with claims like these from some of the commentators: "In a new setting and population, drawn at random from the types of settings employed in the meta-analysis, we would expect the causal effect for that population to be between $\boldsymbol{x}$ and $\boldsymbol{y}$ at least 95% of the time" (Suzuki and VanderWeele); "[L]arge sample size in single trials and meta-analysis of multiple trials may decrease the inherent uncertainty about the

estimated true treatment effect" (Ioannidis); and "A more common approach is to describe the sample in enough detail so that informed persons can evaluate the plausibility of extrapolation to similar sub-populations" (Raudenbush). What populations are these remarks about? And similar in what ways? As Dahabreh notes, there are real "challenges inherent in defining the population from which trial participants are sampled." As with all cases of moving anywhere beyond results on the populations actually studied, what types these are requires a great deal of causal knowledge, much of it just the kind RCTs are supposed to help avoid. Casual description of some characteristics that all the study populations have in common is unlikely to pick out populations that have the same underlying causal structure (or some fixed probability mix over some set of different structures), let alone the same average effect of moderator variables (of which the ATE is a function).

We rarely know exactly ***what*** studies show, because there are always many differences between treatments and controls, and just ***saying*** that the thing we care about was the active ingredient doesn't make it so. Cook's concerns with constructs and construct validity address this point. The description given to the treatment is a construct. So too is the description of what the control intervention is, as well as the description of the population enrolled in the study. It takes a web of theory and evidence to support the claims that these descriptions are the 'right' ones – that they refer to the features that are actually causally relevant. Of course, all studies use constructs and all need to worry about construct validity. But most other study designs are not so ardently defended on the grounds of being theory and assumption free. Cook's introduction of this issue is very welcome since it is almost never mentioned in discussions of what can be learned from RCTs and what it takes to learn it.

6. Even if the trial sample is randomly selected from the population of interest, it is not true that grossing up is immediate, as Raudenbush claims. Consider the cocoa growers in our paper. Causal laws CHANGE when we scale up. Re-weighting is fine, but it is not a general answer. SUTVA frequently does not hold, and in many economic cases, SUTVA cannot hold when we scale up ***as a matter of logic***, for example if an intervention changes supply or demand, which must be met from somewhere, by changes in the behavior of other agents not included in the experiment. That is what equilibrium is about. Scaling up releases causal forces that are absent in the experiment, as Sampson and Dahabreh also argue in their comments.

7. With respect to just how to put RCT results to use outside the study setting, we are happy to recommend the work of Pearl and Bareinboim, albeit with caveats. Their general framework supposes that the relations between an effect and its causes can be represented in potential outcomes equations, like our equation (1), with accompanying directed acyclic graphs, like Sampson's "policy graphs". Within this framework—which excludes the simultaneous causality (e.g. of supply, demand, and price) that is an important part of econometric history and economic practice—Pearl and Bareinboim show which similarities and difference between two populations allow which experimental results from one to be used in which ways to calculate different probabilistic and causal facts in the other. This provides a strong formal argument for our point, which, though it should be uncontroversial seems so often to be made light of, that RCT results from a study population ***can*** be of help in predicting information about other populations, but it takes a host of other assumptions to do so, assumptions that themselves need warranting. Nor need the results, when we get there, look much like the experimental results we started from. Pearl and Bareinboim's theorems show just what those assumptions are for specific inferences.

Note that these assumptions include the very kind of information (e.g. about other independent causes, intermediaries, and moderators) that RCTs are supposed not to require. Pearl seems to dispute this: "RCTs are designed to neutralize the confounding of treatments, whereas our methods are designed to neutralize differences between populations. Researchers may be totally ignorant of the structure of the former and quite knowledgeable about the structure of the latter." But it is not just knowledge of some ***differences*** that one needs in order to use the Pearl/Bareinboim scheme. One needs to know that other relevant facts are ***similar***. As they say elsewhere: "[T]he absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is the same in the two populations." (Pearl and Bareinboim, p 588)

Finally, we should note that, though we appreciate the completeness results of Pearl and Bareinboim, our concerns with integration are broader. They can tell us what facts must hold to predict results in new populations. We are concerned with the ***warrant*** one might have for such predictions. This warrant will in general depend on a web of evidence of very different kinds, including high-, low-, and mid-level theory and conceptualization, where different pieces will have different degrees of credibility. This kind of evidence 'synthesis' takes serious thinking, not just recasting in a formal system. It is what arriving at real scientific conclusions is like.

8. The issues that we raise about standard errors are perhaps "technical", as Dahabreh says, citing Imbens, though note Ioannidis' agreement that effective sample sizes may be ten or a hundred times smaller than the apparent sample sizes, which is exactly our argument, and which is why we think that perhaps a majority of the conclusions from RCTs in economic development are likely to be wrong. It is precisely the ability to generate standard errors that is the true role of randomization in RCTs, as Fisher himself well knew; randomizations are indeed often

"unhappy," and estimates of ATEs far from the truth, but the randomization gives us a way of calculating a standard error which gives us a handle on how far we are likely to be off through "unhappy" randomization (though not how far off we are likely to be due to "unhappy" post-randomization experiences). Hence, if the standard errors are not correct, RCTs have no obvious advantages, since unbiasedness by itself is worth little. The use of the word "technical" is dismissive in this context, meaning that it doesn't matter in practice. But "technical problem" can be the last words of the engineers before the bridge crashes down on their heads. There is a history of long-dismissed "technical" issues in econometrics that turned out to conceal deep misunderstandings, such as non-stationarity, spurious correlations, or the exogeniety of instrumental variables (including randomization.) Alwyn Young's work cited in our paper suggests that current experimental practice in economics gets inference wrong a large fraction of the time. Beyond his concerns, it would not be at all surprising if skewness of the kind we discuss in the paper were common in economic experiments, so that the actual degrees of freedom are many times smaller than the nominal degrees of freedom; indeed, this is a plausible explanation for the wild findings in so many RCTs. Moreover, we see no general argument for Dahabreh's claim that the "impact [of these issues] on observational studies should be about the same as on randomized trials and their probability of occurrence greater." As always, different methods have different strengths and weaknesses in different contexts.

9. Sample size matters, and large trials are better than small trials, but it is pretty much useless to say so unless we know what sample size is enough. And that depends on the causal structure that we are dealing with, on skewness, on relative variances, and on many other things. If we are to make claims about validity based on large sample, we need an argument for that case.

10. We agree with Ioannidis about the desirability of using prior information in the design of an RCT, yet many defenders of RCTs, perhaps particularly in economics, see the lack of assumptions as being their great strength, at least compared with the structural models that were long the standard in the field. The so-called 'credibility revolution' in econometrics rests on the non-parametric and robust nature of its inferences, which are seen as independent of incredible, or at best doubtful, economic theories. Such approaches, including natural experiments and regression discontinuity designs, like well-designed RCTs, often do carry conviction over their domain of validity, but, as is often the case, they estimate an ATE that is valid only in the limited circumstances under which it was obtained, an aspect of the familiar trade-off between getting a result that is likely to be accurate in itself versus getting one that is likely to be useful as evidence for other claims. Put differently, the special requirements of RCTs, regression discontinuity methods, and instrumental variables often give us an estimate of a local magnitude that is not what we originally set out to measure. Credibility is fine, but it would be better if it did not so often involve a shift away from what we want to measure to some other quantity that is dictated by the design. Indeed, if we focus on design, as Imbens argues, we run the risk of prioritizing method over substance, and of ranking results, not by what we learn about the world, but according to a hierarchy of method, which is one of our main targets. We do not believe that economics would be better if the standard paper in the subject adopted the format of the ***New England Journal of Medicine***. Economics is about a great deal more than estimating average treatment effects; it is not epidemiology, and the design-analysis framework will not make economics more credible but turn it into the kind of alchemy that turns gold into lead, not the reverse.

Imbens defends the "credibility" of RCT-based studies, and of wannabe RCT designs, something we find hard to reconcile with the overstated significance levels in a large fraction of

RCTs in economics that Young documents, not to mention the unknown and unrepairable overstatement associated with asymmetry and the Bahadur-Savage theorem. Why these concerns are overblown, given that these studies were published in some of the top journals in economics, and why such results are credible is beyond our comprehension.

Unlike Imbens, we also think that economists (and others) are often confused about what RCTs can and cannot do. The World Bank quote we include is very hard to excuse as sloppy language, and in earlier drafts of our paper, we had several other similar quotes; people often do think that randomization implies balance, and do not understand the possibility of "unhappy randomizations." It is no excuse at all to say that sloppy argument is fine when writing for a lay audience. Rather,here is surely an additional obligation to be clear when writing for those who do not have the training to separate sloppiness from falsehood and whose actions will be guided by a mistaken understanding (as we suspect is the case with many endorsements for the many 'What Works' sites and charitable endeavors that emphasize the importance of RCTs, in which both the UK and US have invested heavily).

11. We concur with Sampson that "the assumptions and causal effects of any given RCT should be theorized within a larger organizational, political, and social structure." This point is seldom specifically addressed. But it is these larger structures that fix which causal pathways are likely and which are not, so they should not be ignored. When they are explicitly noted, they are often treated as 'moderators' in the potential outcomes equation: Give a name to a structure-type and introduce a yes-no moderator variable for it. Formally this can be done. But giving a name to a structure type does nothing towards telling us what the details of the structure are that matter nor how to identify them. In particular, the usual methods for trying to identify moderator variables, like subgroup analysis, are of little help in uncovering what the relevant aspects of a structure are

that afford the causal pathways of interest. Getting a grip on what social structures support similar causal pathways is central to using results from one place as evidence about another, and a casual treatment of them is likely to lead to badly mistaken inferences. The methodology for how to go about this is under developed, possibly because it cannot be well done with familiar statistical methods, and what resources there are across the socio-economic sciences, from ethnographies and interviews to formal modelling in political science and economics, are underutilized in RCT-oriented communities. Dahabreh hazards that medicine has fewer worries here than social science due to the "relative stability of biological structures and disease processes [22], as compared to the constant flux of social and economic structures." But we would warn against ever taking similarity of structure for granted.

12. To Raudenbush's worries about lack of rigor in educational research, we would like to add the concerns recently stressed by Adrian Simpson and cited in our paper about the casualness with which what happens in the control group is reported, even in highly respected 'What Works' cites like the US Department of Education's What Works Clearing House and the UK equivalent, the Education Endowment Foundation. As Cook notes, "Treatment effect claims are always conditional on the comparison group chosen."

13. With respect to Jones and Steel, if we are right, their hopes for empirical tests about which methods are best for answering which questions are doomed. All methods require assumptions to justify the results they are supposed to show. Instrumental variables methods require the absence of causal pathways from the instrument to the effect other than those via the cause under investigation; structural causal modelling needs related assumptions about the factors caused outside the system; to do much work, casual Bayes nets methods need a causally sufficient set of

factors in which causes and effects are probabilistically dependent; RCTs need orthogonality and all that is required to achieve it; and so on. In all these cases it is provable that if the required assumptions are true, the targeted causal conclusion is as well. So which method is best in a given case depends on what we know or are prepared to assume in that case.

14. Even supposing we know the true ATE of a treatment in a population of interest, S. V. Subramanian, Rockli Kim and Nicholas Christakis stress how little help this provides in deciding whether to introduce the treatment either for individuals in the population or for the population. Knowing the variance in individual treatment effects would help a lot more, but that cannot be point estimated from an RCT. Dahabreh contributes here by pointing out that bounds (the Hoeffding-Fréchet bounds) can be calculated; we omitted discussion of those on the grounds that experience suggests that they are often too wide to be informative, see Heckman, Smith and Clements (1997). In this same context, we note the important work by Manski, cited in our paper, on using RCTs to estimate policy-relevant objects other than ATEs.

15. We are happy to see Suzuki and VanderVeele endorse the INUS causality scheme we discussed (due to philosopher Mackie (1965), introduced into epidemiology by Rothman (1976)), which notes that for most effects there are separate clusters of causes each sufficient for a contribution to the effect, each cluster consisting of 'interacting' or 'moderating' factors that are separately necessary to get the contribution. As we noted, the ATE in an RCT is a function of the average value of the treatment's moderator variables. Also, we agree about the importance of tracing the operation of mediators that is central to the 'mediation analysis' work of Suzuki, VanderWeele, and others whom they cite, which is also stressed in the works we cited by Cartwright and Hardie (2012) and Parkkinen (2018). Contrary to Suzuki and VanderWeele's suggestion, however, mediation analyses are not underdeveloped in economics: Mediators are explicitly represented in causal structural models and indeed were one of the main topics of the early development of econometrics by the Cowles Commission; they only disappear in reduced form equations for these models.

Citations

(other than those in the original paper.)

Chang, H. (2004)**. *Inventing Temperature: Measurement and Scientific Progress,*** Oxford: Oxford University Press.

Mill, J. S. (1882). ***A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation.*** (8th edition) New York Harper.

Pearl, J. & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. ***Statistical Science,*** 29(4), 579–595.

Wylie, A. (2011). Critical distance: stabilizing evidential claims in archaeology. In Dawid, P., Twining, W. & Vasilaki, M. (Eds.). ***Evidence, Inference and Enquiry*** (371-394). Oxford: Oxford University Press.