Social Science & Medicine 210 (2018) 37–40

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Commentary

Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: A commentary on Deaton and Cartwright

Thomas D. Cook

Northwestern University and George Washington University, USA

1. Introduction

This paper goes over some of the same details as in Deaton and Cartwright (2018) but aims to succinctly systematize and even extend them. Its argument rests on four fundamental assumptions. First, we understand causation in terms of determining the effects of a manipulable cause rather than in terms of explanation, identifying mechanisms or achieving perfect prediction. Second, our focus is on the causal interpretation of single studies and not of planned or unplanned programs of research, for which some of our points are less relevant. Third, we limit the discussion to the absolute argument for randomized experiments (REs). This postulates that they function as a gold standard so that their mere use guarantees valid causal inference. The alternative, argument is relative; it is that REs justify causal claims better than their non-experimental alternatives and so merit being seen as more gold-like even if not golden. Space alone precludes taking up both the absolute and relative arguments.

Our fourth assumption is the most important. It is that researchers test causal hypotheses in contexts that are more contingent than those that operate in the theoretical world of statistical expectations, where it is assumed that many thousands of trials of a given causal hypothesis take place. Since there is only one trial in most research practice, researchers have to use statistical conventions in order to generate estimates of a causal effect and of its standard error. They also have to test the relationship between a cause and an effect such as each is explicitly named in the causal hypothesis. This requires a justification for moving from the specific causal agent actually manipulated and the specific outcome actually measured to the abstract constructs named in the causal hypothesis. Researchers are also constrained to include in their causal studies actual samples of persons, settings and times even though it is rare for the formulation of the causal hypothesis to mention them. It is also rare for samples to be chosen at random from some clearly specified population, being more often products of convenience and opportunism. Without valid population names for persons and settings in particular, causal estimates cannot be generalized and justifiable statements about factors co-conditioning an effect cannot be made.

This paper identifies the major factors that enhance the quality of causal inferences from REs, expressing them as factors influencing (a) internal validity, the nature of the link between the purported cause and effect – is it "causal" or "correlational"?; (b) statistical conclusion validity – how large and dependable is the estimate of the obtained cause-effect relationship, including whether it is effectively zero; (c) construct validity of the cause and effect – how well do the labels attached to the causal agent and its effect correspond with what researchers have actually manipulated or measured?; and (d) external validity – how generalizable is a causal result when the RE inevitably includes purposively chosen samples of persons, setting and times that are not explicitly included in usual bi-variate formulation of a causal hypothesis?

2. Internal validity: is the association between the possible cause and effect "causal"?

The absolute argument in favor of REs can be framed in a number of ways. One is that the different contrast groups – for convenience we discuss only a single treatment and a single control group – are constituted as randomly drawn samples from the same population. In expectation, they will therefore be initially identical on all measured and unmeasured attributes so that any group differences observed at posttest cannot be due to initial group differences ("selection"). Another argument is that REs logically entail full knowledge of the process of assignment to treatment; it is due to chance and nothing else. As a result, the treatment assignment process is independent of the study outcome and selection is again ruled out. Such independence can also hold with non-experimental studies, but only conditionally on perfect knowledge that the covariates used capture all of the true selection process that is correlated with the effect – an impossibly high bar except for well-implemented regression-discontinuity designs. At a theoretical level, these related rationales based on identical samples and full knowledge of the assignment process are impeccable, forming one leg of the two-legged argument that REs constitute the gold standard for causal research.

In actual research practice, the rationales above are conditional on some assumptions being met. We express them below, noting that they

E-mail address: t-cook@northwestern.edu.

https://doi.org/10.1016/j.socscimed.2018.04.031

Available online 21 April 2018

differ in how often they are met and how easy it is to diagnose they are met. For internal validity, the relevant assumptions are:

Assumption 1 is that a correct random assignment procedure has been chosen.

Assumption 2 is that this correct random assignment procedure has been correctly implemented.

Assumption 3 is that sampling error has not led to "unhappy randomization" – viz., despite random assignment a pre-intervention group selection difference exists that can be mistaken for a treatment effect. Such unhappy randomization is most likely with small samples and thus in research that assigns large and heterogeneous aggregates to treatment. While there are well-known ways of reducing the problem – e.g., by blocking prior to random assignment – they are not always implementable and sometimes result in solutions that cannot be shown to be complete.

Assumption 4 is that attrition from the RE has not differed between the contrast groups, thereby potentially confounding the treatment of interest with selection.

In reports of completed REs, it is commonplace to defend the first two assumptions by describing the random assignment process and its implementation while also testing for balance – whether the contrast group means differ at pretest. In work with smaller samples, it is common to match cases across contrast groups before randomly assigning them and later checking for pretest balance. The fourth assumption concerns differential attrition, and balance tests are relevant here too. If adequately powered statistical tests suggest that pretest imbalance is implausible, the analysis proceeds. If they do not, then statistical procedures have to be used to try to eliminate the resulting selection bias, though it is impossible to know with certainty that all the bias will be removed. Nonetheless, all four internal validity assumptions can be addressed in the RE data. The second leg on which the gold standard argument for REs depends builds on a half century of empirical work that has identified ever more of the assumptions random assignment requires, has led to ever better diagnostic tests and, when these checks fail, has developed corrective procedures that are valid if only by social consensus. As a result, statistical theory and a half century of practice implementing REs render assumptions 1 through 4 not very problematic in careful RE practice; they only marginally undermine the absolute gold standard rhetoric.

3. Statistical conclusion validity: what is the size and dependability of a causal association?

Statistical hypothesis tests are used to link theory about expectations to actual RE practice. These tests provide a point estimate of the size of a causal relationship and also a standard error estimate of the dependability of this size estimate. A test can also be made of whether the obtained causal estimate reliably differs from zero. This probabilistic hypothesis testing framework is indispensable but carries its own set of assumptions that can easily be violated or for which the relevant diagnostic tools are imperfect. Among the relevant assumptions are:

Assumption 5: That a correct statistical test has been chosen, given factors like the distribution of the outcome (continuous or categorical) or how the data are clustered.

Assumption 6: That an alpha level has been chosen for hypothesis testing that can be convincingly defended.

Assumption 7: That the nominal and actual alpha rates correspond. This is to avoid the illiberal use of too many non-independent hypothesis tests – aka "fishing" or "capitalizing on chance".

Standard errors vary with a number of conceptual irrelevancies, none of which enter into how the causal hypothesis is formulated but that nonetheless cannot be avoided. Among them are:

Assumption 8: Rejecting the null hypothesis depends on the homogeneity of the persons, settings and time points sampled – the more homogeneous they are, the more valid is the estimate of the effect and its confidence interval.

Assumption 9: The sample size – the larger this is the more efficient is the design.

Assumption 10: The availability of covariates correlated with the effect measure – the higher their collective correlation the tighter is the estimate of the unobserved true effect.

Assumptions 5 through 10 necessarily condition the interpretation of hypothesis tests. Yet it is rare to find them explicitly included in causal claims as limiting factors. Of course, including them would result in long-winded causal statements full of statistical details readers would see as irrelevant to the bi-variate causal hypothesis under test! Moreover, these assumptions are well known and it is easy for researchers to learn about what to do to deal with them – e.g., by means of pre-study statistical power tests, registering hypotheses and models, knowing how robust each assumption is to violation, and clearly describing the statistical procedures used and the samples included.

Nonetheless, researchers can make mistakes on any of these threat fronts, and do so. Indeed, it is sometimes practically difficult to increase sample sizes, to add covariates, or to measure and model sources of heterogeneity. Even so, careful researcher learn how to deal with most of these threats to statistical conclusion validity even if they do not necessarily rule them all out completely. Assumptions 1 through 10 dent the gold standard rhetoric about REs but do not render it totally implausible. Careful REs may not constitute an absolute gold standard, but they nonetheless offer a close approximation to it if we consider only internal and statistical conclusion validity.

Alas, these are not the only kinds of validity implicated in causal statements. This can best be seen from scientific fields that mostly use REs – e.g., social psychology - but that are nonetheless replete with controversies about causal issues, including about the interpretation and application reach of results from single REs. Some of the controversies touch on issues already covered here – e.g., unhappy randomization due to small sample sizes, or data analyses that capitalize on chance and undermine replication. But other debates concern what the causal manipulation and/or the effect measures "mean", or what the proper range of application of a causal finding is – issues we now take up.

4. Construct validity of the cause and effect – how should manipulations and outcome measures be labelled in simple language?

Most causal statements involve just two entities: A cause and an effect. But REs inevitably include three entities: A treatment group, a comparison group, and an effect measure. This mismatch occurs because REs do not actually test the bi-variate cause/effect connections explicit in the verbal formulation of a causal hypothesis. Instead, they test how the effect measure has changed due to the contrast between the treatment and comparison groups. Contrasts are the real causal agents in REs, not the causal construct named in the hypothesis statement. Unfortunately, causal claims from the very same treatment can differ with the particular comparison group chosen. Effects tend to be larger when some form of a passive comparison is used, e.g., a business-*as*-usual control group; and they tend to be smaller when an active comparison is used, e.g., placebo controls. Treatment effect claims are always conditional on the comparison group chosen.

Assumption 11 is that a defensible argument is available to justify the particular comparison group chosen. Meeting this assumption is not as easy as it sounds, given the many kinds of choice available – business-*as*-usual control groups, waiting-list controls, historical controls, or various forms of active control such as placebo groups or groups that receive an alternative intervention with similar aims to the target treatment group. Moreover, stakeholder groups can differ in the kinds of control they prefer, as in research on mental health interventions where behavioral scientists generally prefer business-*as*-usual controls but psychiatrists favor some form of active control. While the criteria for comparison group choice deserve to be made explicitly, they are still inherently squishy.

The possibility exists of communication between units in the treatment and control group. Such interdependencies are generally not part of the causal hypothesis formulation, but they can inflate or deflate causal estimates depending on the particular social comparison processes that result from comparing the treatment statuses under test.

Assumption 12 is that the control group does not include any dimensions that are meant to be unique to the treatment, for this will reduce the size of the planned treatment contrast.

Assumption 13 is that there is not "compensatory rivalry", as when the control group responds to not getting the treatment by trying harder than it would otherwise have done, also called a "John Henry" effect.

Assumption 14 is that there is not "compensatory equalization", as when an administrator observes the unequal distribution of resources that an RE requires and tries to stifle any anticipated resulting discord by providing extra resources to the control group. The net result is again a possible attenuation of the casual contrast compared to what it would be without such administrator intervention.

Assumption 15 is that the control group does not become demoralized through learning they have not been favored with the intervention – a process that entails a causal direction from the control group to the outcome rather than from the treatment to the outcome, as intended.

Assumptions 12 through 15 depend on study units knowing of the different resources each treatment group gets. The key to dealing with them is, therefore, to ensure that such communication cannot occur. This is often possible, but not always; and dealing with it can increase research costs due to the logistical complications of distance and the possibility that greater distances will engender more heterogeneous samples that may therefore need to be larger to maintain the same efficiency.

The construct validity of the cause would be an issue even if causal conclusions did not depend on the contrast between the treatment and comparison group. Public communication requires attaching an abstract summary label to the treatment even though the treatment particulars will be much more multi-dimensional than the label. Such a mismatch between causal labels and causal operations is most acute with constructs from substantive theory where it is sometimes difficult to get inter-observer agreement on what all the constituent dimensions are; the relevant substantive theory often fails to specify all the intervention components; and in the form the intervention is actually implemented it is highly likely to include components that are not in the relevant theory but that are needed to make implementation practical. These are major reasons why fields like social psychology have active disagreements, despite a high frequency of REs. But the same situation exists in more applied contexts too – even those where where a manual prescribes the intervention particulars to be implemented as the causal agent. What happpens at the ground level will almost always respect the manual content; but implementing all of it will be rare and unplanned elements will often be smuggled in during the implementation process, whether by design as program adaptations or by inadvertence as opportunistic accommodations.

Assumption 16 is that the description/definition of the causal agent is completely captured by the implemented treatment particulars. Otherwise, the offered treatment label cannot be valid.

Assumption 17 is that no other outcome-related features are introduced into the operational treatment that are not part of its abstract description/definition. Otherwise, they might explain the treatment effect and so require a different causal label from that offered.

Assumption 18 is that the label for the causal agent specifies the level of implementation achieved for that agent. This is important because a different causal estimate might result as a function of the treatment "dosage" level. Yet the dosage level is rarely specified as an intrinsic component of how the causal agent is labelled.

Assumption 19 refers to other treatment dimensions that vary between study units. Implementation is always variable in factors other than dosage, thus affecting the reliability of treatment implementation. The less reliable it is, the more likely it is that the cause/effect relationship will be attenuated. Yet the reliability of treatment implementation rarely figures as a factor conditioning the size of the obtained effect

Valid labeling is an issue for the study effect as well as its cause. Causal hypotheses are usually formulated in the abstract and shorthand language of effect constructs like "hours worked" or "academic achievement" or "crime rates". The measures of such constructs are much more complicated in their implementation and require assumptions about the fit between construct and measure. Three stand out.

Assumption 20 is that the description/definition of the effect is completely captured by the measured particulars.

Assumption 21 is that no other outcome-related features are introduced into the operational specification of the effect that are not part of its abstract definition. Inevitably there will be, though, since statements of effects rarely include specific details about how and when they are measured.

Assumption 22 refers to the reliability with which the effect is measured. This is important because unreliability attenuates the size of the causal estimate obtained, making that estimate conditional on measurement error or the quality of attempts to control for it where such control attempts are made.

Assumptions about the construct validity of causes and effects draw attention to how well the cause and effect operations correspond with the cause and effect labels specified in the causal hypothesis. All are generic threats that condition the quality of the claims researchers can make, whether in a non-experiment or RE. However, with REs assumptions 11 through 22 are more problematic than assumptions 1 through 10 because there is less technical knowledge about how to improve construct validity in practice than there is about improving internal validity through better random assignment procedures and improving statistical conclusion validity through better null hypothesis testing. To be fair, psychometrics helps with the construct validity of effects, but there is little to guide thinking about the construct validity of causes. At most there are invocations to measure whether the purportedly most important dimensions of the cause vary with treatment assignment.

5. External validity: how do causal claims generalize to and across persons, settings and times?

Basic research seeks to test universal causal propositions whose applicability is not restricted to any specific population/universe/category/class of persons, settings or times; the hope is that a given causal relationship will hold everywhere and at all times. As research becomes more applied, target populations are more likely to be specified, such as the persons providing services (say, pre-school teachers in the USA) or those receiving them (say, pre-schoolers) or setting targets (say, California Head Start centers) or time-dependent targets, as with the year 2018. Conducting research without person, setting and time samples is impossible. Sometimes, researchers aspire to include samples that "represent" the specifically named targets in their hypothesis formulation in hopes of identifying the range of application of the causal relationship under test. Mostly, though, RE researchers slip into purposive samples from within the domains that are of interest to them and readily forget that their samples are chosen at a given time, in volunteering settings and with purposively chosen samples of those administering and receiving the treatment.

The main conceptual problem is that the best theory of representativeness is rarely possible in REs. This theory requires sampling with known probability from a clearly designated universe. Instead, the under-explicated practice has evolved of purposive choice within target categories in order to select, say, a sample of pre-schoolers in Head Start. The students chosen are likely to be from centers that volunteer for study in a circumscribed geographical area close to the researcher home base, and then to be limited to a single point in time. Some heterogeneity is possible across persons, settings and times and, if the resources for measurement and statistical power are on hand, a test will be made of how well a causal relationship holds across different categories of pre-schoolers, teachers, settings and times. Even so, none of these strategies speaks to formal representativeness. The same basic sampling theory that justifies random treatment assignment and internal validity cannot be used for random sample selection and external validity. Nonetheless, three relevant external validity Assumptions are:

Assumption 23 is that the achieved sample of persons is representative of the intended target population of persons.

Assumption 24 is that the achieved sample of settings is representative of the intended target population of settings.

Assumption 25 is that the achieved sample of historical times is representative of the intended target population of times.

Generalizing causal relationships involves more than just generalizing from samples to intended target populations. It also involves extrapolating casual relationships established with purposive sample of persons, settings and times to novel persons, settings and times with attributes different from those past sampling particulars represent. Stakeholders to RE results want to know whether a similar effect estimate would emerge in the future and with different (but likely overlapping) populations of persons and of settings like those for which they are accountable; and researchers also seek to make claims about the viability of their causal estimate in more general circumstances than those built into their original design. From this extrapolation-based understanding of external validity emerges an especially problematic assumption.

Assumption 26 is that valid extrapolations of the causal results can be made to populations of persons, settings and times that have demonstrably different attributes from those studied to date.

6. Conclusions

We have identified 26 Assumptions on which causal conclusions from REs depend. We could have sliced the pie differently and come up with different numbers. For instance, some statisticians roll Assumptions 12 through 15 together under the name of the SUTVA assumption. We prefer not to, since it fails to capture the different casual signs that can result from the social processes engendered by the non-independence of treatment and control units. Even so, there is nothing sacrosanct about these 26 that are mostly the products of reflection on past practice rather than deductions from some comprehensive theory of causal study design.

Researchers test causal propositions in contextualized settings involving many factors other than the cause and effect constructs that suffice for formulating a causal hypothesis in words and the notions of treatment, control, outcome, chance allocation process and random error that suffice for formulating a causal hypothesis in statistical terms. These other factors speak to the necessity of: (a) implementing and maintaining the random assignment procedure – internal validity; (b) using statistical tests to arrive at treatment effect estimates and their standard errors – statistical conclusion validity; (c) operationalizing the cause and effect constructs that frame the research question – construct validity; and (d) incorporating samples of persons, settings and times into the research that are rarely chosen at random and are mostly opportunistically selected from within a named category of interest – external validity.

It is highly unlikely than any one RE will convincingly meet all of these 26 Assumptions, though the art of causal research design consists in trying to do just this. Given how intrinsically opaque some assumptions are, and given also how judgment-riddled many of the relevant diagnostic tests are, it is difficult to review a completed RE and conclude that its causal estimates are infallible– viz., to claim that the results meet some absolute gold standard because random assignment was used.

It can be argued that, since random assignment speaks only to internal validity, only the first 4 Assumptions we presented are relevant and that we know much about these assumptions and the validity and sensitivity of their diagnostic tests. This is true. But, in practice, hypothesis-testing takes place within a stochastic framework, and so Assumptions 5 through 10 are also relevant for RE practice. Assumptions 5 through 10 are also well known, though, and steps have evolved over time to deal with them. They are not particularly difficult for the careful researcher. This is also true. However, there are now 10 assumptions to deal with, and uncertainty about ruling them all out is bound to be greater than with just the first four. Seen only from the perspective of internal and statistical conclusion validity, it is much easier to argue that RE represents a high standard of casual inference than to argue that it is the gold standard whose use guarantees perfectly valid internal validity.

Researchers do not test causal hypotheses about an unnamed manipulandum and an unnamed effect; each has a general abstract label attached to it in ordinary language, indicating that Assumptions 11 though 22 must be addressed too. Moreover, researchers cannot test causal hypotheses without including samples of persons, settings and times that are often not explicit components of the causal hypothesis formulation. Nonetheless, these samples define the bounds within which a demonstrated causal relationship applies and, unbeknownst to the researcher, they may even condition it. As a result, the external validity assumptions – 23 through 26 – are also relevant to judgments about how well an RE represents the gold standard for causal knowledge. To argue that only the first 4 – or the first 10 – assumptions are relevant depends on logic and statistical theory, but it obscures pragmatism because researchers have no choice but to engage with assumptions 11 through 26 where the assumptions are expressed more vaguely and the diagnostic tests are less specific in their criteria.

Of all the Assumptions, 1 through 4 apply to REs, while Assumptions 5 through 26 apply to non-experiments too. However, assumptions 1 through 4 exist in a practice vacuum devoid of complexities associated with statistical hypothesis-testing, identification of the cause and effect operations in more general terms, and understanding what the achieved samples of persons, settings and times represent in population terms. The statistical and sampling context in which random assignment is inevitably embedded means that the causal knowledge an RE provides is limited by (a) statistical errors and vagaries of sampling error, (b) cause and effect constructs that do not fully map onto the operations designed to index them, and (c) person, setting and time populations that are not related to their achieved samples in any theoretically acceptable way. REs do not seem to deserve gold standard rhetoric when viewed from the four validity types used here.

Nonetheless, with more space we could have considered the argument that REs deserve to be called the relative gold standard – viz., relative to designs without random assignment. Of necessity, that argument would have to deal with Assumptions 1 through 4 and also with theoretical and empirical evidence about the conditions under which specific alternative designs – like regression-discontinuity or comparative interrupted time series, or different kinds of non-equivalent control group designs – dependably reproduce the causal estimates achieved from REs with the same treatment and measurement particulars. Such a review is worth doing, but what we have done instead is to demonstrate the limits of the causal knowledge that any one RE generates by emphasizing the relevance of these 26 assumptions and the low likelihood that they will all be convincingly met in a single study.

Reference

Deaton, Cartwright, 2018. Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med. 210C, 2–21.