How to evaluate a study: the outside view

What is this document?

The Founders Pledge Study Evaluation Checklist describes a long list of considerations that can apply when you're looking at a specific study: internal and external threats to validity, questions about causality, and more. But we don't always have time to look at each study in depth. We're likely to miss important minutiae, and, moreover, we're really concerned with the **weight of evidence**.

In particular, we're often concerned with **causal evidence**, and often want to privilege studies that most compellingly demonstrate a causal effect. We also want to be careful of wasting our time on research that has **obvious red flags** or that are **obviously bullshit**.

This document is a guide for (1) rapidly judging the trustworthiness of a piece of scientific evidence and (2) triaging to avoid wasting time and/or improperly weighting bad evidence.

What are the indicators of good/bad research?

Methodologies that are acceptable as causal evidence

It's common to make causal claims in social science, but unfortunately, it's rarely justified. A common tactic in non-experimental causal inference is to run a regression analysis while claiming to "control for" possible confounders. The claim here is that, if all possible variables that could affect both the independent and dependent variables are accounted for, the only remaining source of variation must be the causal relationship between those two variables. The two main problems with this are (1) it is impossible to account for all, or even most, possible confounders and (2) causation can still run in the other direction.

Ecological or observational studies are almost always unacceptable as causal evidence at Founders Pledge. When considering the full body of evidence about an intervention, social or environmental problem, or other phenomenon, we weight such research at approximately 0.

This is not to say such studies are not useful - but they are almost never causal evidence. If you are looking for causal evidence, **only studies using the following methodologies are acceptable**.

Experimental evidence (High quality)

Randomized controlled trials—experiments—are the gold standard for causal inference. In keeping with the study evaluation checklist, a more rigorous evaluation of a given RCT will involve checking to see that the study is well-designed, that randomization occurs correctly, and that attrition is appropriately handled, but these are generally speaking the most reliable form of causal evidence. Note that the primary potential issue with RCTs is that they are not always **generalizable** - though an RCT is high-quality evidence, the question you will often face is **whether it is applicable to the question at hand**.

The Bangladesh mask RCT is a good example of a well-done experiment.

Quasi-experimental evidence (Medium- to High-quality)

Instrumental variables

Instrumental Variables (IV) designs are statistical designs that essentially simulate an experimental setting by taking advantage of natural variation that can be assumed to be random, or that is at least random with respect to the outcome of interest. The classic example of this is weather. Suppose you want to figure out if people are more productive when they stay home from work. Severe snowfall causes people to work from home, but there's no reason to suspect that it directly impacts productivity, so you might be able to assume that the snowfall-staying home relationship is roughly equivalent to an RCT in which people are assigned to the home/not-home condition — in this case randomly, according to the weather. Be careful about using rainfall, though!

A good example of an IV design is this paper, which uses the existence of a direct flight between Washington, D.C. and various other cities to instrument for lobbying spending.

Regression Discontinuity

Regression discontinuity (RD) designs take advantage of the existence of a more or less arbitrary threshold between two potential treatment conditions, like a geographic boundary or a test score cutoff. The "identification assumption" allowing you to draw a causal conclusion in this setting is that the subjects on either side of the threshold are more or less the same, other than having experienced whatever treatment condition the threshold defines, such as admission to a competitive program (for test scores) or being subjected to a certain policy (for a geographic boundary).

You need to be very careful when evaluating RD studies: a **good rule of thumb** for RD designs is that if you can't clearly see the effect in a diagram, it's probably not a good study; be extremely wary of overfitting. But the best RD studies are extremely convincing.

A good example of an RD design is this carefully done study on the effects of political advertising (see figure 2 on page 36).

Other natural experiments

Natural experiments that do not either fall into the IV or RD categories also often provide evidence that is as good as an RCT, particularly if random assignment is actually involved. This study, for instance, used the Vietnam draft lottery to estimate the earnings penalty to having served.

Causally suggestive evidence (Medium quality)

Dose-response relationships

Dose-response relationships occur, as you might imagine, when some treatment (say, exposure to air pollution) produces an outcome in direct proportion to how much of the treatment is delivered. Studies that claim the existence of a dose-response relationship, like this one that finds a dose-response relationship between household poverty and the magnitude of benefits delivered by Medicaid, are *suggestive* of a causal effect. The reasoning here is something like: "it would be a hell of a coincidence if these variables that just so happened to be related were related in such a way that a dose of one prompts a proportion response in the other." Still, you should be wary of trusting D-R studies too much, as they are often highly technical and may hide key mathematical assumptions.

Synthetic control models

Synthetic control models are a new innovation — in summary, they attempt to simulate the counterfactual path of some subject in the absence of the policy whose impact is being estimated. This is done by constructing a "synthetic control" — a simulated unit composed of weighted data from other units (like states or provinces). The reason synthetic control models are sometimes convincing is the use of "placebo tests": modelers can try many different weightings to see how many of them would have produced the modeled deviation between the treated unit and its counterfactual synthetic control. The best and most convincing synthetic control model to date is this study (see figures 2 and 3) that estimates the effect of cigarette taxes on consumption in California. For an example that gives a null result, see Matt L's MA thesis.

Founders Pledge

A note on time-series models

You'll note that difference-in-differences designs are not listed here as high-quality evidence. That's because they are fundamentally observational studies that offer analysts many degrees of freedom to P-hack their way to statistical significance. Unless they are very strong, they are not good evidence.

Major red flags (in causal or non-causal research)

P-values that are very close to the significance level
Since statistical significance is and has historically been high-stakes, we unfortunately have to take p=0.049 as a signal that there may be some shenanigans going on.

Use of causal language to describe non-causal evidence
In an observational study, the claim that "X increases Y" or "Y results in Z" is sloppy. If there is no causal design, then causal claims are unjustified and you should deprioritize the study on the grounds of authorial competence.

Combative or overly emotional language
Similarly, while this may be superficial, it is only superficially superficial: we don't generally want to waste time on studies whose authors have an axe to grind, because we have *prima facie* evidence against their objectivity.

Poor writing
Again, this is a signal of potential authorial incompetence. Indeed, it could be true that the quantitative work is good — but if we're trying to assess a large number of studies, we should deprioritize those where we have suspicions about competence

Obvious conflicts of interest
It's always good to Google authors, particularly when studies are conducted on particular interventions or organizations. If the authors work for those organizations — or if they have other outstanding conflicts of interests — deprioritize immediately.

Implausible generalization
This is the standard mistake made in most social psychology research, and in lots of lab research more generally. Does the study really show the thing it claims to show? Some claims squeak by without further inspection simply because they are experiments — but not every experiment is transferable to a practical setting. This has to do with the generalizability crisis.

Low-impact or unknown journals
Another sad-but-true thing: some journals will publish anything. When we look at a scientific paper, we are taking some degree of due diligence for granted. But we should keep in mind that (1) some disciplines have lower evidentiary standards than others and that (2) there are many thousands of journals, many of them garbage. If you've never heard of the publication, you should look into it briefly. There are good reasons to trust some journals in a general way — they are the flagship journal for a small or niche discipline, they are edited or managed by high-profile experts, or they have a high impact factor—and others should be treated more skeptically.

Interpreting claims from the physical or natural sciences

Beware of implausibly big claims
It's true that the absurdity heuristic is a cognitive bias — we shouldn't reject claims that seem intuitively strange because they're unfamiliar. But, in the sciences, we need to be *extremely careful* of any claim that suggests a step change in scientific achievement. The canonical example of this is cold fusion, but similar claims arise all the time.

Beware dubious generalization
The claim that it's better to drink hot drinks in hot weather comes from a paper titled "*Body heat storage during physical activity is lower with hot fluid ingestion under conditions that permit full evaporation.*" But that is not really what the study shows: it was a nine-person study under controlled conditions, with controlled water intake. It doesn't show anything in general about "conditions that permit full evaporation" and doesn't say anything in general about heat storage during physical activity. Is it evidence in favor of the proposition indicated in the paper title? Yes — but not very much.

What are observational studies good for?

Getting base rates
Though this study about treaty compliance may make some questionable analytical choices, it's really useful for putting a number on the probability that a treaty will work. Similarly, the observational data on page 7 of Rosie's report on oral health is useful for understanding the scale of the problem.

Getting an idea of where there's no causality
It's *not* the case that no correlation means no causation — but it is evidence in that direction. If there's a theory that X causes Y, but there's no correlation between X and Y, that is suggestive evidence of a lack of a causal relationship. On a more sophisticated level, you can even estimate the *size* of the causal effect that could be detectable.

When is non-quantitative evidence useful?

Learning what questions to ask
Qualitative, anthropological, or historical evidence can point to new avenues for research and raise potential issues that weren't obvious. For this reason, it can be very valuable to review "on-the-ground" work, historical reviews, or case studies.

Developing a theory of change
When evaluating a cause area, reading deeply on the issue can generate ideas for potential interventions or mechanisms of action. This paper about the movement toward asteroid defense is a good example - it provides historical case studies about inside lobbying that can generate ideas for new organizations or initiatives going forward.

Obtaining an existence proof
When evaluating potential theories of change, it's important to know (a) whether a given strategy can work and (b) how exactly "the strategy working" would look. A good source for philanthropic case studies, for instance, is this book listing examples of philanthropic success over the course of the 20th century.

How to apply these standards?

My (Matt L's) recommendation is to **track study quality** when you're doing an investigation of any type. You should keep track of all the studies you use, and you can use either Zotero or Google Sheets to tag individual pieces of evidence as low-, medium-, or high-quality. This procedure has the following virtues:

- It enforces attention to study quality
- It allows you to quantify evidence quality (e.g. how much evidence is high-quality vs. low)?
- It allows you to compare the amount of evidence you have for different claims