Observational causal evaluations with quasi-experimental designs

The goal of this document is to provide helpful resources for OES team members engaged in observational, usually retrospective, causal projects. In particular, this intends to support and augment conversations with agency partners, especially those unfamiliar with designs for such projects. This piece does not intend to provide guidance to be followed during analysis, but intends to outline OES's perspective on observational causal studies. We expect that agency partners will work closely with OES team members on the details of their particular designs.

Observational causal projects differ in important ways from the prospective randomized designs that make up the core of OES work. These studies rely on observational data – data in which the researchers have not themselves assigned the treatment or intervention – to identify causal relationships using a quasi-experimental design (QED). OMB clearly characterizes these observational designs as "impact evaluations" (on page 10, here). Not only do these projects require methods and designs that deviate from recommended methods for analyzing the results of randomized controlled trials (RCTs), they also necessitate modifications to the language we use to communicate results to stakeholders and external audiences.

What is a "quasi-experimental design"?

Observational causal studies often have at their foundations a "quasi-experimental design". This design is a set of assumptions, evidence, and methods that may enable valid causal inferences from the observational data. For example, rather than having a researcher-controlled randomized experiment that provides strong foundations for causal inference, we might encounter a program with a threshold for participation, and those on either side of the threshold might be comparable to each other. This comparability might enable a strong causal comparison. Or, we might have a situation where access to a program is determined by idiosyncratic factors unrelated to outcomes, and those who participate and do not participate are otherwise good comparisons for each other.

It may be very difficult in a QED to rule out one strong plausible alternative explanation for differences that are observed. It may be difficult to rule out *many* plausible alternative explanations, as well. For example, those above and below a program threshold may be fundamentally different in ways that are difficult to capture, or program access may look random, but really be systematic.

Among the most significant challenges to observational causal research is understanding the assignment mechanism. That is, how did the intervention come to be assigned to some units and not others? Because the researchers did not assign the treatment, this mechanism is not known (as

it would be in an experiment) and confounding looms as a threat to causal identification. Thus, every QED should include language about assumptions and caveats to causal interpretation.

Examples of OES observational causal studies assessing the effects of grants to small businesses can be found here and here.

How to discuss QEDs

In the OES abstract and other documents, the strengths and weaknesses of the QED must be identified. The assumptions required for causal interpretation should be clearly stated in both statistical and non-statistical language. For a difference-in-difference design, we might say, "The validity of a causal interpretation of the difference rests on a 'parallel trends' assumption that the trajectories of the [outcome] in the [treatment] and [control] groups would have been identical, in the absence of [treatment]." For example, "The validity of a causal interpretation of the difference rests on a 'parallel trends' assumption that the trajectories of employment in New Jersey and Pennsylvania would have been identical in the absence of the minimum wage increase."

We should invest time and effort in evaluating how well the design's required assumptions are met in practice. This determines whether the design yields causal estimates. At the same time, we should discuss the degree to which we cannot evaluate the assumptions. For example, the parallel trends assumption can never be tested in practice, because it is an assumption about unobservable potential outcomes ("what would have happened in New Jersey had the minimum wage *not* been increased?"). We may be able to show that two states' firms behaved similarly for several years before the policy change, responding to similar economic conditions, but we would want to a) buttress this with more detail about the similar economic conditions during the period after the policy change, and b) be clear that this does not demonstrate that the unobservable trends were, in fact, parallel.

This discussion involves evaluating how well the strengths of the design can be leveraged to overcome the vulnerabilities of the observational data in practice. We should not present observational results as if they are experimental, with a small caveat at the end. For example, appending to a causal claim a phrase like "but we can't be sure that policy X was the cause" or "but we can't be sure this correlation is causal" is insufficient. Summaries *should include* a statement like **"Completed applications increased by 20% after documentation requirements were reduced, but we cannot rule out that other factors may contribute to this increase."**, but more thorough discussion of *why* we cannot rule out those factors is needed in an observational causal study.

Risks of QEDs

When we undertake an observational causal study, we must be very careful to delineate the assumptions required for the inference to be causal. We should provide evidence where possible to assess consistency with those assumptions.

Even if we carefully do so, other stakeholders may interpret our observational work without fully appreciating these constraints on causal interpretations. If we anticipate that others will unduly ignore or gloss over threats to causal inference, we may need to reject the project.

Another risk that partners should expect in QEDs is that the project may not be able to yield convincing causal estimates, *even after the data are delivered*. In an experiment, we can usually produce causal estimates once we have the outcome data. However, in a QED, the background data may tell us that it will be difficult to interpret any estimate causally.

Project process

We use the same project process for an observational causal study as for a randomized experiment. As with other projects, we may draft two analysis plans, one higher level for external publication, and one more detailed for internal planning and guiding reanalysis. However, we recommend against this to avoid inconsistencies in content.

Observational causal designs should be pre-registered to the degree possible. When the outcomes of an observational design already exist (a retrospective observational design), we should be careful not to obtain or link the outcomes to treatment conditions until the design is registered. In cases where this may not be possible, such as when treatment conditions and outcomes are only accessible in a single dataset, we should be careful not to *analyze* the connection between treatment condition and outcome prior to design registration. Rubin (2007) provides a perspective on observational "design *versus* analysis".

We adhere to our usual norms of reanalysis in observational causal studies. An independent reanalyst should be able to take the analysis plan, notes from the primary analysts, and data sets and obtain the same results as the primary analysis.

We recommend sensitivity tests, such as Rosenbaum's Γ to summarize how severe unobserved confounding would have to be in order to change conclusions.

Resources for QEDs

EGAP provides a nice menu of causal methods. Christine Cai also provides a compilation of related resources. Paul Rosenbaum's "Design of Observational Studies" provides an introduction to inference and concepts in observational causal designs, with a particular focus on matching methods.

Designs

Below we describe data environments that may be conducive to certain QEDs. The appropriate method may depend on the estimand that we are targeting, the richness of the data that we have access to, and the treatment assignment mechanism. For example, where a local average treatment effect is acceptable, and a strong encouragement to take up treatment is randomized by researchers, an instrumental variables design suggests itself. On the other hand, where a program

has a sharp participation cutoff by income, many people are just above and below that cutoff, outcomes would change smoothly as a function of income, and a treatment effect specific to the income cutoff is acceptable, a regression discontinuity design suggests itself. However, it is important to note we would usually prefer a randomized experiment among the units around the cutoff to a sharp cutoff itself.

Treatment units	**Control units**	**Time data**	**Other data**	**Method**	**Notes**
1	Many	Many periods before (and after) intervention	Time-series of outcomes and predictors	Synthetic Control	Some variants allow several treated units (synthetic matching, augmented SC, generalized SC)
1	1	≥ 2 periods (before and after)	Outcomes in all periods	Difference-in-differences	
1	0	2 periods		Before-After	Very weak causal identification
Many	Many		Rich covariate data	Matching	
Many	Many	Many periods	Irreversible treatment	Generalized D-i-D	Calloway-Sant'Anna
Many	Many		A threshold or cutoff on an otherwise-smoothly effective predictor	Regression discontinuity	Estimates "local" to cutoff neighborhood
Many	Many		Randomization not of treatment, but of something that then induces treatment	Instrumental variables	Estimates "local" to compliers
Many	Many		Predictors of both assignment and outcome	Doubly-robust estimators (AIPW)	