Declaring and Diagnosing Research Designs[†]

Graeme Blair[‡] Jasper Cooper[§] Alexander Coppock[¶] Macartan Humphreys[††]

First draft: 5/7/2016
This draft: 10/4/2016 12:34

Abstract

The evaluation of research depends on assessments of the quality of underlying research designs. Surprisingly, however, there is no standard definition for what a design is. We provide a framework for formally characterizing the analytically relevant features of a research design. In standard applications, the approach to design declaration we describe requires defining population structures, a potential outcomes function, a sampling strategy, an assignment strategy, estimands, and an estimation strategy. Given a formal declaration of a design in code, Monte Carlo techniques can then be easily applied to a design in order to diagnose properties, such as power, bias, expected mean squared error, external validity with respect to some population, and other "diagnosands." Declaring a design in computer code lays researchers' assumptions bare and allows for clear communication with readers. Ex ante design declarations can be used to improve designs and facilitate preregistration, analysis, and ex post reconciliation of intended and actual analyses. Design declaration is also useful ex post however and can be used to describe and share designs as well as to facilitate reanalysis and critique. We provide an open-source software package, `DeclareDesign`, to implement the proposed approach.

[†]Authors are listed in alphabetical order. This work was supported in part by a grant from the Laura and John Arnold Foundation and seed funding from EGAP – Evidence in Governance and Politics. Errors remain the responsibility of the authors. We thank Peter Aronow, Justin Grimmer, Kolby Hansen, Chad Hazlett, Tom Leavitt, Winston Lin, Matto Mildenberger, Matthias Orlowski, Molly Roberts, Tara Slough, Gosha Syunyaev, Anna Wilke, Erin York, Lauren Young, Yang-Yang Zhou, Teppei Yamamoto, and participants at the Southern California Methods Workshop and the APSA and EPSA 2016 annual meetings for helpful comments. The methods proposed in this paper are implemented in an accompanying open-source software package, `DeclareDesign` (Blair et al., 2016*a*).

[‡]Assistant Professor of Political Science, UCLA. graeme.blair@ucla.edu. `https://graemeblair.com`

[§]Ph.D. candidate in Political Science, Columbia University. jjc2247@columbia.edu. `http://jasper-cooper.com`

[¶]Assistant Professor of Political Science, Yale University. alex.coppock@yale.edu. `https://alexandercoppock.com`

[††]Professor of Political Science, Columbia University. mh2245@columbia.edu. `http://www.macartan.nyc`

Authors and readers of empirical research each have an interest in being able to assess the properties of research designs. In doing so, however, they face two challenges.

First, there are very few tools for assessing the properties of designs. At one extreme, researchers resort to rudimentary power calculators with often hidden assumptions or rely on rules of thumb that may not account for important idiosyncratic features of the research setting. At the other extreme, some scholars conduct fully-fledged simulations requiring advanced programming skills beyond the capabilities of many applied researchers. General-use tools for assessing important properties of designs beyond statistical power, such as bias, are not available.

Second, surprisingly little attention has been paid to the more fundamental question of what constitutes a design. This lack of clarity carries risks both before and after the implementation of a study. If designs are incompletely specified ex ante it is difficult for researchers to assess their strengths and improve them. If they are incompletely specified at the time of analysis, concerns about data snooping may arise. If they remain unspecified after analysis, it may be difficult for other scholars to know how to replicate a study or whether a given type of reanalysis is justified.

In this paper we describe an approach that addresses these two problems by, first, enabling researchers to *declare* research designs[1] mathematically and as computer code objects and, second, to *diagnose* the statistical properties of the design relying on this declaration. We formally define research designs and clarify what features of the design must be declared in order to implement, communicate, and assess its properties. When possible we see advantages to formally characterizing and diagnosing designs before implementation. The resulting design description and diagnosis can then serve many purposes. A researcher may wish to include them as part of a preanalysis plan or a funding request. Whether or not the declaration and diagnosis serves this purpose, we believe that the process of generating them will provide researchers an opportunity to learn about and improve their inferential strategies. Even if only declared ex-post, formal declaration still has advantages; the complete characterization can help readers understand the properties of a research project, facilitate replication, and contribute to re-analysis decisions.

The approach we describe is clearly more easily applied to some types of research than others. In prospective confirmatory work, for example, researchers may have access to all design relevant

[1]We emphasize that the term "declare" does not imply a public declaration or a declaration before research takes place. A researcher may declare the features of designs in our framework for their own understanding and declaring designs may be useful before or after the research is implemented.

information in advance. For some forms of structured exploratory work, for example using cross validation techniques, an exploration strategy may also be described in advance. For other forms of exploratory work, however, researchers may simply not have enough information about possible quantities of interest to declare a design in advance. Although in some cases the design may still be declared ex post, in others it may not be possible to fully reconstruct the inferential procedure after the fact. For instance although researchers might be able to provide compelling grounds for their inferences, they may not be able to describe what inferences would have been made for all possible data. Thus variation in research strategy limits the utility of our procedure for different types of research. In the conclusion, we discuss possible implications of this.

We define research designs through a set of features that include the population, potential outcomes function, sampling procedure, estimands, assignment procedure, and estimators. Not all research designs will include all of these features.[2] We show how each of these features can be defined in mathematical notation as statistical distributions and mappings. In addition, we formalize the notion of "diagnosands," or statistical summaries of the design such as the power of the design, the bias of the estimator, or the expected mean squared error (MSE) of the estimates with respect to an estimand. A design is "θ-complete" in our framework when a diagnosand θ can be estimated from the declared features of the design. We highlight that we do not have a general notion of "complete" design, but rather adopt an approach in which the purposes of the design determine which features must be declared.

Our notion of θ-completeness requires that the question "what is the value of θ?" is answerable. Design descriptions might not be θ-complete for two reasons: not enough information has been provided or θ might be undefined for the design. If θ is power, the probability of correctly rejecting a false null hypothesis, a design might not be power-complete because some crucial piece of information is missing (e.g., the distribution of a test statistic is not specified) or because power is an undefined concept in the design (e.g., the design does not involve hypothesis testing). Additionally, even when a design is θ-complete, θ itself might be a poor indicator of a design's inferential value. For example, if p-values are calculable but incorrect,[3] then the estimate

[2] We discuss in Section 3 how to declare and diagnose designs that do not include sampling procedures, assignment procedures, potential outcomes.

[3] In the sense that the p-values do not accurately represent the probabilities under the appropriate null hypotheses of obtaining test statistics at least as extreme.

of power could be misleading. Thus, good diagnosis requires a judicious choice of diagnosands.

Authors can assess the properties of θ-complete research designs in order to improve designs before implementation. Readers and replication authors can diagnose designs before or after implementation in order to select and critique studies based on their designs rather than their results. In general, we do not provide specific guidance on the set of diagnosands that must be calculable in order for a design to be complete "enough." Domain-specific standards might be agreed upon among members of research communities. A standard set might include power, bias, root mean-squared error, and coverage. Others who concerned about the policy impact of a given treatment might require a design that is θ-complete for an out-of-sample diagnosand such as bias relative to the population average treatment effect.

Diagnosis can be executed analytically for simple designs or through Monte Carlo simulation for more complex designs. We provide an algorithm for simulation that produces diagnosand estimates as well as bootstrapped estimates of simulation uncertainty.

Our framework makes five principal contributions: it enables the diagnosis of designs in terms of their probative value; it enhances research transparency by making design choices explicit; it assists learning about the properties of research designs; it assists in the improvement of research designs through comparison with alternatives; and it provides tools to assist principled replication and reanalysis of published research.

1. Formally Defining a Research Design and Its Diagnosands

How do we know when a study is likely to provide good answers to the questions it poses? When designing research, budget-constrained scholars must choose among alternative sampling strategies, estimators, and (if the study is experimental) treatment assignment schemes. Current practice focuses narrowly on some aspects of a research strategy while neglecting other equally important features. For example, many funding agencies require power calculations but do not require a definition of the estimand. As noted above, it is possible in such cases to design "good" studies that produce over-confidence in the wrong answer, if the estimator is both biased and precise.[4] To understand the expected bias of a study, we must at the very least know its estimand.

[4]Difference-in-difference designs, for example, are a popular choice for policy impact evaluations but have been shown to exhibit high false positive rates (Bertrand, Duflo and Mullainathan, 2004). That is, they commonly exhibit coverage rates that are too low.

By characterizing the analytically-relevant features of a research design we can understand a design's potential to provide useful answers.

1.1 Defining a Research Design

In order to be able to diagnose a research design's statistical properties, design features must be declared exactly. Here, we provide a formalization of research designs at a high level of abstraction and in doing so provide a definition of designs whose analytic features can be simulated and diagnosed. This discussion draws on formal accounts of research designs in Imbens and Rubin (2015) and especially Pearl (2009).

The key feature of the formal definition employs an augmented "Pearlian DAG" (Dawid, 2010) that characterizes relations over three sets of variables: a set of endogenous variables, Y, a set of background variables, X, and a set of possible "manipulands" (or "intervention nodes" in Dawid, 2010), Z. The set Z includes all conditions that are notionally under the control of a researcher, whether or not that control is exercisable in practice (or exercised).[5] The set Y includes all outcomes that are dependent on possible manipulands, including, for example, membership of a study sample, treatment compliance, mediating variables, data missingness, and even estimates. Given this structure, estimands are the quantities a researcher wants to estimate, such as causal quantities like the average treatment effect, but also non-causal quantities such as the average height of a population. These can be defined as summaries of potential values of Y given different values of X and Z, though they may also depend on realizations of data. Summary statistics are calculations that can be made based on the realized data alone. Estimates are summary statistics of realized data that are associated with estimands. Other summary statistics, such as a p-value are associated with an implied hypothesis.[6] Some summary statistics, such as "whether a calculated p-value is less than 0.05," are useful for assessing the properties of a design. The distributions of such "diagnostic statistics" over repeated draws form the basis for "diagnosands." For example, the expected value of the diagnostic statistic "whether a calculated p-value is less that 0.05" is the diagnosand "statistical power."

[5] "Augmented" graphs enumerate these manipulands explicitly whereas non-augmented Pearlian graphs assume that interventions are possible on all endogeneous variables in the system.

[6] In our formulation of a design, hypotheses do not play a distinct role beyond their implicit role in the implementation of tests.

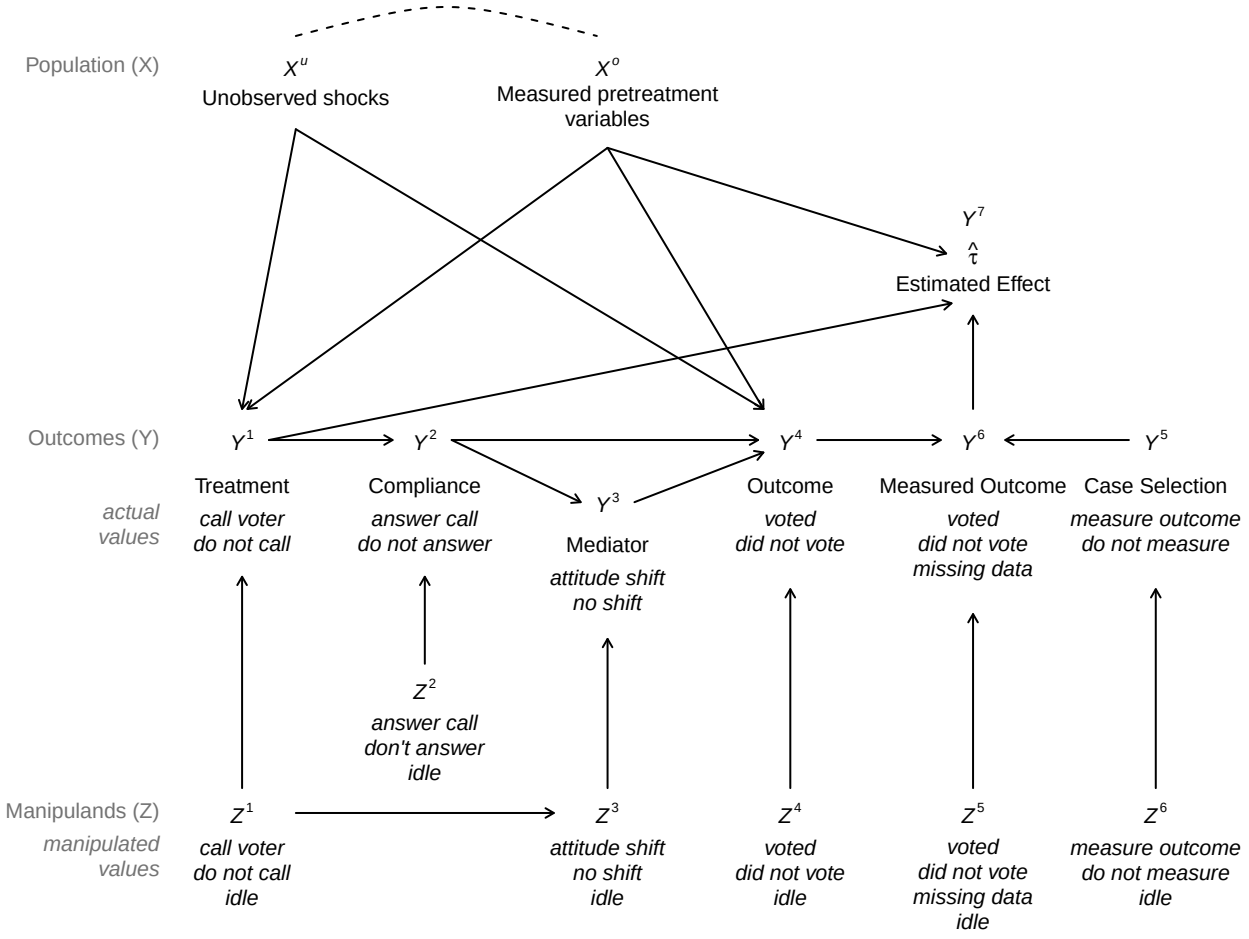

Figure 1: Illustration of a Research Design Represented as a Directed Acyclic Graph (DAG). Black arrows denote possible causal relationships. Each variable is thought of as a vector of outcomes for a population and so spillovers are naturally admitted by the DAG (the potential outcomes of each unit are a function of the entire vector of assignments). Noncompliance is captured by including a compliance potential outcome (Y^2) in the DAG; mediation is captured by including an intermediate potential outcome (Y^3); attrition is captured by including a measurement potential outcome (Y^5). The graph indicates a possibly complex assignment strategy in which assignment of a mediator (Z_3) possibly depends on assignment to a treatment (Z_1).

Figure 1 illustrates a general structure that could represent many research designs focusing on the question of how political campaigns mobilize supporters. The directed acyclic graph (DAG) of this setting includes possible relationships between manipulands Z and background features X to outcomes Y and ultimately to estimates $\hat{\tau}$. Note that missing arrows describe possible exclusion restrictions, but not all of these are necessary (for example X^o could also affect Y^2 directly, or Y^5 in the case of attrition, or it could affect Z^1, if blocking is used). The DAG could be used to represent a research design in which campaign calls are randomly assigned by the

research team or a descriptive design in which the campaign's choices are outside of researcher control (in which case Z^1 would be set to `idle`, meaning it takes on values naturally). We discuss this illustration further in Appendix B.

With this structure in mind, we now formally define the key features of a research design:

1. **Variables**. Let D denote a collection of variables indexed by j and let $\mathcal{V}(D^j)$ denote the space of possible values for variable D^j. Relations between variables in D are represented by a directed acyclic graph (a DAG) which records conditional independencies of the joint distribution of variables. Let $pa(D_j)$ denote the "parents" of D_j, that is, the set of variables on which D_j depends directly.

   We assume that the set D can be partitioned into three classes.

   - **Population**. Set X is a collection of observed and unobserved baseline variables (or "background" variables), that are not treated as being even notionally under the control of researchers. Let p_X denote a probability distribution over possible values of X, $\mathcal{V}(X)$.
   - **Potential Outcomes**. Set Y contains outcomes of interest such as whether a drug is offered by a doctor, and whether a drug is ingested. We let the mapping $f_y^j : \times_{D^k \in pa(Y^j)} \mathcal{V}(D^k) \rightarrow \mathcal{V}(Y^j)$ denote the potential outcomes function of variable Y^j. We use f_y to denote the collection of potential outcome functions. Note that by positing such a function we assume a *functional* causal model (Pearl, 2009): potential outcomes of variables in Y are determined by their parents; any randomness in the potential outcomes is introduced via distributions on X.
   - **Manipulands.** Set Z contains variables that are under the "notional control" of researchers. For each non-terminal Y^j in set Y there is a parent variable Z^j that characterizes potential manipulations of Y^j. We assume that $\mathcal{V}(Z^j)$ is an augmentation of $\mathcal{V}(Y^j)$ that allows each unit to take value `idle` (or "$\oslash$"). For example $\mathcal{V}(Y^j) = \{0,1\}^3 \rightarrow \mathcal{V}(Z^j) = \{0,1,\oslash\}^3$. The idea of "notional control" is that controlled manipulation of the variable is conceivable in the sense described by Holland (1986), and not that such manipulation is practically possible.

Following Pearl, we assume that the potential outcomes mapping satisfies "effectiveness"—meaning that units that are assigned to manipuland conditions that are notionally under a researcher's control receive those precise manipuland conditions[7]— and "modularity" —meaning that Y^j is independent of Z^k for all $j \neq k$ conditional on $pa(Y^j) \cap Y$ (Dawid, 2010); informally, the response of an outcome to a treatment is the same whether or not the treatment is directly manipulated.

2. **Manipuland Assignment** (Probability distribution). Let $p_{Z^j}(\cdot|pa(Z^j))$ denote a probability distribution over the possible values of a manipuland, $\mathcal{V}(Z^j)$, given the parents $pa(Z^j)$. Z^j has no parents when assignment does not depend on background characteristics. In cases in which a manipuland is not manipulated, p_{Z^j} places all probability mass on the condition `idle` for all units. The **manipulation strategy**, p_Z, is a collection of assignment distributions for all $z^j \in Z$. The manipulation strategy describes not just how manipuland conditions are assigned to units but also *whether* manipulands are manipulated. Note that a sampling strategy is a part of the manipulation strategy that determines the units over which data is gathered. Since sampling strategies play privileged roles in designs below we use use p_S to describe the sampling strategy and when there is no risk of confusion, p_Z to refer to non-sampling components of the manipulation strategy.

3. **Data**. Let p_D denote a probability distribution over D induced by p_X, p_Z, and f_y. Let $\mathcal{D}$ denote the set of all possible data ("superdata") and d a particular realization of data. Note that $\mathcal{D}$ includes all potential outcomes given different possible assignments z and background characteristics x.

4. **Estimands** (Function). Let $\tau(\mathcal{D}, d, p_Z)$ denote an estimand and (τ) a collection of such estimands. An estimand is a summary of potential outcomes, recorded in superdata. It may also depend on realizations of assignments, recorded in d, and be sensitive to assignment schemes (p_Z). See Appendix A.1 for a discussion of different classes of estimands.

5. A **summary statistic** (Function) $\phi(d, p_Z)$ is a function and (ϕ) a collection of functions of re-

[7]This assumption does not rule out the possibility of non-compliance in a usual sense: for manipuland conditions outside a researcher's control, the manipuland condition assigned and manipuland condition received should be thought of as separate elements of Y.

alized data and the manipulation strategy; and an **estimate** $\hat{\tau}(d, p_Z)$, is a summary statistic and $(\hat{\tau})$ a collection of summary statistics that has an associated estimand τ.

6. **Strategy**. Let the triple $\Sigma \equiv < p_Z, (\hat{\tau}), (\phi) >$ denote a researcher's strategy. The strategy includes both the manipulation strategy—including both treatment assignments and measurement—and the analysis strategy.

7. **Design**. Let the tuple $\Delta \equiv < p_X, f_y, \Sigma, (\tau) >$ denote a study design, which consists of a set of beliefs about the world p_X, a set of conjectures about potential outcomes functions f_y, a strategy Σ to manipulate and interrogate the world, and a set of target quantities of interest, (τ). Each element in this tuple depends on previous elements in the tuple but not vice versa.

Designs need not include all six elements of a design to be described in this way. Designs without random assignment of a treatment may nevertheless have manipulands related to sampling or data collection. Designs without sampling can be described as designs in which all units are sampled. Descriptive designs include outcomes that have no manipuland parents and are only functions of population variables. We describe in Section 3 how to characterize a range of designs that do not include treatment assignment, sampling, potential outcomes, or even variation in outcomes.

Our definition of a research design captures its analysis-relevant features; it does not describe substantive elements, such as how interventions are implemented or how outcomes are measured. Yet many other aspects of a design that are not explicitly labeled in these features nevertheless enter into this framework if they are analytically relevant. For example, logistical details of data collection such as the duration of time between a treatment being administered and endline data collection do not obviously appear in the ten design features. However, the duration might enter into the potential outcomes function if the longer time until data collection affects subject recall of the treatment.

1.2 Diagnosing the Properties of a Research Design

To allow for *diagnosis* of a design, we introduce two further concepts, both functions of research designs:

1. **Diagnostic Statistic** (Function) $t(d, \Delta)$ is a function of a design and data. Some diagnostic statistics, such as the mean of an outcome, depend on data only. For example, a researcher might define the z-statistic (a summary statistic in the design), given data as $z(d)$. This in turn implies a p-value, for example $p(d) = 1 - \Phi(z(d))$, for a one sided test, where Φ denotes the cumulative normal distribution. A third party examining the design may not agree with the manner in which $p(d)$ is defined in the design, but they can still use $p(d)$ to generate a diagnostic statistic. Although this statistic is not part of the design, it is nevertheless calculable from the data, given the design, as:

$$s(d) = \mathbb{I}(p(d) \leqslant 0.05)$$

Diagnostic statistics can also depend on potential outcomes as well as realized data. For example the statistic:[8]

$$e_{ATE}^{DIM}(d, \Delta) = \mathbb{E}_{p_X}(\tau_{ATE}(f_y) - \hat{\tau}_{DIM}(d))$$

denotes the average difference, over possible populations, between the ATE and the estimated ATE, as estimated using a difference-in-means (DIM) estimator.

2. **Diagnosand** (Function) $\theta(d, \Delta)$ is a summary of the distribution of a diagnostic statistic. In most cases we are interested in the expected value of a diagnostic statistic given the distribution of the data implied by a design and prior uncertainty. For example, given the diagnostic statistics described above, (expected) *bias*[9] in the population treatment effect is: $\mathbb{E}_{f_D} e_{ATE}^{DIM}(d, \Delta)$ and (expected) *power* is: $\mathbb{E}_{f_D} s(d)$.

What diagnosands should researchers choose? Although researchers commonly focus on statistical power, a large range of diagnosands can be examined. These include the expectation and standard deviation of the estimates; the expectation and standard deviation of the estimands

[8]Note on notation. We use $\mathbb{E}$ to denote both expectations and averages: $\mathbb{E}_A x$ denotes the average of x in group A; $\mathbb{E}_f x$ denotes the expected value of x given distribution p_Z:

- $\mathbb{E}_A x = \frac{1}{\#|A|} \sum_{i \in A} x_i$
- $\mathbb{E}_f x = \int x f(x) dx$.

[9]"Expected" is not redundant here under the interpretation of p_X, reflecting prior uncertainty.

(which may vary across draws of the population, for example); statistical power; expected imbalance; and a set of diagnosands defined with respect to the defined estimates and estimands, including bias, root mean squared error, coverage, "type S" error rate, and the exaggeration ratio (Gelman and Carlin, 2014). This set of statistics allows researchers to understand the properties of the estimates across possible realizations of the data and how successful their manipulation and estimation strategies are at estimating estimands. Though these are frequentist properties, they can applied equally to frequentist and Bayesian estimation strategies with judicious choices of diagnosands, as we illustrate below (see Rubin, 1984).

Design diagnosis also provides a framework for assessing the "external validity" of a research design in the limited sense of the ability to make inferences to an out of sample estimand. In some cases, this refers to new subgroups within the population. For example, survey experiments conducted on convenience samples online may or may not reveal insights that generalize to the U.S. population at large. If treatment effects vary substantially across individuals, but the study is conducted on a sample that has relatively high average treatment effects, then the study might exhibit low external validity. The extent to which such discrepancies represent a scientific problem can be explored by changing the sampling strategy and potential outcomes functions.

Diagnosands can also be defined for properties that reach beyond classical statistics. For example one might define a study as "conclusive" if some evidence is observed, whether or not formal hypothesis tests are conducted, an inference as "robust" if the same inference is made under different analysis strategies, or an intervention as having "value for money" if some set of estimates have some minimal magnitude. In these cases a diagnosis reports the chances that a study will be considered conclusive, an inference considered robust, or an intervention deemed to have value for money. These three diagnosands depend on observed data only. More subtle analogues can be defined that also make use of potential data, for example one might define a diagnosand as the chances that an intervention is *correctly* considered value for money.

1.3 What is a Complete Research Design Declaration?

A declaration of a research design that is in some sense complete is required in order to implement it, communicate its essential features, and to assess its properties. Yet existing definitions make clear that there is no single conception of a complete research design that is satisfactory

for all purposes: the Consolidated Standards of Reporting Trials (CONSORT) Statement widely used in medicine includes 22 features and other proposals range from nine to 60.[10]

In current practice, research designs are commonly declared in terms of the elements required to implement the study. The set of features that are often described include the assignment procedure including sampling and treatments as well as the estimation strategy. This corresponds formally to the design Σ, including the manipulation strategy p_Z, estimator $\widehat{\tau}$, and summary statistics ϕ. In some cases, minimal descriptions of potential outcomes mappings f_y in terms of expected effect size or direction are also included. However, though a design described in this minimal way could be implemented, the researcher might not be able to diagnose many of its properties ex ante.

We propose instead a conditional notion of completeness: we say a design is "θ-complete" if a diagnosand θ can be estimated from the declared design. Consider for example the diagnosand statistical power. Power is the probability that a p-value is lower than some value, defined over all possible realizations of the data, conditional on beliefs about the world and an estimator, $\Pr(\phi(d) < \alpha \mid p_X, f_y, \hat{\tau})$. This notion of completeness does not imply a completeness ordering: for example a design that does not specify an estimand may be power-complete but not bias or RMSE complete; a design that does not specify a statistical test may be bias complete but not power complete.

2. Declaring and Diagnosing Research Designs in Practice

We described an approach to defining a research design mathematically in Section 1 and in what follows we demonstrate how a design can be declared as an *object* in computer code and then simulated in order to diagnosing its properties.

2.1 Characterizing Research Designs in Code

Characterizing or "declaring" designs in a common computer-based syntax has three main advantages. First, when a design is transformed into a code object its diagnosands can be quantified. For example, the root mean squared error (RMSE) of RMSE-complete designs is not only defined, it can be numerically estimated through Monte Carlo analysis on a computer. Second,

[10] See "Pre Analysis Plan Template" (60 features); World Bank Development Impact Blog (nine features).

objects created through a common computer language are directly comparable. The core structure of the language we propose provides leeway to accommodate user-defined functions, but also easy-to-use defaults. Third, the ability to re-use the design code throughout the lifecycle of a study presents a practical advantage. The very same code used, say, to declare the assignment mechanism in a randomized trial can also be used to implement sampling given population data, randomize treatment assignment given sample data, and implement analysis given outcome data.

For simplicity the assignment process is divided into two parts in order to privilege sampling procedures. In practice researchers often engage in sampling and make all other assignment decisions, including data collection decisions, conditional on sampling. Privileging this design feature will simplify the declaration of many designs although as seen from the formalization, it is not analytically necessary.

In what follows, we demonstrate how each feature can be defined in code, with an application in which the assignment procedure is known. This could represent an experimental or quasi-experimental design.

p_X **The population**. Defines the population variables, including both observed and unobserved X. In the example below we define a function that returns a normally distributed variable of a given size. Critically, the declaration is not a declaration of a particular realization of data but of a data generating *process*. Researchers will typically have a sense of the distribution of covariates from previous work, and may even have an existing dataset of the units that will be in the study with background characteristics. Researchers should assess the sensitivity of their diagnosands to different assumptions about p_X.

```
my_population <- function(size) { data.frame(u = rnorm(size)) }
population <- declare_population(
  custom_population_function = my_population, size = 100)
```

p_S **Assignment 1: The sampling strategy**. Defines the distribution over possible samples for which outcomes are measured. Formally p_S is a component of p_Z, though it is given the special attention paid to it in many studies. In the example below each unit generated by p_X is sampled with 10% probability. Again `my_sampling` describes a strategy and not an actual sampling.

```
my_sampling <- function(data) {
  rbinom(n = nrow(data), size = 1, prob = 0.1) }
sampling <- declare_sampling(sampling_function = my_sampling)
```

p_Z **Assignment 2: Treatment assignment**. Defines the strategy for assigning variables under the notional control of researchers. In this example each sampled unit is assigned to treatment independently with probability 0.5. The default assumption in our code is that treatment assignment takes place after sampling though as a general matter this need not be the case. In designs in which the sampling process or the assignment process are in the control of researchers, p_z is known. In observational designs, researchers either know or assume p_z based on substantive knowledge.

```
my_assignment <- function(data) {
  rbinom(n = nrow(data), size = 1, prob = 0.5) }
assignment <- declare_assignment(assignment_function = my_assignment,
                                 condition_names = c(0,1))
```

f_y **The potential outcomes**. The potential outcomes function defines conjectured potential outcomes given manipulands Z and parents. In the example below the potential outcomes function maps from a treatment condition vector (Z) and background data u, generated by p_X, to a vector of outcomes. In this example the potential outcomes function satisfies a SUTVA condition—each unit's outcome depends on its own condition only, though in general since Z is a vector, it need not.[11] It also assumes that potential outcomes depends on treatment assignment and not on sampling. Again, the declaration describes the function and not a particular set of potential outcomes.

```
my_potential_outcomes <- function(data) { with(data, Z * 0.25 + u) }
potential_outcomes <- declare_potential_outcomes(
  potential_outcomes_function = my_potential_outcomes,
  outcome_variable_name = 'Y',
  condition_names = c(0, 1))
```

In many cases, the potential outcomes function (or features of it) is the very thing that the study sets out to learn, so it can seem odd to assume features of it. We suggest two

[11] For an example of a function that does not satisfy SUTVA consider $Y = Z + \min(Z \times u)$, for vectors Y, Z, u.

approaches to developing potential outcomes functions that will yield useful information about the quality of designs. First, set a potential outcomes function in which the variables of interest are set to have no effect on the outcome whatsoever. Diagnosands such as bias can then be assessed without having to assume a particular relationship between treatments and outcomes. This approach will not work for some diagnosands such as power or Type-S errors. Second, consider setting a series potential outcomes functions that correspond to competing theories. This enables the researcher to judge whether the design yields answers that help adjudicate between the theories and whether the design has desirable properties (i.e. sufficient power) under the potential outcomes implied by each theory.

τ **The estimands**. The estimand function τ creates a summary of potential outcomes using 'superdata' that can be generated from the elements declared above. In principle the estimand function can also take realizations of assignments as arguments, in order to calculate post-treatment estimands. Below, the estimand takes the mean difference between the potential outcomes for units in a treated condition and units in a control condition.

```
my_estimand <- function(data) { with(data, mean(Y_Z_1 - Y_Z_0)) }
estimand <- declare_estimand(estimand_function = my_estimand,
                             potential_outcomes = potential_outcomes)
```

$\phi, \hat{\tau}$ **The summary statistics** are functions that use information from realized data and the design, but do not have access to the full schedule of potential outcomes. In the declaration we associate estimators with estimands and we record a set of summary statistics that are required to compute diagnostic statistics. In the example below an estimates function takes data and returns an estimate of a treatment effect using regression as well as a set of associated statistics, including the standard error, p-value, and the confidence interval.

```
my_estimates <- function(data) {
  reg <- lm(Y ~ Z, data = data)
  phi <- as.list(summary(reg)$coefficients["Z", ])
  c(est = phi$Estimate, se = phi$"Std. Error", p = phi$"Pr(>|t|)") }
estimator <- declare_estimator(estimates = my_estimates, estimand = estimand)
```

These six features represent the study. In order to assess the completeness of a declaration and

to learn about the properties of the study, we also define functions for the diagnostic statistics, $t(D, Y, f)$, and diagnosands, $\theta(D, Y, f, g)$. These could be coded as a single function for simplicity. An example function for calculating bias as a diagnosand is:

```
diagnosand <- declare_diagnosand(
    diagnostic_statistic_text = "est - estimand",
    summary_function = mean
  )
```

These seven functions could be written in many code languages. In the companion software for this paper, `DeclareDesign` (Blair et al., 2016*a*), we implement it for the widely-used `R` platform.

In practice we advocate declaring simple designs using "design templates" that create full design objects given a small number of critical arguments. In the companion software we include a function `quick_design` that lets one generate one or many designs from a "design template." For example the template `k_arm_design_template` can be used together with `quick_design` to generate one or more multi-arm designs. In section 4 we illustrate the use of such templates to compare multiple designs.

Design diagnosis through simulation places a burden on researchers to come up with realistic values for p_X, p_Z, and f. The utility of any particular diagnosis for making research decisions of course depends on the plausibility of the design declaration: the value of the diagnosis is only as good as the inputs. This problem is familiar from power calculation. Power calculators also make implicit assumptions about p_X, p_Z, and f, yet these choices are typically hidden. A principle advantage of our approach is that such implicit assumptions are rendered explicit.

We advocate that researchers diagnose designs not for their single best guesses of p_X, p_Z, and f but rather for a range of plausible values indicated by past studies, pilot data, and theory. The quality of the research strategy can then be judged by how it performs across this range of scenarios.[12] In a sense, what design declaration offers is not a tool to establish that a design has desirable qualities but a tool to lay bare *under what assumptions* a design has desirable properties.

[12]This is distinct from sensitivity analysis as sensitivity analyses are conducted conditional on realized data.

2.2 Estimating Diagnosands through Simulation

Research design diagnosis—the estimation of diagnosands—can be accomplished in two ways: analytically and through Monte Carlo simulation. For simple designs there may be analytical solutions to the diagnosand, for example statistical power as a function of the sample size. In this case, the researcher can derive the solution and plug in values of those variables in order to report the values of the diagnosands. In other cases, indeed for most moderately complex designs, either there are not closed-form solutions or they are difficult to derive.

For complex designs, we recommend Monte Carlo simulation of research designs to estimate the value of diagnosands. The researcher, using the population function defined in computer code, draws a population, samples units (if relevant), and within that sample assigns treatments (if any) using the assignment function, calculates observed outcomes from the potential outcomes function, estimates from the estimator function, and then estimands, diagnostic statistics, and diagnosands from their respective functions. This nested set of calculations is analogous to running the study over and over. Here, instead of conditioning on a given dataset we also simulate the data-generating process, which enables us explore the properties of the design across potential draws of the data.

We now present the procedure for calculating a single draw of a diagnostic statistic formally:

1. Draw a population x using p_X.
2. Calculate potential outcomes Y using f_y, given x and record these in superdata $\mathcal{D}|^x$.
3. Draw sampling and treatment assignments z using p_S and p_Z.
4. Calculate observed outcomes using f_y and z. This generates dataset d.
5. Calculate estimands using $\tau(\mathcal{D}, p_Z, d)$.
6. Calculate summary statistics, $\widehat{\tau}$, ϕ, using d and p_Z.
7. Calculate a diagnostic statistic t using d and τ, and perhaps p_S and p_Z.

In the simplest case, estimating a diagnosand is straightforward. We conduct steps 1-7 m^{pop} times in order to obtain m^{pop} diagnostic statistics. The distribution of these diagnostic statistics can then be summarized in order to estimate diagnosands, typically by calculating their mean or standard deviation. Some simulations will have more complex structures: within a given draw of x, samples might be drawn m^{samp} times and within each sample, assignments might be allocate

m^{assign} times. Researchers may be interested in diagnosands that depend on this structure, such as what is the standard deviation of sample average treatment effects.

Diagnosands can be estimated with arbitrary precision by increasing the number of draws at each level (population, sampling, assignment). However, simulations are often computationally expensive. In order to assess whether researchers have conducted "enough" simulations to be confident in their diagnosand estimates, we recommend estimating the sampling distributions of the diagnosands via the nonparametric bootstrap. With the estimated diagnosand and its standard error, we can construct a confidence interval to make a decision about whether the range of likely values of the diagnosand compare favorably to reference values such as statistical power of 0.8. We emphasize that this confidence interval reflects both estimation uncertainty (simulation error) and fundamental uncertainty (true variability in the diagnosand, for example across possible population draws).[13]

Simulation has two chief advantages. First, it is straightforward for scholars to diagnose designs without deriving closed-form expressions for diagnosands. They need only declare in code the six features of their design as well as their diagnosands. Second, it allows researchers to quickly diagnose variations of their designs with a few small changes to the computer code.

3. Declaring Common Observational Research Designs

The framework we propose can be used to declare and diagnose a range of research designs typically employed in the social sciences. Whereas the most obvious fits may appear to be randomized experiments, in which researchers control treatment assignment, or quasi-experimental designs, in which treatment processes are known, the scope for design declaration is considerably more general. Below we sketch declarations of designs for cases in which, there is no notion of counterfactuals, no notion of potential outcomes, no experimental control, no known assignment procedure, no null hypothesis testing procedure, and no assumption of variation in the observed outcomes.[14]

[13]This procedure depends on the researcher choosing a "good" diagnosand estimator. In nearly all cases, diagnosands will be features of the distribution of a diagnostic statistic that, given i.i.d. sampling, can be consistently estimated via plug-in estimation (for example taking sample means). Our simulation procedure, by construction, yields i.i.d. draws of the diagnostic statistic within each level.

[14]Many designs will not have sampling procedures. We note that these can also be described as designs in which all units are sampled and thus their declaration and diagnosis is straightforward.

3.1 Descriptive Inference

Many research projects are not centered around the estimation of causal effects. Descriptive research questions often center on measuring a parameter in a sample or in the population, such as the proportion of voters in the United States who support Hillary Clinton. Although seemingly very different from designs that focus on causal inference—because often there are no explanatory variables—the formal differences are not great. In particular descriptive studies may include possible manipulands (for example ways of phrasing a survey question) and so may involve potential outcomes (if for example the measure itself is an outcome). The key difference is that the estimands of descriptive studies depends on outcomes that have been realized and not on counterfactuals. Formally then the estimands are a function of d and not $\mathcal{D}\backslash d$ and are a special case of the general class of designs formalized above

In Appendix C.1 of the appendix, we examine an estimator of candidate support that conditions on being a "likely voter." For this problem the data that help researchers predict who will vote is of critical importance. In the example, analysts declare the full measurement procedure, including the possibly imperfect procedure for determining who is a likely voter, and uses that to assess the risk of falsely concluding that Hillary Clinton's general election support is above 50%.

3.2 Discovery

In some research projects the ultimate hypotheses that are assessed are not known at the the design stage. Some inductive designs are entirely unstructured and explore a variety of data sources with a variety of methods within a general domain of interest until a new insight of some type is uncovered. Yet many can be described in a more structured way. In studying textual data, for example, a researcher may have a procedure for discovering the "topics" that are discussed in a corpus of documents. Before beginning the research, the set of topics and even the number of topics is unknown. Instead, the researcher selects a model for estimating the content of a fixed number of topics (i.e. Blei, Ng and Jordan, 2003) and a procedure for evaluating the model fit that is used to select which number of topics fits the data best. Such a design is inductive, yet the analytical *procedure of discovery* can be described and evaluated.

Exploratory analysis after data is collected is ubiquitous in applied research including model

selection and the detection of data anomalies such as outliers and heteroskedasticity. Discovery is at the heart of this process. In section C.2 of the appendix, we given an example of a design declaration for an exploratory data analysis *procedure* in which in a first stage the researcher explores possible analysis strategies on half of the data and in the second stage apply their preferred procedure to the second half of the data. Split-sample procedures such as this enable researchers to learn about the data inductively while still protecting against Type I errors (Fafchamps and Labonne, 2016). We show how the researcher can evaluate if the procedure is bias-reducing in their context and how much it reduces the power of the design given their hypothesis. Any exploratory procedure in which the domain of exploration (for example, the set of models or tests that will be conducted) and the decision rules (how the researcher selects among models or changes the analysis in response to test values) are known can be explored.

3.3 Matching

In some observational research, assignment processes are not known. Instead, at the design stage, researchers seek to identify conditions under which as-if random assumptions can be made. In such cases, although researchers do not control assignment it is still possible to declare an assumed assignment process. Designs that estimate causal effects using regression with covariates to condition on aspects of an assumed assignment process can similarly be declared and diagnosed.

A matching design declaration provided in section C.3 of the appendix shows how under some assumptions, matching improves mean-squared-error relative to a naive difference-in-means estimator of the treatment effect on the treated (ATT), but can nevertheless remain biased if the matching algorithm does not successfully pair units with equal probabilities of assignment.

3.4 Regression Discontinuity

Some observational strategies employing a potential outcomes framework do not rely on the assumption that the assignment procedure is even as-if random. Consider, for example, the regression discontinuity design in which causal identification is premised on the claim that potential outcomes are continuous at a critical threshold (and not from a claim of random placement of units around a threshold). Although there is no random assignment, it is still possible to describe

both the supposed data generating process and the potential outcomes. Variability in estimates derives from sampling variability even if there is no variation induced by treatment assignments.

In section C.4 of the appendix, we provide an example of a regression discontinuity design that assigns treatment deterministically when units are located beyond a threshold of 0 on the running variable. Diagnosis of such a design makes clear that the estimand involved in many regression discontinuity designs is rarely an average of potential outcomes of all units, but rather an unobservable quantity defined at the limit of the discontinuity. Our framework is entirely amenable to such diverse estimands.

3.5 Model-Based Estimands

Many observational studies seek to make causal claims but do not explicitly employ the potential outcomes framework, instead describing estimands in terms of model parameters. Consider a study that seeks to estimate parameter β from a model of the form:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

In all our examples so far we have explicitly defined an estimand in terms of potential outcomes. What is the estimand here? If we believe that Equation 1 describes the true data generating process then β *is* an estimand: it is the true (constant) marginal effect of x on y. But what if we are wrong about the data generating process? We run into a problem if we want to assess the properties of strategies under different assumptions about data generation if the estimand itself depends on the data generating process.

An approach that can be used without assuming we know the data generating process is to define the estimand as some summary of differences in potential outcomes across conditions, β and then assess how well an OLS model estimates β under different conditions. For example we might define α and β as the solutions to.

$$\min_{(\alpha,\beta)} \sum_i \int (y_i(x) - \alpha - \beta x)^2 f(x) dx$$

Here $y_i(x)$ is the (unknown) potential outcome for unit i in condition x. Estimand β can

be thought of as the coefficient one would get on x if one were to able to regress all possible potential outcomes on all possible conditions for all units (given density of interest $f(x)$). An alternative might be to imagine some analogue of the ATT estimand, for example for some x_i defined on the real line we might define $E(Y_i(x_i) - Y_i(x_i - 1))$ where x_i is the observed treatment received by unit i.

Such estimands require estimators that return scalar estimates of the estimands. In section C.5 of the appendix we declare a design in which the properties of a regression estimate are assessed under the assumption that in the true data-generating process y is in fact a nonlinear function of x. Diagnosis of the design shows that under uniform random assignment of x, the linear regression returns an unbiased estimate of a (linear) estimand, even though the true data generating process is non linear. Interestingly, with the design in hand, it is easy to see that unbiasedness is lost in a design in which different values of x_i are assigned with different probabilities.

3.6 Bayesian Estimation Strategies

In addition to modes of analysis that employ a classic null-hypothesis testing approach to statistical inference, our framework can also be of use to Bayesian strategies. A range of interesting diagnosands can be investigated to conduct ex ante sensitivity analysis and to assist Bayesian model specification. For example, researchers can investigate which kinds of prior and likelihood specifications will lead to the greatest learning, measured either as a shift in the first moments or as a reduction in the second moments when moving from the prior to the posterior distribution over some parameter.

In section C.6 of the appendix, we declare a Bayesian descriptive inference design, in which the researcher seeks to estimate an underlying probability of success in a population using a beta-binomial model and a random sample of successes and failures. Because the underlying DGP is a probit model, the DGP is slightly misspecified and we can see that a small amount of bias arises. The diagnosis also reveals that the maximum a posteriori (posterior mode) is a less biased estimator of the estimand than the posterior mean, due to the skew in the posterior distribution. Flat priors enable the most learning, providing the greatest reduction in posterior variance relative to the prior, and allowing for the largest shift in the first moment of the posterior

relative to the prior. The diagnosis shows that, given his or her assumptions about the DGP, the researcher would do best by employing flat priors and analyzing the posterior mode, rather than the mean.

3.7 Process Tracing

While the strategies described above vary in fundamental ways, they all adhere to a recognizable quantitative mode of analysis in which inferences are drawn by examining patterns across multiple cases. In contrast, many qualitative researchers employ frameworks that may seem incompatible with the type of design declaration we have described, sometimes seeking to make an inference on a single case. For qualitative designs that aim to confirm the presence or absence of a causal relationship (i.e., that are not focused on theory generation), formal design declaration and diagnosis may still be of some use.

There are many distinct approaches to process tracing, but for concreteness consider, a stylized "process-tracing" design similar to ones described for example by Mahoney (2012) or in the appendix to Bennett and Checkel (2014). A researcher selects a case in which some outcome is observed (a revolution, say) and some possible driver is present (a strong middle class, say). The researcher seeks evidence in archives that they believe to be "smoking gun evidence" (Van Evera, 1997) that the driver was indeed important for the outcome—for example they look for evidence that the revolution was financed by domestic industry—and are prepared to draw different inferences depending on what they find in this causal process observation. In this case there is an implicit potential outcomes function, an estimand (whether the driver was a cause or not), a sampling strategy, and an estimation strategy (e.g. conclude that the candidate driver was causal if the causal process observation (CPO) is seen, remain agnostic otherwise). In this design there is only one case and no variation in the observed outcomes. Power is not a meaningful diagnosand in this context, yet other diagnosands can come into play: for example expected error in inferences. As with all the other designs, sensitivity of the diagnostics to researcher assumptions about the causal model can be assessed.

In section C.7 of the appendix we give an example of a declaration of such a design assuming a researcher uses a simple (non Bayesian) inference strategy, updating if and only if they see the smoking gun CPO. As in many qualitative designs, the dataset is small and composed entirely of

boolean data (logical true or false values). Diagnosis of the design shows that, if the researcher's beliefs about the CPO are correct, their inference will be unbiased in the cases in which the CPO is observed but not in those cases in which it is not (and so not overall). This is because the research strategy under analysis does not sufficiently discount the causal theory under investigation when disconfirmatory evidence comes to light. This simple exercise illustrates how features of some qualitative designs (such as the reliability of inference given the updating strategy) can be made assessed using formal declaration.

4. Using Diagnosis to Develop Designs

In designing social science research, researchers face myriad tradeoffs whose features are not obvious. The code-based declaration and diagnosis of designs described in this paper enable authors and readers to learn about the consequences of these different design choices. In what follows, we provide three examples that demonstrate how formal declaration and diagnosis of designs can help assess difficult design choices with respect to sampling, treatment assignment, and analysis.[15] We emphasize that these tradeoffs can also be examined by readers and replication authors independently of one another.

4.1 Illustration of Analysis Decisions: Gains from Covariate Control

The Challenge. Researchers often have to decide whether to use controls when estimating treatment effects from experimental or observational data. Some prefer estimates without controls, some prefer employing controls that they expect to be associated with outcome variables, and some prefer controls only if they find a correlation between the control variable and the explanatory or treatment variable. Among those that use controls, some prefer simple controls and some prefer to interact controls with explanatory variables or treatments.

A lot can ride on what choices are made. We consider a potential outcomes function that allows outcomes to be a function of a covariate X as follows:

$$Y_i(Z_i) = Z_i + \gamma X_i + \nu \epsilon_i$$

[15] See the appendix for instructions on installing `DeclareDesign` and downloading the research design templates illustrated in this section.

where X_i and ϵ_i are distributed with 0 mean and unit variance. To ease interpretation, assume that $\text{Var}(Y(0)) = \text{Var}(Y(1)) = 1$ and that ϵ is orthogonal to X (note, this in turn implies that $\nu = \sqrt{1-\gamma^2}$).

The puzzle for researchers is now: *Should you commit to using controls ex ante? How does the decision depend on γ?* We approach this question by declaring and diagnosing a set of designs.

Design Declaration. To be able to quickly generate many similar designs we create a "design template" that can be used to declares designs as a function of arguments we may wish to vary. In this case the template `make_heterogeneous_fx_design(n, g)`, has one argument for the number of units, (`n`), and another, `g`, for γ. For realism, this template included a wrinkle in which a researcher has access only to some proxy for X, $\tilde{X}$, which implies that models that include $\tilde{X}$ are misspecified. The design estimates three models: Model Z_1 uses regression to take simple difference in means, Z_2 includes a linear control for $\tilde{X}$ and model Z_3 includes both a control for $\tilde{X}$ and an interaction in $\tilde{X}$.

We can then create multiple designs using this template and compare the designs:

```
make_heterogeneous_fx_design <- get_template("covariate_control")
heterogeneous_designs <- quick_design(
  template = make_heterogeneous_fx_design,
  n = 20,
  g = vary(0,.3,.6,.9))

compare_designs(design = heterogeneous_designs)
```

This code will create a matrix with the diagnosis of a design with 20 units, potential outcomes generated under the assumption that γ ranges between 0 and 0.9, and analyses implemented using all three estimators.

Results. Figure 2 shows the behavior of the three estimators for different values of γ. In addition, it shows the behavior conditional on "statistically significant" imbalance on $\tilde{X}$ (dotted lines).

The results show that there can be large gains in power, without introducing bias, when there is a true underlying relationship between the control and the potential outcomes. There are costs however when the true relationship is weak, especially in those cases in which there is imbalance between the covariate and explanatory variable. This highlights the risks of strategies that select covariates based on balance tests without consideration for the relationship between the covariate

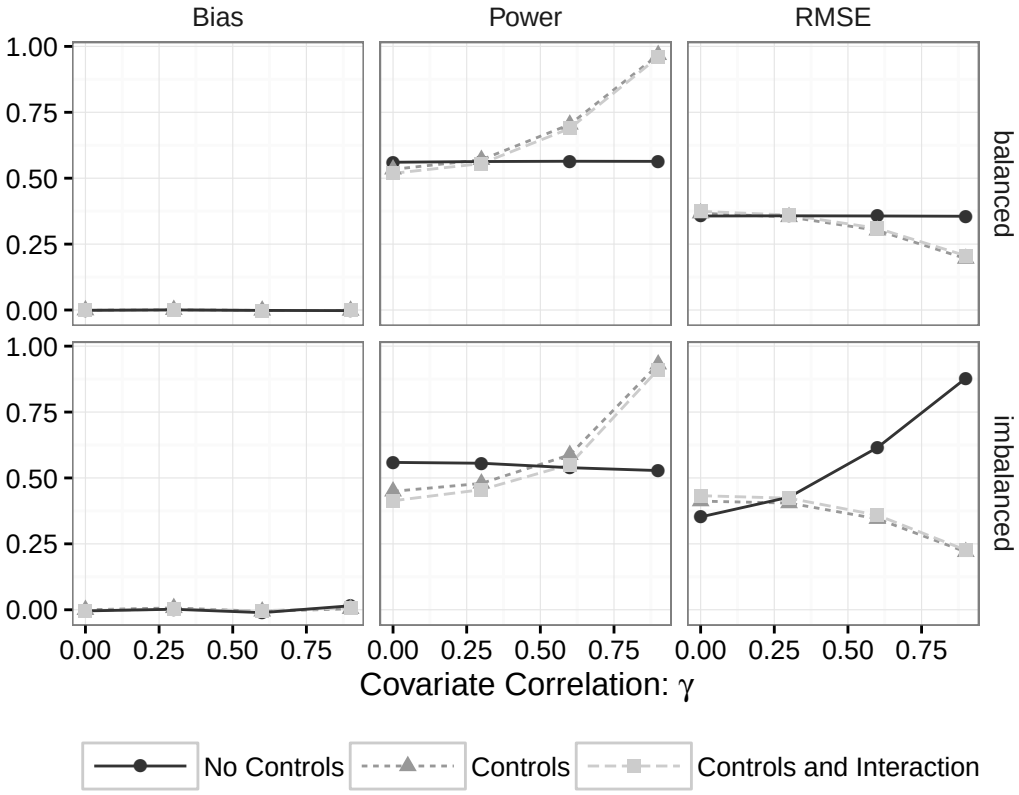

Figure 2: Diagnoses of Designs with Varying Estimators Illustrate the Gains from Covariates. Power and bias from models without controls (black lines; circles), with linear controls for $\tilde{X}$ (dark gray dotted lines; triangles), and with controls interacted with treatment (light gray dashed lines; squares).

and potential outcomes. Finally note that when there is large imbalance on a prognostic covariate, the simple difference in means estimator remains unbiased, but it can have a large mean squared error, reflecting *conditional* bias.

4.2 Illustration of Sampling Decisions: Handling Spatial Spillovers

The Challenge. Researchers often tailor their sampling method in order to best recover inferential targets, such as causal effects. In some settings researchers worry that the effect of an intervention may "spill over" from units that received the intervention to those that did not. To guard against such risks, they might impose a constraint on their sampling strategy, whereby no two units will be sampled if they are deemed too close together. However, such strategies can result in treatment and control groups that are far apart and frustrate efforts to compare like with like.

So their question becomes: *what size 'buffer' to select given risks of bias on the one hand and risks*

of covariate imbalance on the other?

We answer the question by declaring and diagnosing a design that attempts to address spillover risks by introducing buffers.

Design Declaration. Consider an experiment that assigns a treatment to neighborhoods within cities (say, a leafleting strategy), and measures its effects on some outcome (say, voter turnout). The researchers assign one neighborhood within each city to treatment, and one to control. However, they are concerned that the treatment may spillover onto adjacent neighborhoods, so they consider using a 'buffer' to insulate treatment units from control units.

Figure 3 visualizes possible buffered sampling schemes. In the figure each circle represents a neighborhood. In the left panel, two neighborhoods are selected without any regard for their proximity (no spatial buffer). In the middle panel, two neighborhoods are selected such that they are a distance of at least 3 units apart, and in the right panel only neighborhoods that are at least 4 units apart are selected.

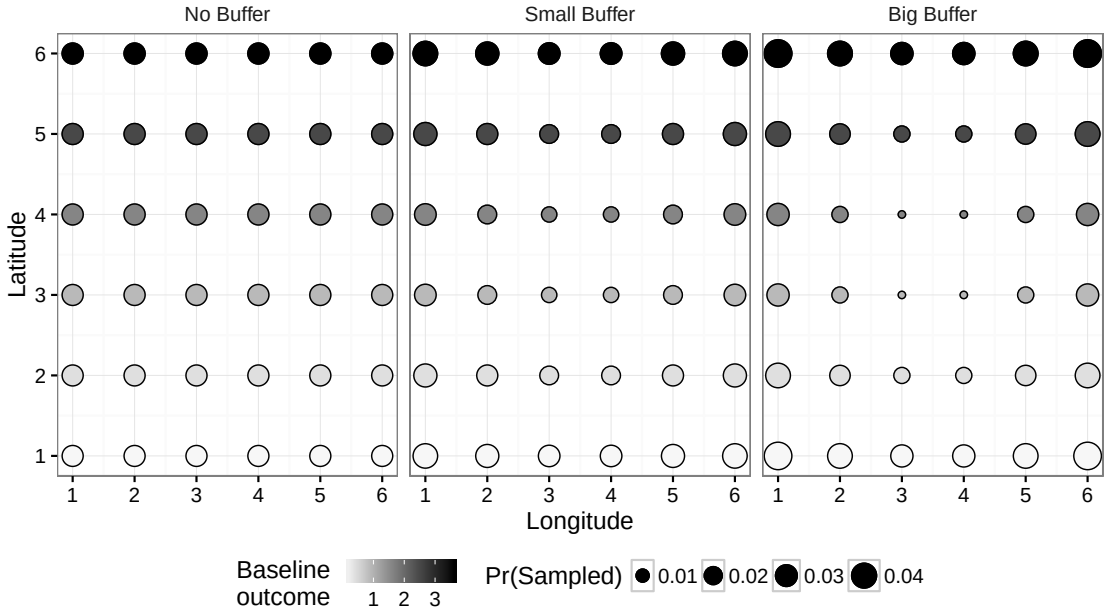

Figure 3: Illustration of the Implications for Sampling Probabilities of Spatial Buffer Sampling Strategies when Baseline Outcomes are Related to Geography. The baseline outcome is highly correlated with latitude (y axis). Three sampling strategies are illustrated: no spatial buffer (left); a strategy in which units are selected only if they do not fall within a small distance (buffer) of another selected unit (middle); and a strategy with a high distance buffer. Units in the center, which are most likely to be close to another unit, are selected with low probability in the "big buffer" case. These units have mid-level baseline outcomes.

Say the researchers also know that units in the north have higher baseline outcomes (e.g. higher turnout rates) than those in the south. This knowledge is represented in Figure 3 with darker shading of areas farther north.

To understand the properties of different designs, researchers can declare designs that include this knowledge in the declaration of potential outcomes and vary sampling strategies. Again we make use of a design template, this time `make_spillover_design`, that takes arguments for buffer size, `buffer`, and the size of the distance effect, `dist_effect`. With a template in hand, multiple designs can be quickly generated and compared:

```
make_spillover_design <- get_template("spatial_spillovers")
spillover_designs <- quick_design(
  template = make_spillover_design,
  buffer = vary(0, 3, 4),
  dist_effect = vary(0,1,3),
  intersect = TRUE)

compare_designs(design = spillover_designs)
```

We can then assess the inferential properties of these different sampling strategies under different assumptions about the strength of spillovers through Monte Carlo diagnosis.

Results. The results are displayed in Figure 4. They illustrate that, whereas large buffers *do* help to avoid bias, they can also increase the expected error of the estimator. Effectively, efficiency is lost because sampling more distant units induces imbalance. In this example, if the researcher expected only weak spillovers at most, she might be better to avoid a spatial buffer altogether: this minimizes the RMSE at the expense of only a small amount of bias. However, this tradeoff is different if there is reason to expect strong spillovers: the reduction in bias from including a spatial buffer is proportionally greater than the increase in RMSE, especially in the shift from no buffer to a small one.

4.3 Illustration of Assignment Decisions: Assigning Multiple Treatments

The Challenge. In experimental work, researchers are frequently faced with a choice between running a 2-by-2 factorial design or a three-arm trial. Different considerations come in to play depending on the estimands of interest and the type of interactive effects researchers might expect. Consider a situation in which a researcher is considering two treatments and is interested

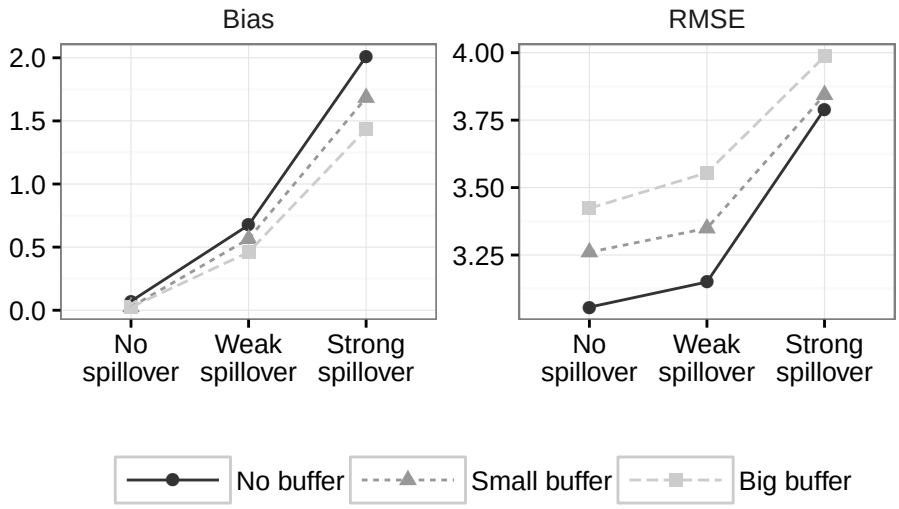

Figure 4: Diagnoses of Designs with Varying Spatial Buffers for Sampling Units Illustrate a Bias-Variance Tradeoff. The bias (left) and root mean-squared-error (right) are displayed for designs with no buffer for sampling units (solid lines; circles), a small distance buffer (dark gray dotted lines; triangles), and a big distance buffer (light gray dashed lines; squares) as a function of the level of spillover effects that are declared in the potential outcomes function (x axis).

in the effect of each treatment *conditional on the other treatment being in the control condition*.

In assessing the two strategies they are conscious that the three arm trial lets them use only two thirds of the data for each comparison whereas the factorial design lets them use 100% of the data for each comparison. This is the point made by Fisher (1992). However to use 100% of the data they need to use data for cases that are in treatment condition for treatment 2 when assessing the effects of treatment 1 and vice versa. Such overlap can introduce risks of bias for assessing the estimands of interest.[16]

The question here is: *Should researchers use a factorial design or a three arm design when their estimand for each treatment is conditional on the absence of the other treatment?*

Again we answer the question by declaring and diagnosing a set of possible designs.

Design Declaration. We assume an experimental population of 500 subjects. Potential outcomes are a function of exposure to treatment 1 (Z1), treatment 2 (Z2), and their interaction.

[16]One might object that the factorial design is not biased—its utility lies in its ability to estimate the effects of treatment 1 in the absence *and* presence of treatment 2. Some of the remarks in Fisher (1992) go in this direction. This response however runs the risk of identifying the estimand as whatever it is that the estimator shoots at. In practice however researchers have estimands in mind, such as average treatment effects, and may select factorial design because of properties such as efficiency, or allowing the possibility to estimate interaction effects, without intending to alter estimands on main effects.

We consider two assignment procedures, one in which subjects are assigned to each cell of a 2×2 with probability $1/4$ and one in which subjects are assigned to a control condition, treatment 1, or treatment 2, each with probability $1/3$. The estimands are as described above. The two estimators are the coefficients on the treatments from from an OLS regression of the outcome on indicators for each treatment.

Given the symmetry of the problem we focus on one treatment effect only. We use two different assignment strategies to generate designs across a range of interaction effects, under the two different assignment procedures:

```
make_factorial_design <- get_template("factorial")
factorial_designs <- quick_design(
  template = make_factorial_design,
  assignment_strategy = vary("two_by_two","three_arm"),
  interaction_coefficient = vary(0.00,0.05,0.10,0.15,0.2,
                                 0.25,0.30,0.35,0.4))

compare_designs(design = factorial_designs)
```

Results. The results of this diagnosis are presented in Figure 5. In the left-most panel, the bias of each design is plotted on the vertical axis, while the size of the interaction is plotted on the horizontal axis. When the true interaction term is equal to zero (i.e., the effect of treatment 1 does not vary with the level of treatment 2), neither design exhibits bias. However, as the interaction between the two treatments is stronger, the factorial design renders estimates of the effect of treatment 1 that are more and more biased relative to the "pure" main effect estimand.

In the center panel, the root-mean-squared error of each design is plotted on the vertical axis. This panel shows that there is a bias-variance tradeoff in this design. When the interaction term is small or close to zero, the factorial design is preferred, because it more powerful: it compares one half of the subject pool to the other half, whereas the three arm design only compares a third to a third. However, as the magnitude of the interaction term increases, the precision gains are offset by the increase in bias documented in the left-panel. In cases of high heterogeneity, the three-arm design is then preferred.

This exercise highlights key points of design guidance. Researchers often select factorial designs because they expect interaction effects: and indeed factorial designs are required to assess

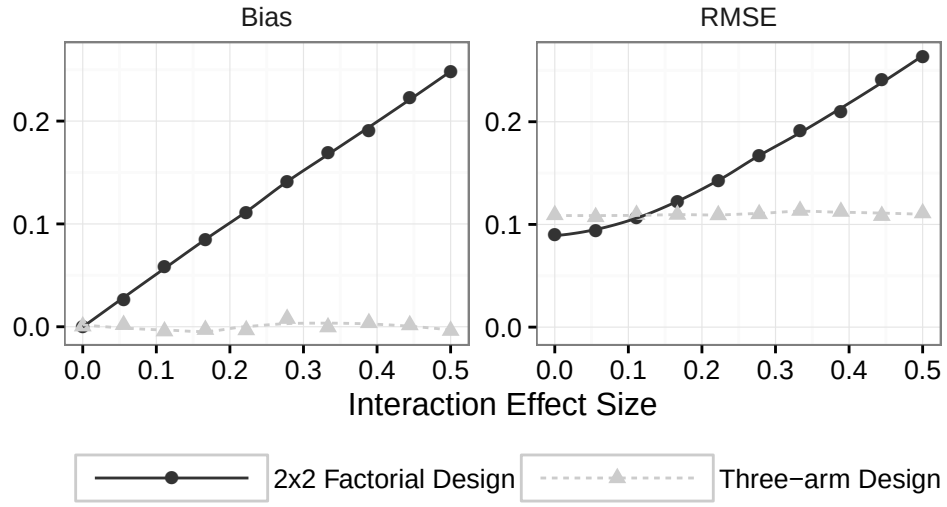

Figure 5: Diagnoses of Designs with Factorial or Three-Arm Assignment Strategies Illustrate a Bias-Variance Tradeoff. Bias (left) and root mean-squared-error (right) are displayed for two assignment strategies, a 2×2 treatment arm factorial design (solid lines; circles) and a three-arm design (dark gray dotted lines; triangles) according to varying values of the interaction effect specified in the potential outcomes function (x axis).

these. However if the scientific question of interest is the pure effect of each treatment, researchers should (perhaps counterintuitively) use a factorial design if they expect *weak* interaction effects.

5. Discussion

We have described a strategy for declaring research designs for which key "diagnosands" can be estimated, given conjectures about the world. How might declaring and diagnosing research designs in this way affect the practices of authors, readers, and replication authors? We believe there are implications for how design choices are made, communicated, and challenged.

5.1 Making Design Choices

The move towards greater transparency places a premium on considering alternative analysis strategies at early stages of research projects, not only because it reduces researcher discretion, but also because it can improve the quality of the final research design. There is nothing new about the idea of determining features such as sampling and estimation strategies ex ante in order to maximize power. In practice, however, many designs are finalized late in the research process, often after the data are collected. This may occur in part due to a lack of tools for

adequately assessing the relevant properties of a research design and for exploring possible analysis strategies before data are collected.

One of the most common tools to evaluate the properties of research designs is the power calculator, which assesses the probability of finding a "significant" result given various assumptions about the data generating process and analysis strategy. Existing power calculators are surprisingly rudimentary: they handle a very small set of special cases, and often do not show how power varies as a result of the many design choices a researcher must make besides sample size. There are no general tools for assessing power or other equally important properties, such as unbiasedness, mean squared error, or coverage.

The lack of tools for design diagnosis is not only a problem for those conducting the research. Readers of empirical studies and the authors of replication studies also need to assess the general properties of research strategies, yet often lack the tools and information needed to do so.

The proposed procedure—declaring and diagnosing research designs—makes it possible and relatively simple for researchers and reviewers to assess a range of statistical properties of research designs and to compare them to alternatives.

We emphasize an obvious caveat to simulation-based defenses of design choices. A simulation based claim to unbiasedness is good only with respect to the conditions of the simulation; for example conditional on the potential outcomes functions posited. In this sense claims for properties of strategies are more robustly made based on analytic results. Often however, the complexity of a given research design prohibits analytic interrogation of diagnosands. Conversely, a simulation based *critique* of a strategy—such a demonstration that a strategy is biased for some estimand—may be powerful even when general analytic results do not exist.

5.2 Communicating Design Choices

Bias in published results can arise for many reasons. For example, researchers may deliberately or inadvertently select analysis strategies because they produce statistically significant results. Proposed solutions to reduce this kind of bias focus on various types of preregistration of analysis strategies by researchers (Rennie, 2004; Zarin and Tse, 2008; Casey, Glennerster and Miguel, 2012; Nosek, 2014; Green and Lin, 2016). Study registries are now operating in numerous areas of social science, including those hosted by the American Economic Association, Evidence in Governance

and Politics, and the Center for Open Science.

However, the effectiveness of design registries in reducing the scope for fishing depends on clarity over which elements must be included in a precommitment document. In practice some registries rely on various checklists and pre-analysis plans exhibit great variation, ranging from lists of written hypotheses to all-but-results journal articles. In our view, the solution to this problem does not lie in ever-more-specific questionnaires, but rather in a new way of characterizing designs whose analytic features can be diagnosed through simulation.

The criteria we propose clarify what is required in order to preregister a study in a way that ensures sufficient analytically relevant detail is provided. Our framework provides a set of generalizable procedures that promise to standardize the way in which designs are preregistered. These procedures will make it easier for researchers and third parties to understand when a plan is consistent with the standards required for effective preregistration. Rather than asking: "are the boxes checked?" the question becomes: "can it be diagnosed?" A design can only be diagnosed when sufficient detail has been provided to analytically characterize diagnosands or to conduct Monte Carlo simulations of the implementation of the design from beginning to end.

Declaration of a θ-complete design also enables a final and infrequently practiced step of the registration process, in which the researcher "reports and reconciles" the final with the planned analysis. Declaring the features of a design ex ante and ex post makes possible the identification of deviations from an analysis plan. Understanding how the two diverge is a central part of assessing whether the results should be viewed as exploratory or confirmatory.

5.3 Challenging Design Choices

The independent replication of the results of studies after their publication is an essential component of the shift toward more credible science. Be it verification and reanalysis of the original data, or reproduction of results through the collection of fresh data, replication provides incentives for researchers to be clear and transparent in their analysis strategies, and can build confidence in the robustness of findings.[17]

A declaration of a design that can be simulated, implemented, and diagnosed facilitates replication by rendering the design itself transparent. Indeed replication (sometimes called pure

[17] For a discussion of the distinctions between these different modes of replication, see Clemens (2015).

replication, analytic replication, or verification) using the same data becomes technically trivial. For this exercise the question becomes not whether the code produces what the authors claim, but whether the code is correct.

More subtly, a complete declaration can also inform the re-analysis and critique of published research and allow for a different approach to reanalysis. A standard practice for replicators engaging in reanalysis is to propose a range of alternative strategies and assess the robustness of the *data*-dependent estimates to different analyses. A more coherent strategy, if it were possible, would be to assess the robustness of the analysis strategy to different ways in which the data may have been generated. The problem with the standard approach to reanalysis is that when divergent results are found, third parties do not have clear grounds to decide which results to believe. This issue is compounded by the fact that, in changing the analysis strategy, replicators risk departing from the estimand of the original study, possibly providing different answers to different questions. In the worst case scenario, it can be difficult to determine what is learned both from the original study and from the replication.

Design declaration enables a new type of replication: the "design replication." In a design replication, a scholar restates the essential design characteristics to learn about what the study *could have* revealed, not just what the original author reports *was* revealed. This helps to answer the question: under what conditions are the results of a study to be believed? By providing a structure to compare the abstract properties of alternative analyses, design declaration provides grounds to support alternative analyses on the basis of the original authors' intentions and not on the basis of the degree of divergence of results. Conversely, it provides authors with grounds to question claims made by their critics. We provide an example of a design replication of a study for which data is currently not available in Blair et al. (2016*b*). In that replication we illustrate how the strategy employed by Björkman and Svensson (2009) could under some reasonable data generating processes give rise to biased results. We emphasize that this exercise does not *demonstrate* bias rather it simply helps locate possible sources of bias.

Table 1 shows possible situations that may arise. In a declared design an author might specify situation A: a particular set of claims on the structure of potential outcomes and an estimation strategy. A critic might then question the claims on potential outcomes (for example questioning SUTVA) and/or question estimation strategies (for example arguing for the need to include or

	Author's assumed potential outcomes	Alternative claims on potential outcomes
Author's proposed estimation strategy	A	B
Alternative estimation strategy	C	D

Table 1: Diagnosis Results Given Alternative Assumptions on Potential Outcomes and Alternative Estimation Strategies. Four scenarios encountered by researchers and reviewers of a study are considered depending on whether the potential outcomes function or the estimation strategy differs from the author's original strategy.

exclude some control variables from an analysis), or both.

In this context here are several possible criteria for admitting alternative estimation strategies:

- **Home ground dominance.** If ex ante the diagnostics for situation B are better than for A then this gives grounds to switch to B. That is, a critic can demonstrate that an alternative estimation strategy outperforms an original estimation strategy even under the data generating process assumed by an original researcher, then they have strong grounds to propose a change in strategies. Conversely if an alternative estimation strategy produces different results, conditional on the data, but does not outperform the original strategy given the original assumptions, this gives grounds to question the reanalysis.

- **Robustness to alternative potential outcome functions.** If the diagnostics in situation B are as good as in A but are better in situation D than in situation C this provides a robustness argument for altering estimation strategies.

- **Potential outcomes plausibility.** If the diagnostics in situation A are better than in situation B, but the diagnostics in situation D are better than in situation C, then this is cause for worry and the justification of a change in estimators depends on the plausibility of the different assumptions on potential outcomes.

As an illustration of the application of these principles, consider a situation in which a researcher produces an estimate of an average treatment effect. A critic notes that the treatment is highly correlated with a covariate, not included in the original analysis, and that significance is lost once the control is included. The researcher might then counter that although results are sensitive to the inclusion of the control, the new strategy does not satisfy home ground dominance—that is, given prior assumptions about potential outcomes, the diagnostics from the

new estimation strategy are not better than those from the original strategy. The critic could then describe an alternative potential outcomes function and demonstrate either that the new strategy is more robust to alternative potential outcomes functions or that it is preferable on the basis of potential outcome plausibility—for example by using the data to demonstrate that the covariate is prognostic of potential outcomes contrary to researcher assumptions. In all cases, transparent arguments can be made by formally comparing the original design to a modified design.

While such criteria will not eliminate disputes they should at least help focus the discussion on the analytically relevant issues.

5.4 Limitations and Risks of the Proposed Approach

We describe a procedure for characterizing and diagnosing designs before implementation. Ex ante declaration and diagnosis of designs can help researchers improve their properties. It can make it easier for readers to evaluate a research strategy prior to implementation and without access to results. It can also make it easier for designs to be shared and to be critiqued. Our proposed framework and software aims to facilitate these steps. However, the creation of a set of tools to evaluate the completeness and quality of research designs also creates a set of risks. We outline four.

The first risk is that evaluative weight gets placed on essentially meaningless diagnoses. Given that design declaration includes declarations of conjectures about the world it is possible to choose numbers so that a design passes any diagnostic test set for it. Fortunately, however, the advantage of the formal declaration is that the basis for the diagnoses can be examined and new diagnostics can be generated quickly given alternative specifications of data generating processes while keeping other design elements intact. Even still, the risk remains that if the grounds for diagnoses are not inspected, designs may be favored because of the optimism of the designers rather than because of the inherent qualities of the design.

A second risk is that research gets evaluated on the basis of a narrow but perhaps inappropriate set of diagnosands, such as power, bias, or RMSE. In fact, the appropriateness of the diagnosand depends on the purposes of the study. The optimal bias-variance tradeoff for example might depend on whether the interest is in assessing properties of a specific case or whether a study is contributing to a larger literature. To help guard against this risk we provide a range

of diagnosands as defaults in our software and allow users to define their own diagnosands. In this way, the evaluative grounds for research may be widened for example by making it easier for researchers to demonstrate the value of a research design that carries risk of bias but has other valuable properties.

A third risk is that as the evaluation of formal properties of a design become easier, evaluative weight shifts away from the substantive importance of a question being answered. A similar concern has been raised regarding the "identification revolution" where a focus on identification risks crowding out attention to the importance of questions being addressed (Huber, 2013). Similarly there could be a risk that less attention is paid to measurement issues, which largely fall outside our framework. It is also possible however that simplification of the evaluation of formal properties of a design allow for a shift in attention towards examining other properties of a design such as measurement strategy or substantive and theoretical relevance. More creatively, it may also be possible to think of substantive importance as a diagnosand—for example one could declare as a diagnosand the likelihood that the research will contribute new knowledge to a given question (whether or not it is excellent statistical properties).

A fourth risk is that the variation in the suitability of design declaration to different research strategies that we outlined above is taken as evidence of the relative superiority of different types of research strategies. We believe that the range of strategies that can be declared and diagnosed is wider than what one might at first think possible, and we sketch above outlines for declarations of descriptive, experimental, observational, quasi-experimental, and qualitative strategies. We argue that there is value in formally declaring designs when this is possible. There is no reason to believe, however, that all strong designs can be declared either ex ante or ex post. An advantage of our framework, we hope, is that it can help clarify when a strategy can or cannot be completely declared. In cases in which a strategy can not be declared, nondeclarability is all that the framework provides, and in such cases we urge caution in drawing broader inferences about design quality.

References

Bennett, Andrew and Jeffrey T. Checkel, eds. 2014. *Process tracing*. Cambridge: Cambridge University Press.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How much should we trust difference-in-difference estimates?" *The Quarterly Journal of Economics* 119(1):249 – 275.

Björkman, Martina and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda." *Quarterly Journal of Economics* 124(2):735–769.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2016*a*. "DeclareDesign Version 1.0." Software package for R, available at http://declaredesign.org.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2016*b*. "Using DeclareDesign to Replicate Bjorkman and Svensson (2009).".
URL: *http://declaredesign.org/replications/bjorkman-svensson.html*

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3(Jan):993–1022.

Bourguignon, Francois. 1979. "Decomposable income inequality measures." *Econometrica: Journal of the Econometric Society* 47(4):901–920.

Casey, Katherine, Rachel Glennerster and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." *The Quarterly Journal of Economics* 127(4):1755–1812.

Clemens, Michael A. 2015. "The Meaning of Failed Replications: A Review and Proposal." *Center for Global Development Working Paper* 399.

Dawid, A. Philip. 2010. "Beware of the DAG!" *Journal of Machine Learning Research Workshop and Conference Proceedings* 6:59–86.

Fafchamps, Marcel and Julien Labonne. 2016. Using Split Samples to Improve Inference about Causal Effects. Technical report National Bureau of Economic Research Working Paper No. 21842.

Fisher, Ronald A. 1992. The arrangement of field experiments. In *Breakthroughs in Statistics*. Springer pp. 82–91.

Gelman, Andrew and John Carlin. 2014. "Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9(6):641–651.

Green, Donald P. and Winston Lin. 2016. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *PS: Political Science and Politics* 49(3):495–499.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–960.

Huber, John. 2013. "Is theory getting lost in the "identification revolution"?" *Monkey Cage* blog post.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods and Research* 41(4):570–597.

Nosek, Brian A. et al. 2014. "Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices." Transparency and Openness Committee Report.

Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.

Rennie, Drummond. 2004. "Trial registration." *JAMA: the Journal of the American Medical Association* 292(11):1359–1362.

Rubin, Donald B. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *The Annals of Statistics* 12(4):1151–1172.

Van Evera, Stephen. 1997. *Guide to methods for students of political science*. Ithaca: Cornell University Press.

Zarin, Deborah A. and Tony Tse. 2008. "Moving towards transparency of clinical trials." *Science* 319(5868):1340–1342.

Appendix

Contents

A. Types of Estimators and Estimands

B. Illustrations of Design Formalization

C. Declaration and Diagnosis of a Bayesian Estimation Strategy

D. Reproducing the Examples in Section 3

E. Reproducing the Examples in Section 4

A. Types of Estimators and Estimands

This section expands on the discussion in Section 1 on estimators and estimands.

A.1 Estimands

This formalization helps us distinguish between analytically distinct classes of estimand.

- A "**post-treatment**" estimand is an estimand that is conditional on the particular assignment to treatment (in a particular sample). For example assuming a binary treatment, and let $Y^{(\mathbb{1}(i))}$ denote the assignment vector Y in which only unit i is assigned to treatment and let $Y^{(\varnothing)}$ denote the assignment in which no units are assigned to treatment. Let Y^1 denote a treatment and Y^2 an outcome of interest. Conditional on a realization of X, the average effect on the treated is defined as

$$\tau_{ATT}(d, \mathcal{D}) = \mathbb{E}_{i:y_i^1=1}(Y_i^2(Z^{1(\mathbb{1}(i))}|x) - Y_i^2(Z^{1(\varnothing)})|x)$$

  Note that here the definition of the population of interest for the estimand is given by Y^1 not Z^1—that is by the set that are in treatment, whether or not they are included in the experimental population. The estimand however is based on the potential outcomes that would be observed had all units been experimentally assigned to different conditions, whether or not they were (Z^1).

  Under our formulation sampling is itself a type of treatment. Thus the "sample average treatment effect," SATE, is itself a post-treatment estimand, that depends on the outcomes of a sampling decision, and the sample average treatment effect on the treated, SATT, depends on both sample selection and treatment assignment. These are all post-treatment estimands in our framework though some may be defined conditional on the assignment of some treatments and with respect to possible assignments of other treatments. To illustrate the *expected* SATT depends on a particular outcome from sampling but not on a particular treatment assignment.[18]

- A "**population estimand**" depends upon a particular population but not a particular as-

[18]Though one could also define an expected SATT that is conditional on a treatment assignment in a population but not on a particular sample.

signment. For example, we could also define a notion of an *expected* SATT in a population that is does not condition on a sample or an assignment, as $\tau_{ESATT} = \int\int \tau_{SATT} dp_{Z^1} dp_{Z^2}$. Where Z_1 is a sampling decision and Z_2 is a treatment decision. This particular population estimand depends on the sampling scheme and the assignment scheme, though other population estimands, such as the population average treatment effect, might not.

Given uncertainty on X we can also define expected population estimands: $\tau_{EPATT} = \int \tau_{EPATT} df_X$. These estimands are themselves of different types:

- A "**design independent estimand**" is an estimand that does not depend on the design, p_Z. For example the PATE, can be defined as $E_N(Y_i^2(Z^{1(\mathbb{1}(i))}) - Y_i^2(Z^{1(\varnothing)}))$ where E_N is the mean over units in a population. The PATE, thus defined, is a design-independent population estimand.
- A "**universal estimand**" for group N is an estimand that is defined for all subgroups of N.
- A "**decomposable estimand**" is an estimand that can be written in the form: $\tau_N = \sum_{i \in N} \frac{1}{N} \tau_i$.[19]

Note that every decomposable estimand is universal. The estimand $\tau_N = \mathbb{E}_{i \in N}(Y_i(1) - Y_i(0))$ is universal and decomposable. The estimand $\tau_N = \max_{i \in N}(Y_i(1) - Y_i(0))$ is universal but not decomposable. The estimand $\tau_N = \frac{1}{2} \frac{1}{\#\{i:X_i=0\}} \sum_{\{i:X_i=0\}}(Y_i(1) - Y_i(0)) + \frac{1}{2} \frac{1}{\#\{i:X_i=1\}} \sum_{\{i:X_i=0\}}(Y_i(1) - Y_i(0))$ is neither decomposable nor universal.

A.2 Estimators

A **design independent estimator** is one that depends on the realized data only. For example the difference in means estimator depends on the data only.

$$\hat{\tau}_{DIM}(D) = E_{i:Y_i^1=1} Y_i^2 - E_{i:Y_i^1=0} Y_i^2$$

For an example of an estimator that is not design independent, consider an inverse probability weighted estimator for a two arm design. Let $\pi_i^j = \int \mathbb{I}(Z_i^1 = j) dp_{Z^1}$ denote the probability of individual i to be in treatment condition j. Let S_i indicate whether $\pi_i^j \in (0,1)$ for all j. Then let

[19] For a discussion of aggregative and decomposible indices, see for example Bourguignon (1979).

$w_i^j = \sum_{k:Y_k^1=j, S_k=1} \pi_k^j \frac{1}{\pi_i^j}$ denote the inverse probability weight. Then the inverse probability weight estimator can be written:

$$\hat{\tau}_{IPW}(D, p_Z) = \sum_{i:Y_i^1=1, S_k=1} w_i^1 Y_i^2 - \sum_{i:Y_i^1=0, S_k=1} w_i^0 Y_i^2$$

Note that the estimator only uses data from units for which there is a positive probability of being in every condition, though that does not necessarily mean that the associated estimand is for that subgroup. Here π_i^j depends on the design through p_{Z^i} and w_i additionally depends on the realization of Z^1. Thus the "data" required to define the IPW estimator includes all the weights that *would* result from all assignments to treatment.

Standard errors and confidence intervals can also generated by estimators. The confidence interval estimated by inverting randomization inference tests uses a design-dependent estimator as is the confidence interval estimated using Neyman standard errors. The confidence interval generated using the HC2 robust estimator for standard errors, though equivalent in some cases to a Neyman estimator is data-based.

B. Illustrations of Design Formalization

We illustrate the formalization in two parts, focusing first on the causal structure, and second on estimands and diagnosands.

B.1 Illustration: Study structure

Figure 1 above illustrated the structure of studies implied by this formalization and the effectiveness and modularity assumptions. As noted, this example encompasses both experimental and non-experimental studies depending on whether `idle` is set for all subjects. A number of features of Figure 1 are worth highlighting.

- For each outcome variable Y^j there is a manipuland Z^j that points to Y^j but only to Y^j in the Y class of variables. Z variables may point to each other however, representing the possibility that a researcher may condition one assignment on the realization of another assignment. Thus modularity is represented here as exclusion restrictions in the DAG.

- Perhaps confusing in terms of notation, treatment is labeled as Y^1 (though in other discus-

sions it is labeled as X, D, Z, or W to distinguish it from outcomes, which take Y labels). This is done to allow for a distinction, available for all the Y variables, between values taken due to direct manipulation and values taken without manipulation (i.e. `idle`). In observational reserach designs, all units are set to `idle` for treatments.

- We include $\hat{\tau}$ in the graph to highlight the fact that an estimate is itself a potential outcome in which case it could in principle be counted among the set of random variables Y.

- $\hat{\tau}$ may be affected by X^o directly (for example using an estimator with covariates) but is affected by X^u only through the measured outcome, Y^6.

- Measurement is itself an outcome with an associated manipuland. This allows one to include in superdata what would be measured if an unmeasured variable were measured.

- The estimand τ is not a node on the DAG.

B.2 Illustration: Estimands and Diagnosands

To illustrate the calculation of estimands and diagnosands, we consider a population with $N = 3$ units. A manipuland Z^1 determines whether each unit is put in a treatment condition $Y_i^1 = 1$ or a control condition $Y_i^1 = 0$ or is left untouched by the researcher (in which case they could self select into either of these two conditions, possibly as a function of the assignment of other units).

Treatment status determines an ultimate outcome of interest Y^2. The key quantities of interest include:

- The estimand of interest is the average across all units i, of the difference between the condition in which unit i only is assigned to treatment (and all other units are assigned to control) and the condition in which all units are assigned to control. Call this τ.

- A summary statistic of interest, $\hat{\tau}$ is the difference between the average outcome in treated ($y^1 = 1$) and untreated units ($y^1 = 0$). Here this is an estimate of τ using data on items whose value on Z^1 was not idle, though one could also estimate τ using all data.

- A diagnostic statistic of interest is the difference between $\hat{\tau}$ and τ.

- A diagnosand of interest is the expected value of this diagnostic statistic, which is the bias of the difference in means estimator for the average causal effect.

Consider now an assignment scheme, p_Z, in which exactly one of the three units is randomly assigned to treatment, one is assigned to control, and one is not assigned by the researcher at all. There are six possible assignment vectors of this form. Assume that p_Z assigns a 1 in 6 probability to each of these assignment vectors. A single draw could give rise to assignment $Z^1 = \{1, 0, \oslash\}$, for example.

Table 2 illustrates ten possible datasets d that could arise given different realizations of Z^1, the first six of which are given positive probability under p_Z; the last four are *potential* realizations, even though they are never assigned under the randomization scheme. These rows, together with the assignments not included, form the superdata, $\mathcal{D}$.

The second and third columns of Table 2 show that Y^1 takes the values given by Z^1 when Z^1 is not idle. However, when Z_1 is idle the value taken by Y^1 can depend on *which* units are assigned. In this illustration, unit 3 is idle in both assignments 1 and 5 but takes on different values of Y^1. Similarly, potential outcomes, Y^2, are a function of the entire vector Y^1. We see in this example that there is no imposition of SUTVA; in particular unit 1 has a different value of Y^2 under assignments 1 and 2.

Returning to the quantities of interest, note:

- The estimand can be calculated by comparing outcomes across different assignment vectors. It is found by taking the average of y_1^2 in assignment 7, less y_1^2 in assignment 10, y_2^2 in assignment 8, less y_2^2 in assignment 10, and so on. In this example the estimand is 4, indeed the treatment effect for each unit (when all others are in control) is 4.

- Estimates can be calculated from Y^1 and Y^2 directly and can be calculated even for assignments that receive zero probability under p_Z. Note however that estimates cannot be calculated for the uniformity trial (assignment 10).

- The diagnostic statistic is calculated separately for each realization, but depends in part on the unobserved realizations 7 - 10. The diagnostic statistic in the final column is not a potential outcome since it depends on τ.

- The expected value of the diagnostic statistic is the diagnosand. This is taken with respect to the probability distribution p_Z. Here it is the expected difference between $\hat{\theta}$ and θ and takes the value $-8/6$.

The diagnosand tells us that the estimator is biased. The bias comes from the fact that unit 2 exhibits spillover effects when it is assigned to control and both other units are treated. A design that controlled unit 1 and unit 3 only and ignored unit 2 would not suffer from bias.

The introduction of possible idle conditions lets us be more precise about relatively subtle features of the definition of estimands. To illustrate, consider a researcher that introduced assignment 1 and used data on units 1 and 2 only. Say the researcher claimed to be interested in the sample average treatment effect; they estimate an effect of 0. But what is the estimand? In our calculations we used assignment 10 as the base condition for defining the estimand, but the researcher could instead respond that assignment 9 is more appropriate; it is after all the assignment in which the controlled units are both in the control condition and which the non controlled unit is in the treatment condition. With that base condition the bias is not so great. One might worry however that the uncontrolled unit is only in the treatment condition because of the particular assignment of controlled units to treatment. With a different assignment (assignment 5), the non controlled unit behaves differently. The key point is that the outcomes in study units can depend on which units enter a study and can in turn affect the definition of the estimand and the definition of a sample estimand can make use of population potential outcomes.

We also see effects of self selection that are different from the usual concern. If data on all units were used, further bias could be introduced even though all units in this example have the same treatment effects and indeed have the same outcomes in the uniformity trial and in the assignments in which they alone are selected. The reason is that unassigned units whose potential outcomes are affected by spillovers could self-select into or out of treatment. To see this note that, if data from self selecting units were used, the only difference in estimates would be under assignment 5, in which case the estimated effect would be 2 rather than 4, because idle unit 3 self-selected to control and experienced changes in potential outcomes due to the assignment of unit 2.

	p_{Z^1}	Z^1		Y^1		Y^2		$Y^3 = \widehat{\tau}$		t
Assignment Index	Probability of Assignment	Manipuland		Actual Assignment		Potential Outcome		Estimate Diff.-in-means		Diagnostic − Stat. $\widehat{\tau} - \tau$
1.	$\frac{1}{6}$	$\begin{bmatrix} 1 \\ 0 \\ \oslash \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$	$\rightarrow$	0	$\rightarrow$	-4
2.	$\frac{1}{6}$	$\begin{bmatrix} 1 \\ \oslash \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 4 \\ 0 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
3.	$\frac{1}{6}$	$\begin{bmatrix} \oslash \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 4 \\ 0 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
4.	$\frac{1}{6}$	$\begin{bmatrix} \oslash \\ 0 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$	$\rightarrow$	0	$\rightarrow$	-4
5.	$\frac{1}{6}$	$\begin{bmatrix} 0 \\ 1 \\ \oslash \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
6.	$\frac{1}{6}$	$\begin{bmatrix} 0 \\ \oslash \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 4 \\ 4 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
7.	0	$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
8.	0	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix}$	$\rightarrow$	4	$\rightarrow$	0
9.	0	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$	$\rightarrow$	0	$\rightarrow$	0
10.	0	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\rightarrow$	NaN	$\rightarrow$	NaN
								$\mathbb{E}(\widehat{\tau}) = \frac{16}{6}$		$\theta_{\text{bias}} = \mathbb{E}(t) = -\frac{8}{6}$

Table 2: Illustration of the Outcomes of a Research Design as a Function of Potential Random Assignments. Ten possible assignments of a manipuland Z^1 are displayed. The estimand $\tau = 4$ is defined as the average difference between potential outcomes for a unit in an assignment in which it is assigned to treatment and all other units assigned to control and its potential outcomes under an assignment in which all units assigned to control (here, assignment 10). NaN indicates that the estimates and diagnostic statistics are undefined in assignments 9 and 10, due to division by zero because there are no units assigned to the treatment condition. The expectation of the estimator $\mathbb{E}(\widehat{\tau})$ and the bias diagnosand θ_{bias} are displayed at the bottom of the table.

C. Diagnoses for the Examples in Section 3

The code examples can be downloaded from the internet and run using the free, open source statistical package R. First, install the DeclareDesign software as follows:

```
source("https://declaredesign.org/install.R")
```

Code for running the examples is below. The complete replication code for Section 3 and Section 4 is available at https://declaredesign.org/paper/replication.Rmd. Further details on the R software package, including other examples and documentation, can be found at declaredesign.org.

C.1 Descriptive Inference

```
# Download design object from design library
design <- get_design("descriptive_inference")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	0.488
mean(estimate)	0.509
sd(estimate)	0.056
power	0.070
bias	0.021
RMSE	0.060
coverage	0.931

C.2 Discovery

```
# Download design object from design library
design <- get_design("discovery")
# Diagnose the design
diagnose_design(design = design)
```

estimator_label	diagnosand_label	diagnosand
estimator_right	mean(estimand)	0.500
estimator_right	mean(estimate)	0.504
estimator_right	sd(estimate)	0.185
estimator_right	bias	0.004
estimator_right	RMSE	0.185

estimator_label	diagnosand_label	diagnosand
estimator_right	coverage	0.952
estimator_right	power	0.771
estimator_right	type S rate	0.003
estimator_split_sample	mean(estimand)	0.500
estimator_split_sample	mean(estimate)	0.635
estimator_split_sample	sd(estimate)	0.322
estimator_split_sample	bias	0.135
estimator_split_sample	RMSE	0.349
estimator_split_sample	coverage	0.697
estimator_split_sample	power	0.606
estimator_split_sample	type S rate	0.023
estimator_wrong	mean(estimand)	0.500
estimator_wrong	mean(estimate)	1.000
estimator_wrong	sd(estimate)	0.045
estimator_wrong	bias	0.500
estimator_wrong	RMSE	0.503
estimator_wrong	coverage	0.000
estimator_wrong	power	1.000
estimator_wrong	type S rate	0.000

C.3 Matching

```
# Download design object from design library
design <- get_design("matching")
# Diagnose the design
diagnose_design(design = design)
```

estimator_label	diagnosand_label	diagnosand
dim	mean(estimand)	1.000
dim	mean(estimate)	3.395
dim	sd(estimate)	0.080
dim	bias	2.395
dim	RMSE	2.396
dim	coverage	0.000
dim	power	1.000
dim	type S rate	0.000
matching	mean(estimand)	1.000
matching	mean(estimate)	1.517
matching	sd(estimate)	0.069
matching	bias	0.517
matching	RMSE	0.522
matching	coverage	0.000
matching	power	1.000
matching	type S rate	0.000

C.4 Regression Discontinuity

```
# Download design object from design library
design <- get_design("regression_discontinuity")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	0.000
mean(estimate)	0.009
sd(estimate)	0.230
bias	0.009
RMSE	0.230
coverage	0.948
power	0.052
type S rate	1.000

C.5 Model-Based Estimands

```
# Download design object from design library
design <- get_design("model_based_estimand")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	2.000
mean(estimate)	2.143
sd(estimate)	0.422
bias	0.143
RMSE	0.446
coverage	0.964
power	0.983
type S rate	0.000

C.6 Bayesian Estimation Strategies

```
# Download design object from design library
design <- get_design("bayesian_estimation")
# Diagnose the design
diagnose_design(design = design)
```

estimator_label	diagnosand_label	diagnosand
flat_prior	Avg. % Reduction in Variance (Prior vs. Posterior)	-0.981
flat_prior	Avg. Maximum A Posteriori	0.210
flat_prior	Avg. Posterior Mean	0.216
flat_prior	Bias in Maximum A Posteriori	0.000
flat_prior	Bias in Posterior Mean	0.005
flat_prior	Coverage Probability of Posterior Mean	0.825
flat_prior	Average Shift in Mean (Prior vs. Posterior)	-0.284
flat_prior	True Population Proportion	0.210
info_prior	Avg. % Reduction in Variance (Prior vs. Posterior)	-0.967
info_prior	Avg. Maximum A Posteriori	0.216
info_prior	Avg. Posterior Mean	0.221
info_prior	Bias in Maximum A Posteriori	0.005
info_prior	Bias in Posterior Mean	0.011
info_prior	Coverage Probability of Posterior Mean	0.816
info_prior	Average Shift in Mean (Prior vs. Posterior)	-0.279
info_prior	True Population Proportion	0.210

C.7 Process tracing

```
# Download design object from design library
design <- get_design("process_tracing")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
Estimand	0.500
Est based on SG	0.524
Bias	0.024
Conditional bias when K seen	0.000
Conditional bias when K not seen	0.026

D. Diagnoses for the Examples in Section 4

D.1 Gains from Covariate Controls

```
# Download design object from design library
design <- get_design("covariate_control")
# Diagnose the design
diagnose_design(design = design)
```

estimator_label	diagnosand_label	diagnosand
M1	mean(estimand)	0.750
M1	mean(estimate)	0.997
M1	sd(estimate)	0.406
M1	bias	-0.003
M1	RMSE	0.406
M1	coverage	0.950
M1	power	0.646
M1	type S rate	0.005
M2	mean(estimand)	0.750
M2	mean(estimate)	0.997
M2	sd(estimate)	0.415
M2	bias	-0.003
M2	RMSE	0.415
M2	coverage	0.951
M2	power	0.628
M2	type S rate	0.007
M3	mean(estimand)	0.750
M3	mean(estimate)	0.997
M3	sd(estimate)	0.422
M3	bias	-0.003
M3	RMSE	0.422
M3	coverage	0.953
M3	power	0.614
M3	type S rate	0.007
B	mean(estimand)	0.750
B	mean(estimate)	0.010
B	sd(estimate)	0.707
B	bias	0.010
B	RMSE	0.707
B	coverage	0.957
B	power	0.043
B	type S rate	1.000

To download the template used in this section, type,

```
template <- get_template("covariate_control")
```

D.2 Handling Spatial Spillovers

```
# Download design object from design library
design <- get_design("spatial_spillovers")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	1.000
mean(estimate)	0.944
sd(estimate)	4.241
bias	-0.056
RMSE	4.241
coverage	0.000
power	1.000
type S rate	0.407

To download the template used in this section, type,

```
template <- get_template("spatial_spillovers")
```

D.3 Assigning Multiple Treatments

```
# Download design object from design library
design <- get_design("factorial")
# Diagnose the design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	0.000
mean(estimate)	0.000
sd(estimate)	0.178
bias	0.000
RMSE	0.178
coverage	0.953
power	0.047
type S rate	1.000

```
# Download design object from design library
design <- get_design("three_arm")
# Diagnose design
diagnose_design(design = design)
```

diagnosand_label	diagnosand
mean(estimand)	0.500
mean(estimate)	0.499
sd(estimate)	0.111
bias	-0.001
RMSE	0.111

diagnosand_label	diagnosand
coverage	0.950
power	0.994
type S rate	0.000

To download the template used in this section, type,

```
template <- get_template("factorial")
```