A general framework for learning about research designs[†]

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

1/15/2018

Abstract

Researchers need to select high quality research designs and communicate those designs to readers. Both tasks are difficult. We provide a framework for formally characterizing the analytically relevant features of a research design. In standard applications, the approach to design declaration that we describe requires defining a model of the world (M), an inquiry (I), a data strategy (D), and an answer strategy (A). Declaration of these features in code provides sufficient information for researchers and readers to use Monte Carlo techniques to diagnose properties such as power, bias, external validity, and other "diagnosands." Declaring a design lays researchers' assumptions bare. Ex ante design declarations can be used to improve designs and facilitate preregistration, analysis, and reconciliation of intended and actual analyses. Ex post design declarations are also useful for describing, sharing, reanalyzing, and critiquing existing designs. We provide open-source software, `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, Erin Hartman, 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 EPSA 2016, APSA 2016, and EGAP 18 meetings for helpful comments. The methods proposed in this paper are implemented in an accompanying open-source software package, `DeclareDesign` (Blair et al. 2016).

[‡]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`

Empirical political scientists routinely face two research design problems. First, we need to select a high-quality design, given resource constraints. Second, we need to convince readers and reviewers of the design's high quality.

To select strong designs, we often rely on rules of thumb, rudimentary power calculators, or generic best practices from the methodological literature that may assume ideal conditions that do not hold in real-world applied settings. These relatively informal practices sometimes result in the selection of suboptimal designs, or worse, designs that are simply too weak to deliver useful answers.

To convince others of the quality of our designs, we defend them with citations to previous studies that used similar approaches; power analyses that may rely on assumptions unknown even to ourselves; or in rare cases, ad hoc simulation code.

In this paper we describe an approach to address these problems. We first provide a framework to describe the distinct elements of a research design. The *MIDA* framework asks researchers to provide information about their background model (M), their inquiry (I), their data strategy (D), and their answer strategy (A). We then introduce 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. We say a design is "diagnosand-complete" when a diagnosand can be estimated from the declared features of the design. We do not have a general notion of a *complete* design, but rather adopt an approach in which the purposes of the design determine which features must be declared. Many different designs can be defined in terms of their diagnosand-completeness: including causal inference strategies employing observational, experimental or qualitative methods, as well as descriptive and exploratory strategies.

Using this framework researchers can *declare*[1] research designs mathematically and as computer code objects and then *diagnose* the statistical properties of the design relying on this declaration. We recommend researchers diagnose their designs using Monte Carlo simulation, which can produce both estimates of diagnosands and of simulation uncertainty. We implement this framework in the companion `R` package `DeclareDesign`.

[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.

The formal characterization and diagnosis of designs before implementation can serve many purposes, at different stages of a research cycle. First, it lets researchers learn about and improve their inferential strategies. Later, a researcher may wish to include design declaration and diagnosis as part of a preanalysis plan or a funding request. Even if only declared ex-post, formal declaration still has benefits. The complete characterization can help readers understand the properties of a research project, facilitate transparent replication, and contribute to re-analysis decisions.

We do not expect there to be general agreement on the set of diagnosands that must be calculable in order for a design to be complete "enough." Rather, domain-specific standards might be agreed upon among members of particular research communities. A standard set might include power, bias, root mean-squared-error, and coverage. Others concerned about the policy impact of a given treatment might require a design that is diagnosand-complete for an out-of-sample diagnosand, such as bias relative to the population average treatment effect.

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 information prior to launching their study. For exploratory work, by contrast, 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 they would have drawn had different data had been realized. Thus variation in research strategy limits the utility of our procedure for different types of research.

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

1. Research Designs and Diagnosands

We consider a general description of a research design as the specification of a problem and a strategy to answer it. Our framework is an elaboration of ideas proposed by King, Keohane and Verba (1994, p. 13), who enumerate four components of a research design: a theory, a research question, data, and an approach to using the data. Formalized, this framework can be used to structure a general procedure for assessing research designs' inferential properties, by, for example, assessing the bias or precision of results from a design given the theory, the data collection process, the research question, and inferential strategy. We extend this four component framework using recent advances in the theory of causal inference. Specifically, we follow Pearl's (2009) approach to structural modeling, which gives a syntax for mapping design components to design outputs. We combine this causal modeling approach with the potential outcomes framework as presented in Imbens and Rubin (2015), which many political scientists use to clarify their inferential targets.

1.1 Elements of a Research Design

The specification of a problem requires a description of the world and the question to be asked about the world as described. The answering requires a description of what information is used and how conclusions are reached given the information.

At its most basic we think of a research design, Δ, as including four elements $< M, I, D, A >$:

1. A **model**, M, of how the world works. In general following Pearl's definition of a probabilistic causal model we will assume that a model contains three core elements. First, a specification of the variables X about which research is being conducted. This includes endogenous and exogenous variables (V and U respectively) and the ranges of these variables. In the formal literature this is sometimes called the *signature* of a model (e.g., Halpern 2000). Second, a specification of how each endogenous variable depends on other variables (the "functional relations" or "potential outcomes"), F. Third, a probability distribution over exogenous variables, $P(U)$.

2. An **inquiry**, I, about the distribution of variables, X, perhaps given interventions on some variables. In many applications I might be thought of as the estimand. Using Pearl's

notation we can distinguish between questions that ask about the conditional values of variables, such as $\Pr(X_1|X_2 = 1)$ and questions that ask about values that would arise under interventions: $\Pr(X_1|do(X_2 = 1))$.[2] We let a^M denote the answer to I provided by the model. Under model M, answer a^M is generated with probability $P_M(a^M)$.

3. A **data** strategy, D, generates data d on X. Data d arises, under model M with probability $P_M(d|D)$. Note that implicitly the data strategy includes sampling strategies and assignment strategies, which we denote with P_S and P_Z respectively.[3]

4. An **answer** strategy, A, that generates answer a^A using data d. Under model M, answer a^A is generated with probability $P_M(a^A|D, A)$.

A key feature of this bare specification is that if M, D, and A are sufficiently well described, the answer to question I has a distribution $P_M(a^A|D)$; moreover one can construct a distribution of comparisons of this answer to the correct answer, under M, for example by assessing $P_M(a^M - a^A|D)$. One can also compare this to results under different data or analysis strategies, $P_M(a^M - a^A|D')$ and $P_M(a^M - a^{A'}|D)$, and to answers generated under alternative models, $P_M(a^{M'} - a^A|D)$, as long as these possess signatures that are consistent with inquiries and answer strategies.

Many social scientists will be familiar with a statistical framework that distinguishes between an estimand, an estimator, and an estimate. In our terms, an estimate is an answer a^A. An estimator is a procedure that is jointly described by the Data Strategy D and the Answer Strategy A. An estimand is the "true" answer in inquiry I, which in practice is thought of as the answer a^M that the Inquiry I receives from some background Model M. The overlapping and imperfect mapping of the estimand-estimator-estimate framework to the *MIDA* framework highlights the special utility of *MIDA*: the distribution of an estimator is a product of both how the data are collected and how they are analyzed; an estimand is a summary of a *theoretical* model that may or may not be correct.

MIDA captures the analysis-relevant features of a design, but it does not describe substantive

[2] The distinction lies in whether the conditional probability is recorded through passive observation or active intervention to manipulate the probabilities of the conditioning distribution. For example, $\Pr(X_1|X_2 = 1)$ might indicate the conditional probability that it is raining, given that Jack has his umbrella, whereas $\Pr(X_1|do(X_2 = 1))$ would indicate the probability with which it would rain, given Jack is made to carry an umbrella.

[3] Measurement strategies are also a part of data strategies though these can be thought of as a form of sampling—that is, the decision over the nodes on which data will be gathered.

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 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 enter into the model if the longer time until data collection affects subject recall of the treatment.

1.2 Diagnosands

The ability to calculate distributions of answers, given a model, opens multiple avenues for assessment and critique. How good is the answer you expect to get from this strategy? Would you do better with a different data strategy? With a different analysis strategy? How good is the strategy if the model is wrong in some way or another?

To allow for this kind of *diagnosis* of a design, we introduce two further concepts, both functions of research designs. These are quantities that a researcher or a third party could calculate with respect to a design.

1. A **Diagnostic Statistic** is a summary statistic generated from a "run" of a design—that is, the results given a possible realization of variables, given the model and data strategy. A diagnostic statistic may or may not depend on the model as well as realized data. For example the statistic: $e =$ "difference between the estimated and the actual average treatment effect (ATE)" depends on the model (where "actual" is defined under the model for a given run). The statistic $s = \mathbb{1}(p \leqslant 0.05)$, interpreted as "the result is considered statistically significant at the 5% level," does not depend on the model but it does presuppose an answer strategy that reports a p value.

   Diagnostic statistics have a distribution that results from the fact that both the model and the data generation, given the model, may be stochastic.

2. A **Diagnosand** is a summary of the distribution of a diagnostic statistic. For example, (expected) *bias* in the estimated treatment effect is $\mathbb{E}(e)$ and statistical *power* is $\mathbb{E}(s)$.

To illustrate, consider the following design. A model M specifies three variables X, Y, Z in some population (the signature) and some functional relationships between them that allow for the possibility of confounding (for example, $Y = bX + Z + \epsilon_Y; X = Z + \epsilon_X$, with $Z, \epsilon_X, \epsilon_Z$

distributed standard normal). The question of interest is "what is the average effect of a unit change in X on Y in the population?" Note that this question depends on the signature of the model, but not the functional equations of the model (the answer provided by the model does of course depend on the functional equations). Consider now a data strategy D, in which data is gathered on X and Y for n randomly selected units. An answer a^A, is then generated using ordinary least squares as the answer strategy, A.

Is this a good research design? One way to answer this question is with respect to the diagnosand "expected error." Here the model's functional equations provide an answer, a^M to the inquiry (for any draw of β), and so the distribution of the expected "error," *given the model*, $a^A - a^M$, can be calculated.

In this example the expected performance of the design may be poor because the data and analysis strategy do not handle the confounding described by the model. In comparison, better performance may be achieved through an alternative data strategy (e.g., where D' randomly assigned X to n units before recording X and Y) or an alternative analysis strategy (e.g., A' conditions on Z). These design evaluations depend on the model, and so one might reasonably ask how performance would look were the model different (for example if it allowed for spillovers or effect heterogeneity).

In all cases the evaluation depends on the assessment of a diagnosand, and comparing the diagnoses to what could be achieved under alternative designs.

1.3 Choice of Diagnosands

What diagnosands should researchers choose? Although researchers commonly focus on statistical power, a larger range of diagnosands can be examined and may provide more informative diagnoses of design quality. We list and describe some of these in Table 1, indicating for each the design information that is required in order to calculate them.

This set of statistics allows researchers to understand the properties of the estimates across possible realizations of the data and how successful their data and analysis strategies are at estimating estimands. Though these are frequentist properties, many of the diagnosands can be used to assess Bayesian estimation strategies (see Rubin 1984), and as we illustrate below there are diagnosands unique to Bayesian strategies.

Diagnosand	Description	Required: M	I	D	A
Power	Probability of rejecting null hypothesis of no effect	✓		✓	✓
Estimation Bias	Expected difference between estimate and estimand	✓	✓	✓	✓
Sampling Bias	Expected difference between population average treatment effect and sample average treatment effect (Imai, King and Stuart 2008)	✓	✓	✓	
RMSE	Root mean-squared-error	✓	✓	✓	✓
Coverage	Probability that estimand falls within confidence interval	✓	✓	✓	✓
SD of Estimates	Standard deviation of estimates	✓		✓	✓
SD of Estimands	Standard deviation of estimands	✓	✓	✓	
Imbalance	Expected distance of covariates across treatment conditions (Mahalanobis 1936; Gu and Rosenbaum 1993)	✓		✓	
Type S Rate	Probability estimate has incorrect sign, if statistically significant (Gelman and Carlin 2014)	✓	✓	✓	✓
Exaggeration Ratio	Expected ratio of absolute value of estimate to estimand, if statistically significant (Gelman and Carlin 2014)	✓	✓	✓	✓
Value for money	Probability that the estimated effect is at least as large as x	✓		✓	✓
Robustness	Joint probability of rejecting the null hypothesis across multiple tests	✓		✓	✓

Table 1: Examples of diagnosands and the elements of the Model (M), Inquiry (I), Data Strategy (D), and Answer Strategy (A) required in order for a design to be diagnosand-complete for each diagnosand.

Diagnosands can also be defined for design properties that are often discussed informally but rarely subjected to formal investigation. For example one might define an inference as "robust" if the same inference is made under different analysis strategies, or an intervention as having "value for money" if a set of estimates have at least a specified minimal magnitude. A diagnosis would then report the chances that an inference is considered robust or an intervention is deemed to have value for money.

1.4 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: the Consolidated Standards of Reporting Trials (CONSORT) Statement widely used in medicine includes 22 features and other proposals range from nine to 60 components.[4]

We propose a conditional notion of completeness: we say a design is "diagnosand-complete" for a given diagnosand if that diagnosand can be calculated from the declared design. Thus a design that is diagnosand complete for one diagnosand may not be for another. Consider for example the diagnosand "statistical power." Power is the probability that a p-value is lower than

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

some critical value. Thus, power-completeness requires that the answer strategy return a p value. It does not, however, require a well defined estimand (hence the lack of a checkmark under I on Table 1). Bias- or RMSE-completeness in contrast do not require a hypothesis test, but do require the specification of an estimand.

Our notion of diagnosand-completeness does not encompass all of the information relevant to research design. It instead clarifies the assumptions under which a design has good inferential properties.[5]

2. Existing Methods for Diagnosing Research Designs

Researchers commonly assess designs using one of three methods: analytical formulae for simple studies such as a sample of n units to estimate a population parameter (e.g., Cohen 1977; Haseman 1978; Muller and Peterson 1984; Muller et al. 1992; Lenth 2001); bespoke Monte Carlo simulation code written to diagnose specific studies; and pre-existing software tools. Methods falling into the third category are widely used and available as Web apps, general statistical software (e.g., `easypower` for `R` and Power and Sample Size for Stata), and standalone software (e.g., Optimal Design, G*Power, nQuery, SPSS Sample Power). Despite their popularity, these pre-existing tools cannot calculate key diagnosands for even relatively simple designs.

We conducted a census of the currently available computational diagnostic tools for research designs. We assessed tools listed in four reviews of the literature (Kreidler et al. 2013; Guo et al. 2013; Groemping 2016; Green and MacLeod 2016), in addition to the first thirty results from Google searches of the terms "statistical bias calculator," "statistical power calculator," and "sample size calculator."[6] Thirty of the 143 tools we identified were able to diagnose inferential properties of designs, such as their power.[7]

Using these 30 tools, we attempted to diagnose the following hypothetical research design. A

[5]Diagnosand-completeness only conveys what aspects of a design must be declared in order for some diagnostic feature to be queried, but it does not convey what information is required to make such diagnosis *believable*. A bias-complete design that excludes the possibility of bias from Hawthorne effects may be declared. Whether the estimated bias of the design is credible or not depends on the credibility of the model used to generate the diagnostic statistics. Different research communities set different standards for what constitutes sufficient information to make such conjectures about the world plausible. With respect to effect sizes, for example, some organizations may want to see how diagnoses vary across the entire range of conceivable effects, while others may require researchers to conduct a relevant meta-analysis or even a baseline survey in order to bolster the assumptions feeding into their design declarations.

[6]We found no admissable tools using the term "statistical bias calculator."

[7]See Online Appendix Section S2 for further details on the tool survey.

team of researchers wants to assess the effectiveness of a new voter turnout strategy. Their **M**odel allows the effectiveness of the campaign to vary depending on the size of the voting precinct, but their **I**nquiry is nevertheless the average treatment effect. Because of budget constraints, the team's **D**ata strategy involves sampling five precincts at random, then randomly assigning 10 households to treatment within each precinct, i.e., blocking by voting precinct. This procedure generates different probabilities of assignment across blocks, so the team is considering two **A**nswer strategies: a block-level fixed effects estimator (BFE) or an estimator that incorporates both inverse probability weights and block fixed effects (IPW-BFE).

The team seeks to answer three diagnostic questions:

1. What is the power of each estimator?
2. What is the bias of each estimator with respect to the average treatment effect?
3. Given the answers to 1 and 2, which approach should the team pre-register as the main analysis?

Using the tools surveyed for this article, the team would be unable to answer any of these questions correctly, because those tools cannot incorporate key features of the design. As shown in Table 2, none of the tools was able to diagnose the design while taking account of: the posited correlation between block size and potential outcomes; the sampling strategy; the exact randomization procedure; the formal definition of the estimand as the population average treatment effect; or the use of inverse-probability weighting.[8] As a result, no tool was able to calculate the power for the IPW-BFE estimator. Moreover, no tool was able to calculate the design's bias, root mean-squared-error, or coverage.

We compared the power calculations from these 30 tools to the true power of a simulated version of the researcher's design under assumptions about the data-generating process, which we calculated in `R` using the companion software to this paper, `DeclareDesign`. (Blair et al. 2016).Starting from a large finite population of interest in which each unit has a treatment and control potential outcome, we first calculate the true population average treatment effect (PATE), then randomly sample five blocks, assign ten units within each block to the treatment and the

[8]The one tool (GLIMMPSE) that was able to account for the blocking strategy encountered an error and was unable to produce diagnostic statistics.

(a) Declare Elements of Designs

	Design feature	No.
(M)	Effect and block size correlated	0/30
(I)	Estimand	0/30
(D)	Sampling procedure	0/30
(D)	Assignment procedure	0/30
(D)	Block sizes vary	1/30
(A)	Probability weighting	0/30

(b) Diagnosis Capabilities

Diagnosis	No.
Power (DIM estimator)	28/30
Power (BFE estimator)	13/30
Power (IPW-BFE estimator)	0/30
Bias (*any* estimator)	0/30
Coverage (*any* estimator)	0/30
SD of estimates (*any* estimator)	0/30

Table 2: Existing tools cannot declare many core elements of designs and, as a result, can only calculate some diagnosands. Panel (a) indicates the number of tools that allow declaration of a particular feature of the design as part of the diagnosis. In the first row, for example, 0/30 indicates that no tool allows researchers to declare correlated effect and block sizes. Panel (b) indicates the number of tools that can perform a particular diagnosis.

rest to control and then estimate the treatment effects using naive difference-in-means (DIM),[9] BFE and IPW-BFE. The parameters from this simulation exercise are used to calculate the power of the design using the 30 identified tools.

While almost all (28/30) of the tools were able to provide an estimate of the design's power when using the DIM estimator, fewer than half (13/30) were able to provide an estimate of the power using the BFE estimator, and none were able to provide a power estimate for the design when using the IPW-BFE estimator. The tools substantially exaggerated power estimates for the DIM and BFE estimators – by an average of 15 and 13 percentage points, respectively. The tools overestimate power because they assume the estimators are unbiased. However, simulations show the estimates produced by the DIM and BFE estimators are lower than the true effect on average, due to the negative correlation between a unit's treated potential outcome and its probability of assignment to treatment. Because the assessed tools base power calculations on the true underlying effect, which is larger than the estimates provided by those two answer strategies ($E[a^M] > E[a^A]$), they exaggerate the design's power.

Using the companion software, we show that the IPW-BFE estimator is better powered and less biased (in terms of the PATE) than the BFE estimator. However, power is a misleading indicator of the efficiency of the IPW-BFE strategy: it is better powered because it produces biased variance estimates that lead to a coverage probability that is too low. In terms of RMSE and the standard deviation of estimates, the IPW-BFE strategy does not outperform the BFE estimator.

[9] We include the DIM estimator because it is the only case that many of the tools assessed can handle and thereby enables direct comparison of their performance.

The exercise thus highlights why power and sample size calculations alone are insufficient to fully assess the tradeoffs between these relatively simple design alternatives.

In sum, existing tools cannot incorporate the information required to assess the probative value of research designs correctly. This shortcoming does not derive from any statistical weakness in the tools. Rather, they lack a framework specifying what design features must be declared to fully diagnose the design's inferential properties. Moreover, these tools' pre-defined diagnostic methods abstract from core features of particular designs, and thus fail to faithfully approximate the answers realworld designs would provide. We argue below that computer-based simulation of designs provides better answers across a variety of research contexts.

3. Declaring and Diagnosing Research Designs in Practice

A design that can be described mathematically (as in Section 1) can also be declared in computer code and then simulated in order to diagnose its properties. The core advantage of simulation over diagnosis through analytic solutions is that diagnosands can be quantified even where closed-form solutions do not exist or are difficult to derive. The top panel of Table 3 shows how to declare a design in code using the companion software to this paper, `DeclareDesign` (Blair et al. 2016). The resulting set of objects (`P_U`, `F`, `I`, `p_S`, `p_Z`, and `A`) are functions. A single simulation calls each of these functions successively as shown in steps 1-5. A design diagnosis conducts m simulations, then summarizes the resulting distribution of diagnostic statistics in order to estimate the diagnosand.

Diagnosands can be estimated with higher levels of precision by increasing m. 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 characterize our uncertainty 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 the standard error reflects both estimation uncertainty (simulation error) and fundamental uncertainty (true variability in the diagnosand, for example across possible population draws).[10]

[10]This procedure depends on the researcher choosing a "good" diagnosand estimator. In nearly all cases, diag-

	Design Declaration	Code
M	Declare background variables $P(U)$	`P_U <- declare_population(x_1 = rnorm(N), N = 100)`
	Declare functional relations F	`F <- declare_potential_outcomes(Y ~ x_1 + Z)`
I	Declare inquiry I	`I <- declare_estimand(PATE = mean(Y_Z_1 - Y_Z_0))`
D	Declare sampling p_S	`p_S <- declare_sampling(n = 50)`
	Declare assignment p_Z	`p_Z <- declare_assignment(m = 25)`
A	Declare answer strategy, A	`A <- declare_estimator(Y ~ Z, estimand = I)`
	Declare design, <M, I, D, A>	`D <- declare_design(P_U, F, I, p_S, p_Z, A)`

	Design Simulation (1 draw)	Code
1	Draw a population u using $P(U)$	`u <- P_U()`
2	Calculate an answer a^M to I using F and u	`uv <- F(u)` `a_M <- I(uv)`
3	Draw data, d, given sampling and treatment assignments specified in D and data realizations as determined by F and u	`d_1 <- p_S(uv)` `d <- p_Z(d_1)`
4	Calculate answers, a^A using A and d:	`a_A <- A(d)`
5	Calculate a diagnostic statistic t using a^A and a^M	`t <- a_A["est"] - a_M["estimand"]`

	Design Diagnosis (m draws)	Code
	Declare a diagnosand	`bias <- declare_diagnosands(bias = mean(est - estimand))`
	Calculate a diagnosand	`diagnose_design(D, diagnosands = bias, sims = m)`

Table 3: A procedure for declaring and diagnosing research designs using the companion software `DeclareDesign` (Blair et al. 2016). The top panel includes each element of a design that can be declared along with code used to declare them. The middle panel includes the steps in words and code in order to simulate that design. The bottom panel includes the procedure to diagnose the design.

Our companion software facilitates design diagnosis for beginner to intermediate coders in `R`. Those with no coding experience can use the online design declaration and diagnosis wizard, available at `DeclareDesign.org`. The website also contains instructions for implementing this framework in Stata. In the Appendix, we provide a simple example and explains how each step corresponds to the *MIDA* framework.

Design diagnosis through simulation does place a burden on researchers to come up with a substantive model, M. Since researchers presumably want to learn about the model, declaring it in advance may seem to beg the question. Yet declaring a model is unavoidable when diagnosing

nosands 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.

designs. In practice it is already familiar to any researcher who has calculated the power of a design, which requires the specification of effect sizes. The seeming arbitrariness of the declared model can be mitigated by assessing the sensitivity of diagnosis to alternative models and strategies, which is relatively straightforward given a diagnosand-complete design declaration. Just as power calculators focus attention on minimum detectable effects, design declaration offers not only a tool to demonstrate a design's desirable qualities but also lays bare *under what assumptions* a design has desirable properties.

In the next three sections, we outline how research designs that aim to answer causal, descriptive, and exploratory research questions can be declared and diagnosed in practice.

3.1 Causal Inference

The approach to design diagnosis we propose can be used to declare and diagnose a range of research designs typically employed to answer causal questions in the social sciences.

Observational Regression-Based Strategies. Many observational studies seek to make causal claims, but do not explicitly employ the potential outcomes framework, instead describing inquiries in terms of model parameters. Consider a study that seeks to estimate parameter β from a **M**odel of the form $y_i = \alpha + \beta x_i + \epsilon_i$. What is the estimand here? If we believe that this model 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 model? We run into a tautology if we want to assess the properties of strategies under different assumptions about data generation when the inquiry itself depends on the data generating model.

We can declare an **I**nquiry as some summary of differences in potential outcomes across conditions, β. 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 $f(x)$).[11] Our **D**ata

[11] An alternative might be to imagine some analogue of the ATT estimand, for example for an 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.

strategy will simply consist of the passive observation of units in the population, and we assess the performance of an **A**nswer strategy employing an OLS model to estimate β under different conditions.

To illustrate, we declare a design using the `R` package `DeclareDesign` and assess the properties of a regression estimate under the assumption that in the true data-generating process y is in fact a nonlinear function of x (for the full declaration, see Online Appendix Section S1.8). 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.

Process Tracing. While many qualitative researchers employ frameworks that may seem incompatible with the type of design declaration we have described, our approach may still be of use to qualitative designs that aim to confirm the presence or absence of a causal relationship (i.e., that are not focused on theory generation). Consider a stylized "process-tracing" design similar to ones described for example by Mahoney (2012) or in the Online 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 (CPO).

Declared in terms of MIDA, the **M**odel in such a study could stipulate a population $P(U)$ of N cases. The unobserved variable $T \in \{A, B, C, D\}$ gives the causal type of each case. In combination with the potential outcomes function F, the type variable creates a mapping between the presence or absence of the causal driver $X \in$ { No strong middle class, Strong middle class } and the presence or absence of the outcome $Y \in$ { No revolution, Revolution }. A types only have revolutions when there is no strong middle class, B types only have revolutions when there is a strong middle class, C types never have revolutions, and D types have them irrespective of the middle class's strength. A clue $K \in$ { Revolution not financed by domestic industry, Revolution financed by domestic industry } is generated with probability .2 only if the case

is a B type and the causal driver is present. The **D**ata strategy involves selecting one case at random for process-tracing, specifically one in which there is a middle class and a revolution. This leads to the **I**nquiry: is the case a B type, given the CPO? Or, formally, $Pr(T = B \mid K)$. A priori, since the case can only be a B or a D type given that X and Y are both present, the researcher might assign equal probabilities to the case being of either type. Suppose that the **A**nswer strategy involves inferring with certainty that the case is a type B when the clue is observed and remaining agnostic when it is not, such that $Pr(T = B \mid \neg K) = .5$. With these elements in hand, it is relatively straightforward to generate a distribution of diagnostic statistics $t = a^A - a^M$. Our implementation of this procedure using the `R` package `DeclareDesign` in Online Appendix Section S1.1 shows that the researcher's inference will be unbiased in the cases in which the CPO is observed ($E[t^K] = 0$), but not in those cases in which it is not ($E[t^{\neg K}] \neq 0$]), and so not overall ($E[t] \neq 0$]). The bias arises from the non-Bayesian property of the answer strategy: the researcher does not sufficiently discount the causal theory under investigation when disconfirmatory evidence comes to light.

Selection-on-Observables with Matching. In many observational research designs, the processes by which units are assigned to treatment are not known with certainty. In selection-on-observables designs, the unknown assignment procedure may, for example, be approximated by matching units on their observable traits to justify an assumption of as-if random assignment. Diagnosis in such instances can be helpful as a tool to explore the conditions under which such assumptions are justified. In Online Appendix Section S1.2, we declare a design with a **M**odel in which three observable random variables are combined in a probit process that assigns the treatment variable, Z. The **I**nquiry pertains to the average treatment effect of Z on the outcome Y among those actually assigned to treatment, which we estimate using an **A**nswer strategy that reconstructs the assignment process to calculate a^A. Our diagnosis shows that matching improves mean-squared-error ($E[(a^A - a^M)^2]$) relative to a naive difference-in-means estimator of the treatment effect on the treated (ATT), but can nevertheless remain biased ($E[a^A - a^M] \neq 0$) if the matching algorithm does not successfully pair units with equal probabilities of assignment.

Regression Discontinuity. While in selection-on-observables designs researchers do not typically know the assignment process, in other observational settings researchers may know how assignment works without necessarily controlling it. In regression discontinuity designs 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). The declaration of such designs involves a **M**odel that defines the unknown potential outcomes functions mapping average outcomes to the running and treatment variables. Our **I**nquiry concerns the average difference in potential outcomes as they limit toward the threshold of the running variable at which the assignment variable changes values. The **D**ata strategy involves passive observation and collection of the data. The **A**nswer strategy is a polynomial regression in which the assignment variable is linearly interacted with a fourth order polynomial transformation of the running variable. In Online Appendix Section S1.3, we declare and diagnose such a design. A key point to arise from the simulation is 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. Assessing the external validity of this design can be complicated: unless one postulates unobservable counterfactuals (such as the 'treated' outcome for a unit located below the treatment threshold), it is difficult to declare designs that are bias-complete with respect to the population or even sample average treatment effects.

Experimental Design. Experimental research may call particularly for design declaration and diagnosis because researchers are typically in direct control of many features of the design, beginning with assignment of treatments. A common choice faced in experimental research is between employing a 2-by-2 factorial design or a three-arm trial where the "both" condition is excluded. Consider a researcher studying two treatments who is interested in the effect of each treatment *conditional on the other treatment being in the control condition*. Should she choose a factorial design or a three-arm design? Focusing for simplicity on the effect of a single treatment, we declare two designs under a range of alternative models to help assess the tradeoffs. For both designs, we consider **M**odels $M_1, ..., M_K$, where we set the interaction between treatments to 0 for M_1, and increment it by $.5/(K-1)$ for each $M_{k \in 2,...,K}$. Our **I**nquiry is always the average treatment effect of treatment 1 given all units are in the control condition for treatment 2. We have two alternative **D**ata strategies under consideration: d' using an assignment strategy p'_Z, in which subjects are assigned to a control condition, treatment 1, or treatment 2, each with probability $1/3$; and d'' using p''_Z to assign subjects to each cell of a 2×2 with probability $1/4$. The **A**nswer strategy in both cases involves a regression of the outcome on both treatment indicators.

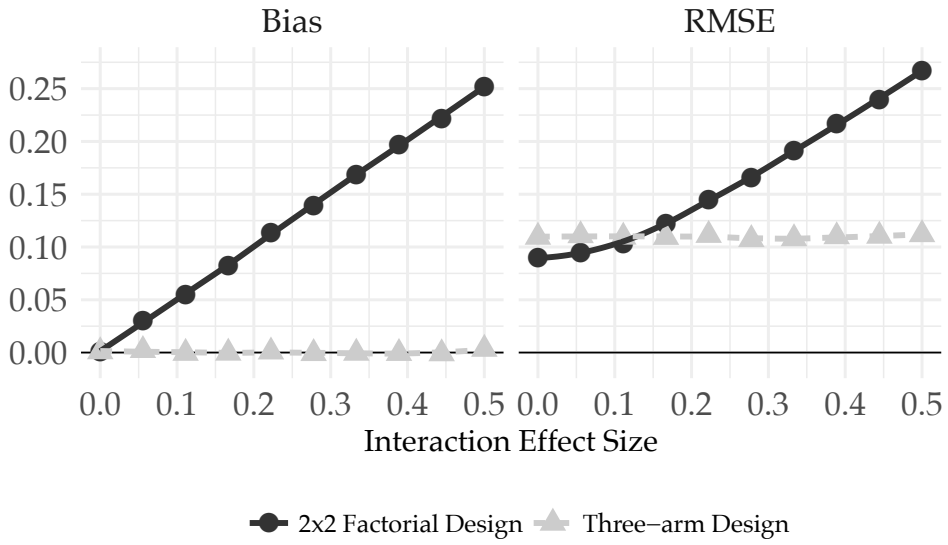

Figure 1: Diagnoses of Designs with Factorial or Three-Arm Assignment Strategies Illustrate a Bias-Variance Tradeoff. Bias (left) and root mean-squared-error (right) are displayed for two assignment strategies, a 2×2 treatment arm factorial design (black solid lines; circles) and a three-arm design (gray dashed lines; triangles) according to varying interaction effect sizes specified in the potential outcomes function (x axis).

We declare and diagnose this design and find that neither design exhibits bias when the true interaction term is equal to zero (Figure 1 left panel). The details of the declaration can be found in Online Appendix Section S1.4. 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. Moreover, there is a bias-variance tradeoff in choosing between the two designs (Figure 1 right panel). When the interaction term is small or close to zero, the factorial design is preferred, because it is 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 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.

3.2 Descriptive Inference

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 the Democratic candidate for president. Although seemingly very different from designs that focus on causal inference—because often there are no explanatory variables—the formal differences are not great.

Survey Designs. 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 Online Appendix Section S1.5, we declare a **M**odel in which latent voters are likely to vote for a candidate, but unlikely to reveal to interviewers their true propensity to vote. The **I**nquiry concerns the true underlying support for the candidate, while the **D**ata strategy involves a random population sample. The **A**nswer strategy involves looking at support for the candidate among likely voters. The design can be diagnosed to assess the risk of falsely concluding that the general election support of the democratic candidate is above 50%, given assumptions about how people report their voting proclivities.

Bayesian Descriptive Inference. 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. In Online Appendix Section S1.6, we declare a Bayesian descriptive inference design. The **M**odel stipulates a latent probability of success for each unit, and makes one binomial draw for each from this probability. The **I**nquiry pertains to to the latent probability, and the **D**ata strategy involves a random sample of relatively few units. There are two alternative **A**nswer strategies under consideration: in the first, the researcher stipulates uniform priors, with a mean of .5 and a standard deviation of .29; in the second, the priors place more probability mass at .5, with a standard deviation of .11. The design can be diagnosed not only in terms of its bias, but also as a function of quantities specific to Bayesian estimation approaches, such as the expected shift in the location and scale of the posterior distribution relative to the prior distribution. The diagnosis shows that the informative prior approach yields more certain and more biased inferences than the uniform prior approach. In terms of the bias-variance tradeoff, the informative priors decrease the posterior standard deviation by 40% relative to the uniform priors, but increase the bias by 33%.

3.3 Designs for 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 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.

We examine an exploratory data analysis *procedure* in which the researcher explores possible analysis strategies in a first stage on half of the data and in the second stage applies her preferred procedure to the second half of the data. Split-sample procedures such as this enable researchers to learn about the data inductively while protecting against Type I errors (for an early discussion of the design, see Cox 1975). In Online Appendix Section S1.7, we declare a design in which the **M**odel stipulates that education is a confounder for the effect of income on the outcome of interest, Y, while the **I**nquiry pertains to the unconfounded effect of income on Y. The **D**ata strategy simply involves passively recording the variables of interest. We compare three **A**nswer strategies: the "right" and "wrong" models, which do and don't condition the analysis of income on the concurrent effect of education, on the one hand, and on the other, a split-sample procedure that estimates effects on one half of the sample by selecting from three candidate models the one that has the best goodness of fit when estimated on the other half of the sample. The design is complete for a range of diagnosands (power, bias, RMSE, Type-S, etc.). The split-sample procedure reduces bias and power relative to selection of the "wrong" estimator. Researchers can declare and diagnose any exploratory procedure for which they can describe the domain of exploration (for example, the set of tests that will be conducted) and the decision rules (how the researcher selects among models or changes the analysis in response to test values).

4. Putting Declarations and Design Diagnosis to Use

We have described and illustrated a strategy for declaring research designs for which "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 describe implications for how designs are chosen, communicated, and challenged.

4.1 Making Design Choices

The move towards increasing credibility of research in the social sciences places a premium on considering alternative data strategies and analysis strategies at early stages of research projects, not only because it reduces researcher discretion, but more importantly because it can improve the quality of the final research design. While there is nothing new about the idea of determining features such as sampling and estimation strategies ex ante in order to maximize power, for example, in practice many designs are finalized late in the research process, after data are collected. Frontloading design decisions is difficult not only because existing tools are rudimentary and often misleading, as illustrated in Section 2, but because it is not clear in current practice what features of a design must be considered ex ante.

We provide a framework for identifying *which* features affect the assessment of a design's properties, declaring designs and diagnosing their inferential quality, and frontloading design decisions. Declaring the design's features in code enables direct exploration of alternative data and analysis strategies using simulated data; evaluating alternative strategies through diagnosis; and exploring the robustness of a chosen strategy to alternative models. Researchers can undertake each step before study implementation or data collection.

4.2 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 et al. 2015; 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. Bias may also arise from reviewers basing publication recommendations on statistical significance. Results-blind review processes are being introduced in some journals to address this form of bias (e.g. Findley et al. 2016).

However, the effectiveness of design registries and results-blind review in reducing the scope for either form of publication bias depends on clarity over which elements must be included to describe the design. In practice some registries rely on 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 requirement that design declarations be diagnosand-complete can clarify for researchers and third parties what aspects of a study need to be specified in order to meet standards 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.

Declaration of a diagnosand-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. Identifying how and whether the features of a design diverge between ex ante and ex post declarations highlights deviations from the pre-analysis plan. The magnitude of such deviations determines whether results should be considered exploratory or confirmatory. At present, this exercise requires a review of dozens of pages of text, such that differences (or similarities) are not immediately clear even to close readers.

4.3 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. Replication — whether verification, reanalysis of the original data, or reproduction using fresh studies — provides incentives for researchers to be clear and transparent in their analysis strategies, and can build confidence in findings.[12]

In addition to rendering the design more transparent, diagnosand-complete declaration can allow for a different approach to the re-analysis and critique of published research. A standard

[12]For a discussion of the distinctions between these different modes of replication, see Clemens (2017).

	Author's assumed **M**odel	Alternative claims on **M**odel
Author's proposed **A**nswer strategy	1	2
Alternative **A**nswer strategy	3	4

Table 4: Diagnosis Results Given Alternative Assumptions about the Model and Alternative Answer Strategies. Four scenarios encountered by researchers and reviewers of a study are considered depending on whether the model or the answer strategy differs from the author's original strategy and model.

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. The problem with this approach 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.

A more coherent strategy facilitated by design simulations would be to use a diagnosand-complete declaration to conduct "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 emphasizing abstract properties of the design, design replication 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.

Table 4 illustrates situations that may arise. In a declared design an author might specify situation 1: a set of claims on the structure of the variables and their potential outcomes (the model) and an estimator (the answer strategy). A critic might then question the claims on potential outcomes (for example questioning a no-spillovers assumption) or question estimation strategies (for example arguing for inclusion or exclusion of a control variable from an analysis), or both.

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

- **Home Ground Dominance.** If ex ante the diagnostics for situation 3 are better than for 1 then this gives grounds to switch to 3. That is, if 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 Models.** If the diagnostics in situation 2 are as good as in 1 but are better in situation 4 than in situation 3 this provides a robustness argument for altering estimation strategies.

- **Model Plausibility.** If the diagnostics in situation 1 are better than in situation 2, but the diagnostics in situation 4 are better than in situation 3, then this is cause for worry and the justification of a change in estimators depends on the plausibility of the different assumptions about potential outcomes.

Without a declared design—in particular the model and inquiry—none of these three criteria can be evaluated, complicating the defense of claims for both the critic and the original author.

We illustrate an application of these principles through a design replication of Björkman and Svensson (2009). Importantly, we are able to conduct a results-independent design replication because sufficient detail to simulate the design is provided in the original article and its supporting materials. The independence of the replication to the data not only makes replication possible in a context where the data is not yet publicly available, but more significantly focuses attention on features of the design rather than results.

We draw upon the "robustness to alternative models" criterion to claim that an alternative answer strategy would be superior to the original strategy employed by the authors, in the sense that the alternative approach exhibits less bias under plausible conjectures about the world. In the original study, Björkman and Svensson (2009) investigate whether community-based monitoring can improve health outcomes in rural Uganda. They focus on improvements in two important indicators: child mortality, defined as the number of deaths per 1000 live births among under-5 year-olds, taken at the catchment-area-level; and weight-for-age z-scores, which are calculated by subtracting from an infant's weight the median for their age from a reference population, and dividing by the standard deviation of that population. In the original design, the authors estimate a positive effect of the intervention on weight among surviving infants. However, they also find that the treatment greatly decreases child mortality.

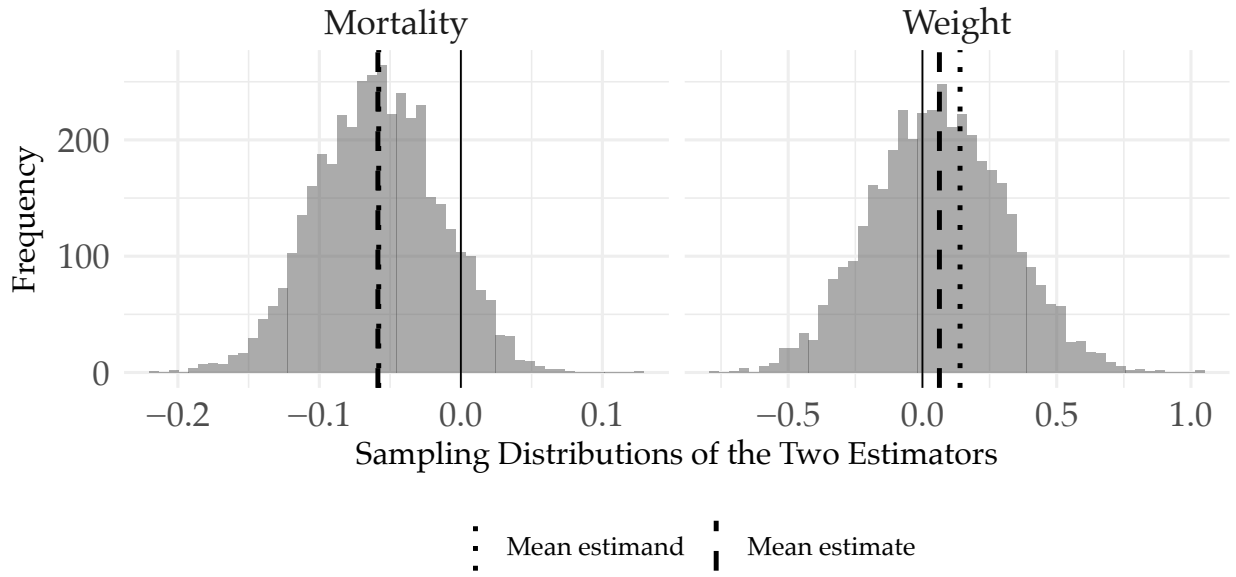

Figure 2: Data-independent replication of estimates in Björkman and Svensson (2009). Histograms display the frequency of simulated estimates of the effect of community monitoring on infant mortality (left) and on weight-for-age (right). The dashed vertical line shows the average estimate, the dotted vertical line shows the average estimand.

The weight of infants in control areas whose lives would have been saved if they had been in the treatment cannot be observed. We posit that unobserved variables, "family health" and "community health," may determine both whether infants survive early childhood and whether they are malnourished. We discuss these claims in detail in our design replication of Björkman and Svensson (2009) in Online Appendix Section S3. Figure 2 illustrates how the existence of an effect on mortality can pose problems for the unbiased estimation of an effect on weight-for-age when the two outcomes are correlated by community or family health. The histograms represent the frequency with which the design gives different answers to the inquiry about the effect of community monitoring on infant mortality and weight-for-age. The differences arise because the random sampling and assignment procedures select and assign different units on each run of the design. The dotted vertical line represents the true average effect, whereas the dashed line represents the average answer, i.e. the answer we expect the design to provide given our assumptions. Under our proposed model of the world the estimates of the effect on weight-for-age are biased downwards because it is precisely those infants with low health outcomes whose lives the treatment saves. This pulls down the treatment group's average weight outcome.

An alternative answer strategy is to attempt to subset the analysis of the weight effects to a group of infants whose survival does not depend on the treatment. In the original study, for example, the effects on survival are much larger among infants younger than two years old. If indeed the survival of infants above this age threshold is unaffected by the treatment, then it is possible to provide unbiased estimates of the weight-for-age effect by subsetting to this group (assuming effect homogeneity). In terms of bias, such an approach does at least as well if we assume that there is no correlation between weight and mortality, and better if such a correlation does exist. It thus satisfies the "robustness to alternative models" criterion.

A reasonable counter to this replication effort might be to say that the alternative answer strategy does not meet the criterion of home ground dominance with respect to RMSE: the power loss from subsetting to a smaller group may outweigh the bias reduction that it entails. In both cases, transparent arguments can be made by formally declaring and comparing the original and modified designs. While such criteria will not eliminate disputes they should at least help focus the discussion on the analytically-relevant issues.

4.4 Risks

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 here. 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 generated 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 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. 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 design that carries a 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.[13] Similarly there could be a risk that less attention is paid to measurement issues, which largely fall outside our framework. Simplification of the evaluation of formal properties of a design could instead, however, allow for a shift in attention towards examining other properties of a design such as measurement strategy or substantive and theoretical relevance.[14]

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. When a design cannot be declared, nondeclarability is all the framework provides, and in such cases we urge caution in drawing conclusions about design quality.

5. Conclusion

How can researchers assess the properties of research designs and improve them before implementation? Available tools do not allow scholars to faithfully characterize the features of common applied research designs. These tools often provide misleading assessments of design properties and sometimes are not able to provide an assessment at all. The approach described here, though simple, allows researchers to fully characterize, and thus to diagnose their designs in a manner consistent with their assumptions and plans. Of course, even a simulation-based claim to

[13] 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).

[14] 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 has good statistical properties).

unbiasedness that incorporates all features of a design is still only good 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—a demonstration that a strategy is biased for some estimand—may be powerful even when general analytic results do not exist.

The procedure for characterizing and diagnosing designs that we describe may produce multiple benefits. 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. Ex post, it can also make it easier for designs to be shared, improved, and critiqued.

References

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

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. "DeclareDesign Version 1.0." Software package for R, available at http://declaredesign.org.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3:993–1022.

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. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31(1):326–342.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cox, David R. 1975. "A note on data-splitting for the evaluation of significance levels." *Biometrika* 62(2):441–444.

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky and Thomas B. Pepinsky. 2016. "Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study." *Comparative Political Studies* 49(13):1667–1703.

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.

Green, Peter and Catriona J. MacLeod. 2016. "SIMR: an R package for power analysis of generalized linear mixed models by simulation." *Methods in Ecology and Evolution* 7(4):493–498.

Groemping, Ulrike. 2016. "Design of Experiments (DoE) & Analysis of Experimental Data." R Package.
URL: *https://CRAN.R-project.org/view=ExperimentalDesign*

Gu, Xing S. and Paul R. Rosenbaum. 1993. "Comparison of multivariate matching methods: Structures, distances, and algorithms." *Journal of Computational and Graphical Statistics* 2(4):405–420.

Guo, Yi, Henrietta L. Logan, Deborah H. Glueck and Keith E. Muller. 2013. "Selecting a sample size for studies with repeated measures." *BMC Medical Research Methodology* 13(1):100.

Halpern, Joseph Y. 2000. "Axiomatizing causal reasoning." *Journal of Artificial Intelligence Research* 12:317–337.

Haseman, Joseph K. 1978. "Exact sample sizes for use with the Fisher-Irwin test for 2 x 2 tables." *Biometrics* pp. 106–109.

Huber, John. 2013. "Is theory getting lost in the "identification revolution"?" Retrieved: January 8, 2018.
URL: *http://themonkeycage.org/2013/06/is-theory-getting-lost-in-the-identification-revolution/*

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the Royal Statistical Society: Series A (statistics in society)* 171(2):481–502.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Kreidler, Sarah M., Keith E. Muller, Gary K. Grunwald, Brandy M. Ringham, Zacchary T. Coker-Dukowitz, Uttara R. Sakhadeo, Anna E. Barón and Deborah H. Glueck. 2013. "GLIMMPSE: online power computation for linear models with and without a baseline covariate." *Journal of Statistical Software* 54(10).

Lenth, Russell V. 2001. "Some practical guidelines for effective sample size determination." *The American Statistician* 55(3):187–193.

Mahalanobis, Prasanta C. 1936. "On the generalised distance in statistics." *Proceedings of the National Institute of Sciences of India* pp. 49–55.

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

Muller, Keith E. and Bercedis L. Peterson. 1984. "Practical methods for computing power in testing the multivariate general linear hypothesis." *Computational Statistics & Data Analysis* 2(2):143–158.

Muller, Keith E., Lisa M. Lavange, Sharon Landesman Ramey and Craig T. Ramey. 1992. "Power calculations for general linear multivariate models including repeated measures applications." *Journal of the American Statistical Association* 87(420):1209–1226.

Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, Christopher D. Chambers, Gilbert Chin, Garret Christensen et al. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348(6242):1422.

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

Below, we demonstrate how each diagnosand-relevant feature of a simple design can be defined in code, with an application in which the assignment procedure is known. This could represent an experimental or quasi-experimental design.

$P(U)$ **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 <- declare_population(N = 1000, u = rnorm(N))
```

Each `declare` step creates a function, in this case a function that returns a data set of N observations with a variable named `u` drawn from a random normal distribution. For example, the population step $P(U)$ could have equivalently been created using the following function:

```
my_population_function <- function(N) { data.frame(u = rnorm(N)) }

my_population <- declare_population(
  handler = my_population_function, N = 1000)
```

$D(1)$ **Assignment 1: The sampling strategy**. Defines the distribution over possible samples for which outcomes are measured, p_S.

In the example below each unit generated by p_X is sampled with 10% probability. Again `my_sampling` describes a sampling strategy and not an actual sample.

```
my_sampling <- declare_sampling(n = 100)
```

$D(2)$ **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 <- declare_assignment(m = 50)
```

F **The structural equations, or potential outcomes function**. The potential outcomes function defines conjectured potential outcomes given interventions 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.[1] 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 <- declare_potential_outcomes(Y_Z_0 = u, Y_Z_1 = u + .25)
```

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 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 of 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.

I **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 <- declare_estimand(ATE = mean(Y_Z_1 - Y_Z_0))
```

A **The answer strategies** 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 estimator function takes data and returns an estimate of a treatment effect using the difference-in-means estimator, as well as a set of associated statistics, including the standard error, p-value, and the confidence interval.

```
my_estimator <- declare_estimator(Y ~ Z, estimand = my_estimand)
```

We then declare the design, which in this case primarily describes the order of the features, though it could include other changes to the data such as subsetting or adding variables.

```
my_design <- my_population + my_potential_outcomes + my_estimand +
             my_sampling + my_assignment + my_estimator
```

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

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, $\theta(D, Y, f, g)$. For simplicity, the two can be coded as a single function. For example, to calculate the bias of the design as a diagnosand is:

```
diagnosand <- declare_diagnosands(bias = mean(est - estimand))
```

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